Digital Human Sciences

New Objects – New Approaches

Edited by
Sonya Petersson

STOCKHOLM UNIVERSITY PRESS

Digital Human Sciences

New Objects—New Approaches

Edited by Sonya Petersson

STOCKHOLM
UNIVERSITY PRESS

Published by
Stockholm University Press
Stockholm University
SE-106 91 Stockholm, Sweden
www.stockholmuniversitypress.se

Supporting Agency (funding): The Board of Human Science,
Stockholm University

First published 2021
Cover designed by Christina Lenz, Stockholm University Press

Stockholm Studies in Culture and Aesthetics (Online) ISSN: 2002-3227

ISBN (Paperback): 978-91-7635-147-5
ISBN (PDF): 978-91-7635-144-4
ISBN (EPUB): 978-91-7635-145-1
ISBN (Mobi): 978-91-7635-146-8

DOI: https://doi.org/10.16993/bbk

Suggested citation:
Petersson, Sonya (ed.). 2021. *Digital Human Sciences: New Objects—New Approaches*. Stockholm: Stockholm University Press. DOI: https://doi.org/10.16993/bbk. License: CC-BY 4.0.

Stockholm Studies in Culture and Aesthetics

Stockholm Studies in Culture and Aesthetics (SiCA) (ISSN 2002-3227) is a peer-reviewed series of monographs and edited volumes published by Stockholm University Press. SiCA strives to provide a broad forum for research on culture and aesthetics, including the disciplines of Art History, Heritage Studies, Curating Art, History of Ideas, Literary Studies, Musicology, and Performance and Dance Studies.

In terms of subjects and methods, the orientation is wide: critical theory, cultural studies and historiography, modernism and modernity, materiality and mediality, performativity and visual culture, children's literature and children's theatre, queer and gender studies.

It is the ambition of SiCA to place equally high demands on the academic quality of the manuscripts it accepts as those applied by refereed international journals and academic publishers of a similar orientation. SiCA accepts manuscripts in English, Swedish, Danish, and Norwegian.

Editorial Board

Titles in the series

1. Rosenberg, T. 2016. *Don't Be Quiet, Start a Riot! Essays on Feminism and Performance*. Stockholm: Stockholm University Press. DOI: https://doi.org/10.16993/baf. License: CC-BY 4.0
2. Lennon, J. & Nilsson, M. (eds.) 2017. *Working-Class Literature(s): Historical and International Perspectives*. Stockholm: Stockholm University Press. DOI: https://doi.org/10.16993/bam. License: CC-BY 4.0
3. Tessing Schneider, M. & Tatlow, R. (eds.) 2018. *Mozart's* La clemenza di Tito: *A Reappraisal*. Stockholm: Stockholm University Press. DOI: https://doi.org/10.16993/ban. License: CC-BY 4.0
4. Petersson, S., Johansson, C., Holdar, M. & Callahan, S. (eds.) 2018. *The Power of the In-Between: Intermediality as a Tool for Aesthetic Analysis and Critical Reflection*. Stockholm: Stockholm University Press. DOI: https://doi.org/10.16993/baq. License: CC-BY 4.0
5. Hayden, H. 2018. *Modernism as Institution: On the Establishment of an Aesthetic and Historiographic Paradigm*. Stockholm: Stockholm University Press. DOI: https://doi.org/10.16993/bar. License: CC-BY 4.0
6. Lennon, J. and Nilsson, M. (eds.) 2020. *Working-Class Literature(s): Historical and International Perspectives. Volume 2*. Stockholm: Stockholm University Press. DOI: https://doi.org/10.16993/bbf. License: CC-BY 4.0

7. Petersson, S. (ed.) 2021. *Digital Human Sciences: New Objects—New Approaches*. Stockholm: Stockholm University Press. DOI: https://doi.org/10.16993/bbk. License: CC-BY 4.0

Peer Review Policies

Guidelines for peer review

Stockholm University Press ensures that all book publications are peer-reviewed. Each proposal submitted to the Press will be sent to a dedicated Editorial Board of experts in the subject area for evaluation. The full manuscript will be reviewed by chapter or as a whole by two external and independent experts.

A complete description of Stockholm University Press' peer-review policies can be found on the website: http://www.stockholmuniversitypress.se/site/peer-review-policies/

The Editorial Board of Stockholm Studies in Culture and Aesthetics applied a single-blind review during manuscript assessment. The Editor Sonya Petersson stepped down momentarily from her role in the Editorial Board during the review process of the book to avoid a conflict of interest.

Recognition for reviewers

Stockholm University Press and the Editorial Board would like to extend a special thanks to the reviewers, who contributed to the process of editing this book, for their work and time spent on reviewing the manuscript of this book.

Table of Contents

PART THREE: NEW OBJECTS—NEW APPROACHES

List of Figures and Tables

Figures

Tables

Introduction: Digital Human Sciences as a Field of Research

Sonya Petersson and Uno Fors

The ongoing digitization of culture and society and the ongoing production of new digital objects in culture and society require new ways of investigation, new theoretical avenues, and new multidisciplinary frameworks. In order to meet these requirements, this volume digs into questions concerning, for example: the epistemology of data produced and shared on social media platforms; the need for new legal concepts that regulate the increasing use of artificial intelligence in society; and the need for combinatory methods to research new media objects such as podcasts, web art, and online journals in relation to their historical, social, institutional, and political effects and contexts. Gathered around the perspective of what we call digital human sciences—a field of research that includes the humanities, the social sciences, and law—the chapters in this volume emanate from scholars situated within a broad range of disciplines. Nevertheless, they meet in the mutual objectives of researching the present digitization of culture and society and of discussing ways of doing research within this field of study.

Purpose of the Volume

Our purpose is twofold. First, we aim to discuss and develop methods and approaches for investigating digital society, digital culture, and digital media objects. Second, we aim to develop the new research field of the digital human sciences through the methodological discussion.

How to cite this book chapter:
Petersson, Sonya, and Uno Fors. "Introduction: Digital Human Sciences as a Field of Research." In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 1–20. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.a.

Our definition of *digital human sciences* is inclusive. "Digital human sciences" describes a multidisciplinary field of research aimed at the study of digital objects and environments and their significance for human beings and society. This includes, but is by no means limited to: investigations of agents and their roles in digital society; social and legal aspects; questions of liability and ethics; and interaction between human beings in digital systems and between human beings and digital entities. The field of the digital human sciences thus transgresses not only disciplinary borders but also borders between faculties (the humanities, the social sciences, and law). It could include objects of study such as internet history, online museums, computer games, social tweeting practices, AI systems in public administration or medical service, and web forums for political activism.

The question now is how we methodologically approach such topics and objects of study. Do they require a combination of digital tools and traditional close reading, a fusion of philosophical reasoning and web archaeology, or media theory, traditional sociology, and digital epistemology in a sort of union with contributions from law and information technology too? What kinds of exchanges over traditional disciplinary borders are beneficial for methodological developments and experiments? Where are the dividing lines between material infrastructure, intellectual infrastructure, and methodological procedures, and how should they be dealt with?

Each chapter both reflects on these types of general methodological questions and concretizes them in a particular case study. The volume is thus tied together by the running methodological discussion, while its particular cases are built around the diversity and specificity of social, cultural, medial, and legal objects, discourses, and practices in digital culture and society.

Beginnings

This volume is one of the results from the digital human sciences initiative at Stockholm University in 2016. A group of scholars from the humanities, social sciences, and law faculties was then appointed to form a committee for promoting, funding, and organizing research activities around the study of digital culture and society.[1] Initially, one of the central tasks was to map out the field. What is meant by "digital human sciences," what types of research objects and topics does it circumscribe, and how does it overlap with or depart from related fields of research? While the definition above sheds light on the first two questions, the third question will be addressed in the next section.

Digital Human Sciences and Related Fields of Research

The field of the digital human sciences is part of what is often and in the broadest sense referred to as digital studies. We align our project to the general objective of Arthur and Marilouise Kroker in the introduction to the by now classic reader *Critical Digital Studies*, as it encourages investigating culture and society in relation to:

> the potential of digital devices for shaping the ways in which we understand the world and communicate with one another to the unexplored implications of technological innovations—mobile media, cloud computing, social networking, augmented reality, 3D printing, drone technology—for both illuminating and, perhaps, sometimes constraining the human condition.[2]

As much as we subscribe to the importance of examining the various ways in which digital culture and society inflect the "human condition," we nevertheless avoid the term "digital studies." Its umbrella character could too easily be confusing.[3] First of all, it is often invoked in relation to academic milieus, infrastructures, and research output that could just as well and more aptly be related to the digital humanities. This is, for instance, the case when

---

1 See the Digital Human Sciences website for more information: https://dhv.blogs.dsv.su.se. The Digital Human Science Committee at Stockholm University 2016–20 included Uno Fors and Petter Karlström from the Department of Computer and Systems Science, Cecilia Magnusson Sjöberg from the Department of Law and Informatics, and Christer Johansson and Sonya Petersson from the Department of Culture and Aesthetics.

2 Arthur Kroker and Marilouise Kroker, eds., "Introduction," in *Critical Digital Studies* (Toronto: University of Toronto Press, 2013 [2008]), 3.

3 Cf. Bernard Stiegler's all-embracing use of the term in "Call for Digital Studies," paper composed in 2012, Digital Studies Network website, https://digital-studies.org/wp/call-for-digital-studies/.

universities in the Anglophone world use the label "digital studies" for programs and courses that teach computing skills in the context of culture, the arts, and the humanities.[4] The same is true of the scholarly journal *Digital Studies/Le Champ numérique*, published for the Alliance of the Digital Humanities Organizations.[5] Its French parallel title is further evocative of what Camille Roth calls "numerical humanities" and characterizes as a subfield of the digital humanities employing quantitative research methods.[6]

Moreover, the label "digital studies" could also very well include many of the other fields of research that intersect with and inform the digital human sciences: software studies, new media studies, human and computer interaction, digital cultural heritage, and—to a certain extent—media archaeology.[7] The hesitation about media archaeology has to do with how "the digital" sometimes enters and functions within the field. Whereas, for instance, studies of digital cultural heritage generally investigate the digital preservation, storage, and accessibility of art and cultural artifacts, media archaeology is often inclined to treat "the digital" as an interface to which previous (and future) media are

4 E.g., "The Digital Studies MA," Chicago University website, https://digitalstudies.uchicago.edu/about/digital-studies-ma; "Digital Studies Certificate Program," Northwestern University website, https://sps.northwestern.edu/advanced-graduate-certificate/digital-studies/; "Digital Studies in the Arts and Humanities," University of Maryland website, https://dsah.umd.edu/; "Digital Studies," University of Pittsburgh website, https://www.greensburg.pitt.edu/academics/majors-minors/digital-studies; "Digital Studies," University of Wisconsin-Madison website, https://digitalstudies.wisc.edu.

5 "About this Journal," *Digital Studies/Le Champ numérique* website, https://www.digitalstudies.org.

6 Camille Roth, "Digital, Digitized, and Numerical Humanities," *Digital Scholarship in the Humanities* 34, no. 3 (2019): 617–619.

7 See, e.g., Matthew Fuller, ed., *Software Studies: A Lexicon* (Cambridge, MA: MIT Press, 2008); Mark B. N. Hansen, *Feed-Forward: On the Future of Twenty-First-Century Media* (Chicago, IL: University of Chicago Press, 2015); Johanna Drucker, *Graphesis: Visual Forms of Knowledge Production* (Cambridge, MA: Harvard University Press, 2014); Oliver Grau, ed., *Imagery in the 21st Century* (Cambridge, MA: MIT Press, 2013); Yanni Alexander Loukissas, *All Data Are Local: Thinking Critically in a Data-Driven Society* (Cambridge, MA: MIT Press, 2019); Erkki Huhtamo and Jussi Parikka, eds., *Media Archaeology: Approaches, Applications, and Implications* (Berkeley, CA: University of California Press, 2011).

interconnected. In the first case, "the digital" is an integrated part of the object of study. In the second case, it functions as a sort of lens—often in terms of the Foucauldian "archive"—through which the object of study (digital as well as analog media) is remediated and excavated.[8] This difference is pivotal to the digital human sciences and will be further discussed below.

Digital human sciences and digital humanities

As indicated above, the field of digital human sciences connects to the by now well-established field of digital humanities. The latter is a body of research that is far-reaching, inclusive, under constant development, and admittedly hard to pin down in a few sentences.[9] By calling it a field, we follow what we take to be the point of Blackwell's *New Companion to Digital Humanities*, which originally appeared in 2004 and reappeared in a new edition in 2015. The new edition reassesses the first edition's consideration of digital humanities as a "discipline in its own right." Instead, the editors speak of a field:

> It remains debatable whether digital humanities should be regarded as a "discipline in its own right," rather than a set of related methods […]. In retrospect, it is clear that the decision this group

[8] Cf. Jussi Parikka, *What Is Media Archaeology?* (Cambridge: Polity, 2012), 113–135.

[9] For the diversity and evolving character of the field, see: Barbara Bordalejo and Roopika Risam, eds., "Introduction," in *Intersectionality in Digital Humanities* (Leeds: Arc Humanities Press, 2019), 1–8, for questions about how social identity intersects the digital humanities and challenges its traditional "big tent" metaphor; Stuart Dunn and Kristen Schuster, eds., "Research Methods in the Digital Humanities: General Introduction," in *Routledge International Handbook of Research Methods in Digital Humanities* (London: Routledge, 2020), 1–9, for the important question of how "methodology grounded in negotiation" (as distinct from reconciliation) ties together idiographic and nomothetic fields of research and thus moves beyond interdisciplinarity only among humanists; and Agiati Benardou et al., eds., "Introduction: A Critique of Digital Practices and Research Infrastructures," in *Cultural Heritage Infrastructures in Digital Humanities* (Abingdon, Oxon: Routledge, 2018), 1–14, for an argument developing a view of digital infrastructures as "ecosystems" both by their own evolving nature and by their use beyond the academy, in the cultural heritage sector at large.

> of editors, prompted by their publisher, took in naming the original *Companion* changes the way we refer to this *field*: we stopped talking about "humanities computing" and started talking about "digital humanities."[10]

We understand "discipline" to target established and by tradition and institutional infrastructures' more fixed spheres of knowledge, such as art history, comparative literature, computer and systems science, and so on, while "field" is a more flexible denomination. Fields could be held together by related objects of study, related research perspectives, or, to take the example from the quote, related methods, which all have the function of uniting scholars otherwise working in different disciplines.[11] In this respect, the digital human sciences are, just like the digital humanities, a field of research. But in two specific regards the digital human sciences depart from the digital humanities: the one concerning that which holds the field together—"a set of related methods" versus the object of study—and the other concerning the transgression of faculty borders.

First, the digital human sciences take as their object of study the ongoing digitization of culture and society, including the relations between digital entities and human beings, while the digital humanities do not necessarily construe the object of study in digital terms. Instead, the digital humanities are broadly associated with employing digital tools to deal with their objects of study and with a narrower history of computational text analysis. The latter is usually exemplified by Roberto A. Busa's (printed) *Index Thomisticus*, the result of a computational processing of Thomas

[10] Emphasis added. Susan Schriebman, Ray Siemens, and John Unsworth, eds., "Preface," in *A New Companion to Digital Humanities* (Chichester: John Wiley and Sons, 2016 [2004]). For a different account of the *Companion*'s role in the formation of the digital humanities as a field, see Matthew Kirschenbaum, "What Is Digital Humanities and What's It Doing in English Departments?" in *Debates in the Digital Humanities*, ed. Matthew K. Gold (Minneapolis, MN: University of Minnesota Press, 2012), 4–5.

[11] Cf. Janina Wildfeuer et al., eds., "Multimodality: Transdisciplinary Thoughts and the Challenge of Diversity – Introduction," in *Multimodality: Disciplinary Thoughts and the Challenge of Diversity* (Berlin: De Gruyter, 2019), 16–21.

Aquinas's writings in search for word stems that began in the 1940s. A similar endeavor is the text corpora Press-65, developed during the 1960s by the Swedish linguist Sture Allén.[12]

As for the contemporary digital humanities' emphasis on digital tools, the core tendency is to highlight the intersection of the various disciplines of the humanities and "the digital" as residing in the methodological and epistemological scholarly engagement with digital technologies and infrastructures as instruments of research. New and (often) creative ways of employing digital tools for humanist ends are hailed as reconfiguring older epistemologies, as enabling new research questions, and as paving the way for multimodal rather than predominately textual representations of knowledge.[13] Differently put, "the digital" is conceived of as standing in a more or less instrumental relation to the humanities; it is a tool of research, but not necessarily an object of research.[14] This is why distinguished digital humanities scholars such as David M. Berry and Anders Fagerjord, as well as Patrik Svensson, stress

---

12 For more information on Press-65 and Språkbanken (the language bank), hosted by Gothenburg University, see Lars Borin, "About Språkbanken," Språkbanken website, https://spraakbanken.gu.se/eng/about-us/about-spr%C3%A5kbanken. See also Koraljka Golub et al., "Digital Humanities in Sweden and Its Infrastructure: *Status Quo* and the *Sine Qua Non*," *Digital Scholarship in the Humanities* 35, no. 3 (September 2020): 547–556.

13 Cf. Katherine N. Hayles, *How We Think: Digital Media and Contemporary Technogenisis* (Chicago, IL: University of Chicago Press, 2012), 23–79; Patrik Svensson and David Theo Goldberg, eds., "Introduction" and "The Field of Digital Humanities," in *Between Humanities and the Digital* (Cambridge, MA: The MIT Press, 2015), 1–16; Patrik Svensson, *Big Digital Humanities: Imagining a Meeting Place for the Humanities and the Digital* (Ann Arbor, MI: University of Michigan Press, 2016); Stefan Gelfgren and Julia Pennlert, "En (digital) humaniora? Potential, dilemma och kritik," in *Digital humaniora – Humaniora i en digital tid*, eds. Per-Olof Erixon and Julia Pennlert (Gothenburg: Daidalos, 2017); David M. Berry and Anders Fagerjord, *Digital Humanities: Knowledge and Critique in a Digital Age* (Cambridge: Polity Press, 2017).

14 More precisely, digital research tools are often argued to indirectly both embrace and transgress mere instrumentality, exactly because of the digital humanities' recognition of their mutual effects on the objects of study and the process of knowledge production. Cf. Benardou et al, "Introduction," 3, 5.

the need for keeping a sharp focus on research questions originating in the humanities even if elaborated and inflected by the use of digital tools.[15] This point is actually a part of the field's growing ambivalence toward, and critical reconsideration of, the effects of its embracing of digital methods of investigation, collaboration, representation, and distribution of research output. One concern is that the focus on tools and technologies has resulted in weakening the role of the particular research questions of the different disciplines of the humanities. Another is that the focus on digital tools and technologies risks overemphasizing efficacy, large-scale projects, and easy dissemination, and may therefore come to serve neoliberal, administrative, or managerial agendas before scholarly aims.[16] In the present volume, both Amanda Wasielewski's and Teresa Cerratto Pargman and Cormac McGrath's chapters treat these issues in the contexts of computational methods in art history and learning analytics in higher education. In this regard, both chapters exemplify how the digital human sciences and the digital humanities overlap despite fundamental differences.

As distinct from the digital humanities, the digital human sciences may or may not systematically employ digital tools and technologies as instruments of research, or they may do so to a greater or lesser extent. From the digital human sciences point of view, instruments of research cannot define the field,[17] especially

---

[15] Berry and Fagerjord, *Digital Humanities*, 50; Svensson, *Big Digital Humanities*, 26. The same spirit informs Eileen Gardiner and Ronald G. Musto's dissociation of digital humanities from humanities computing: "Our perspective in these pages, however, is not with humanities computing, but with the digital humanities, with harnessing computing power to facilitate, improve, expand and perhaps even change the way humanists work." Computational tools are thus turned to humanistic ends—perhaps with the effect of altering them, but nonetheless working in their service. Eileen Gardiner and Ronald G. Musto, *The Digital Humanities: A Primer for Students and Scholars* (Cambridge: Cambridge University Press, 2015), 4–5.

[16] Svensson and Goldberg, "The Field," 10; Richard Grusin, "The Dark Side of the Digital Humanities: Dispatches from Two Recent MLA Conventions," *differences: A Journal of Feminist Cultural Studies* 25, no. 1 (2014): 79–92; Berry and Fagerjord, *Digital Humanities*, 7, 56–57, 141.

[17] Notwithstanding the fact that many digital humanities scholars define their field in terms of digital tools and technologies as instruments of research, the field cannot be reduced to solely a set of digital methods. It

since today, in all areas of study, it is hard to imagine scholars working completely without digital tools, systematically employed or not. That said, the scholarly use of digital tools and technologies should still be recognized in its complexity. Especially since a good deal of the tools and technologies in academic practices are somewhat "hidden," naturalized, or functioning as what Matthew Fuller and Andrew Goffey call "gray" media. The latter include media technologies normally unnoticed by human perception—such as algorithms—or so habituated that they are not always recognized as mediating devices in their own right—such as databases and workspaces.[18] Therefore, the lesson to be learned from the digital humanities' highlighting of the use of digital tools and technologies is to be suspicious of media naturalization. When the explicit employment of digital tools, technologies, and infrastructures is made salient by being under scrutiny and exploration, these discussions can by extension serve to also uncover and illuminate the infrastructures and technologies operating in the "gray" domain.

So far, we have outlined the first principal difference between the digital human sciences and the digital humanities, in terms of the former's emphasis on the present digitization of culture and society as the object of study (with digital as well as analog methods) and the latter's emphasis on digital tools. The second difference is that the field of the digital human sciences casts its net on the level of faculties and includes the humanities, the social sciences, and law, while the digital humanities traditionally include the disciplines of the humanities.[19] Berry and Fagerjord

---

obviously includes much more, such as critical theory and collaborative infrastructural projects between the academy, the art world, and the cultural sector at large. Cf. Bordalejo and Risam, "Introduction"; Benardou et al., "Introduction"; Kathryn Brown, ed., *The Routledge Companion to Digital Humanities and Art History* (New York: Routledge, 2020).

[18] Matthew Fuller and Andrew Goffey, *Evil Media* (Cambridge, MA: MIT Press, 2012), 1, 11–13. See also Loukissas, *All Data Are Local*, 55–88 for a related discussion of how digital visualization "can help us see data rather than seeing through them" (82) in the context of library collections migrating online.

[19] As in note 17, while we maintain that the predominance of the arts and humanities in digital humanities is a strong tendency in terms of tradition, the field cannot be reduced solely to this tradition. Artistic practices—as

manage to put the research activities of a wide-reaching field succinctly: The digital humanities are, "broadly speaking, the application of computation to the disciplines of the humanities."[20] The most important point is not, however, which particular disciplines are contained in the humanities or whether the social sciences or the faculty of law are included in the digital humanities traditionally conceived. Instead, the all-important point is how to conceive of issues of multidisciplinarity and its cognate, interdisciplinarity. Again, there are good reasons to learn from the experiences of the digital humanities, since they have a history of scholarly collaborations on multiple levels. This includes the basic transgression of the boundary between the humanities and systems science/engineering and, as Patrik Svensson keeps emphasizing, a "liminal" position between the various disciplines of the humanities.[21] Svensson describes this position as simultaneously a question of material infrastructure (including digital platforms for multidisciplinary journals and research organizations) and a question of intellectual infrastructure—that is, knowledge production benefiting from input from more than one discipline.[22]

Multidisciplinarity, interdisciplinarity, and transdisciplinarity

The distinction above between material and intellectual multi- or interdisciplinary infrastructures points both to their interdependence and to a pressing need for further distinctions. First, what

distinct from the study of the "arts" in arts and humanities—are occasionally included, as in Thomas Bartscherer and Roderick Coover, eds., *Switching Codes: Thinking Through Digital Technology in the Humanities and the Arts* (Chicago, IL: University of Chicago Press, 2011), as well as the social sciences and the legal domain, which is evident from Roth, "Digital, Digitized, and Numerical Humanities"; Alex H. Poole and Deborah A. Garwood, "Digging into Data Management in Public-Funded, International Research in Digital Humanities," *Journal of the Association for Information Science and Technology* 71, no. 1 (January 2020): 84–97; Ryan Whalen, ed., *Computational Legal Studies: The Promise and Challenge of Data-Driven Research* (Cheltenham: Edward Elgar Publishing, 2020).

20 Berry and Fagerjord, *Digital Humanities*, 3.

21 Svensson, *Big Digital Humanities*, 33.

22 For a broader view on research infrastructures/infrastructures for research, see Benardou et al., "Introduction," and comments in previous note 9.

is the difference between multi- and interdisciplinary approaches? Second, to qualify knowledge as multi- or interdisciplinary, is it enough that it is the product of scholars with different disciplinary backgrounds sharing the same platforms? And, third, are there slight or radical differences between knowledge resulting from a mixture of inputs from different disciplines and knowledge resulting from employing a method from another discipline?

Julie Thompson Klein's definitions of multi- and interdisciplinary collaborative and/or intellectual encounters over traditional disciplinary boundaries are helpful in qualitatively distinguishing between the different sorts of knowledge output these encounters provide. *Interdisciplinarity* concerns the "integration of information, data, methods, tools, concepts and/or theories from two or more disciplines or bodies of specialized knowledge."[23] The key issue here is the "integration" of knowledge units as fostering a "more holistic understanding of a question, topic, theme, or problem by individuals or teams."[24] Interdisciplinarity differs from *multidisciplinarity*, which concerns the "juxtaposition of separate disciplinary inputs" and "fosters breadth of knowledge and diversity of approaches."[25] Consequently, multidisciplinary approaches are additive; separate units of knowledge are assembled rather than integrated around a core question.

As a field, the digital human sciences are multidisciplinary. They tie together scholars or groups of scholars from various disciplines in the humanities, the social sciences, and law. Also, the present volume should, on an overarching level, be characterized as a multidisciplinary, collective knowledge output, as it gathers scholars from, for instance, art history, computer and systems science, law, library and information science, and media and communication studies. Within this multidisciplinary venue, many of the contributions additionally have an interdisciplinary approach to their subjects. Johan Jarlbrink's chapter integrates ethnographic perspectives and digital forensics against a backdrop

[23] Julie Thompson Klein, *Interdisciplining Digital Humanities: Boundary Work in an Emerging Field* (Ann Arbor, MI: University of Michigan Press, 2015), 15.
[24] Klein, *Interdisciplining*, 15.
[25] Klein, *Interdisciplining*, 15.

of media and communication studies. Stanley Greenstein and Cecilia Magnusson Sjöberg respectively build their cases around the integration of digital informatics and law. And Teresa Cerratto Pargman and Cormac McGrath's chapter is the result of mixing perspectives from education and computer and systems science.

Since all contributions deal with the same question—about methods and approaches to digital objects and phenomena in culture and society—there is yet another relation between disciplines to consider. In our chapters, the shared question gets repeatedly reset and revisited by different disciplinary or interdisciplinary frameworks. Hence, the question is *transdisciplinary* in character. The latter differs both from interdisciplinarity and multidisciplinarity in, still according to Thompson Klein, designating "an overarching synthesis associated with new conceptual frameworks."[26] In this holistic sense, "the digital" in the digital humanities would be the transdisciplinary object *per se*, as the common denominator of the field. We, however, understand this sense of transdisciplinarity to be too reductive. Our ambition is rather to acknowledge the various reconfigurations that occur in (inter)disciplinary studies of a shared problem, concept, or other object of research. Our sense of transdisciplinarity thus comes closer to Mieke Bal's theory of conceptual travels, which emphasizes "negotiation," "transformation," and "reassessment" on each stage of a concept's trajectory through different (inter) disciplinary frameworks.[27]

All in all, our shared question is transdisciplinary in its travel through the different chapters of this multidisciplinary volume. In each chapter, the question is subjected to an (inter)disciplinary framework that treats it differently and answers it differently.

Outline of the Volume

The volume is structured in three parts, following the chapters' approaches to the question of how the present digitization of culture and society is submitted to, as well as giving rise to,

[26] Klein, *Interdisciplining*, 20.

[27] Mieke Bal, *Travelling Concepts in the Humanities: A Rough Guide* (Toronto: University of Toronto Press, 2002), 39.

methodological and epistemological strategies and constraints. In the first part, the chapters turn toward *consequences of method* or the types of knowledge and results that emanate from different research practices. In the second part, the chapters have in common their approach to *method as being an object of study in itself*, in discussions that critically assess how personal data processing methods are employed and functioning in digital society. In the third part, the chapters are concerned with *demonstrating method* in examinations of digital objects.

Epistemologies of objects, methods, and research practices

Part one comprises five chapters that examine the different ways in which digital objects, analog as well as digital methods of investigation, and other research practices, condition and inflect knowledge outputs.

First out, Jonas Andersson Schwarz discusses digitally mediated and socially networked texts, exemplified by Twitter conversations. Andersson Schwarz's chapter offers the concept of "social big data" as a shorthand for the complex nature of an information entity ("data") that is semiotic as well as material, and socially situated on the individual as well as on the institutional level, besides being retrievable in enormous quantities ("big"). Another of Andersson Schwarz's points is that such data is not only socially mediated by the producer but also conditioned by the researcher's situationally embedded knowledge. This recognition of the complex nature of "social big data" forms the basis of the author's discussion of traditional criteria for information validation. The following chapter, by Jonas Stier, shares one of Andersson Schwarz's objectives, namely to problematize the idea that data and, in Stier's case, digital technologies are somehow clean, directly representational, and thus "objective." More precisely, Stier's chapter examines "inherent social and cultural biases" in educational and political discourses on digitization and digital innovations. Stier presents a typology of "discursive blind spots," including technocentrism and normativism, homo- and heterocentrism, and ego- and ethnocentrism. The leading argument is that such biases and lacunae can be overcome by combining a set of integrative approaches or conducting research along interdisciplinary as well as intercultural and "intermethodological" lines.

Next follow two chapters that deal with visualization and visual terminology. The chapter by Julia Pennlert, Björn Ekström, and David Gunnarsson Lorentzen introduces a set of visual, or "teleoptical," metaphors for understanding and analyzing processes of computer-assisted readings. The chapter builds on three case studies of Instagram posts, Twitter conversations, and online scholarship in order to examine how metaphorical thinking inflects the research process. The picture suggested is that of "lenses" and "filters" between, on the one hand, the researcher and the instrument and, on the other, between the instrument and the object of study. In other words, the overall idea is to analyze how conceptual tools reciprocally shape the interfaces of digital objects and instruments, and intervene with the process of knowledge production. Approaching the visual sphere from another end, Karolina Uggla's chapter reviews the analytical terminology in the field of information visualization. The object of study is thus the words that name the process of turning data into images such as charts and graphs. "Infoviz" is both quantitatively and qualitatively permeating digital society. If the first concerns the sheer number of visualizations in such diverse domains as newsfeed and scholarship (the latter illustrated by Figures 3–6 in the preceding chapter), the second concerns the conceptual legacy of theorizing "information" or "data" as visual translations of numerical units. Uggla demonstrates how the field is broadly divided in subfields based on semiotic and "designerly" approaches that models the "same" visual objects somewhat differently. Such differences in how the visual object is conceptually modeled by terminological choices is particularly of note in Uggla's two "interpretative frameworks," dealing with social semiotics, and emotions and ethics, respectively.

Closing the first part, Amanda Wasielewski's chapter targets and questions the reluctance in the discipline of art history to employ digital methods. Wasielewski recognizes two factors as especially challenging. First, that the objects central to art history are images and that the methods for digital image analysis are more complicated than the methods for computational text analysis. Second, that there is today a lack of preexisting art historical data to study with digital methods (which may not be the case in the future). If these are the challenges Wasielewski identifies, her chapter also provides examples of studying “Internet art” and artists’ digital traces in methodological frameworks that eschew simplifications such as digital methods on the one hand and traditional ones on the other: “it is useful to think of methodologies as techniques that can be combined and mixed together rather than camps ready to wage war on one another.”

Legal and ethical dilemmas

The second part gathers three chapters that deal with legal and ethical issues arising from the digital handling and processing of personal data in public and private administration and in higher education.

The first two chapters, by Stanley Greenstein and Cecilia Magnusson Sjöberg, both reject traditional legal approaches in favor of “proactive” (as distinct from reactive) law. This perspective, the authors argue, is better suited to dealing with the increasing use of digital technology in the contexts of, for instance, social security, social insurance, and other public administration. Greenstein especially points out that, since AI technology is presently put to a multiplicity of societal uses, it needs to be addressed broadly as a practice encompassing both the risk of manipulating human behavior and the legal function of protecting individuals. One of Greenstein’s points is to learn from the cognitive sciences in order to properly analyze potential risks with AI. Together with perspectives from legal informatics, the insights from the cognitive sciences provide the author’s basis for reassessing the “traditional legal science approach.” Magnusson Sjöberg’s chapter focuses on issues of transparency in AI-based personal data processing, and highlights the need to have already implemented such considerations in the planning stages of systems design. Magnusson Sjöberg proceeds from making a distinction between her key term transparency and “openness,” and explains the distinction’s importance in the case of, for example, access to public sector information: public institutions can be governed by principles of openness but nevertheless fail to provide transparency due to insufficient access rights, or, if the legal rights are in place, insufficient implementation of them—hence the need to focus on legal proactivity.

Teresa Cerratto Pargman and Cormac McGrath's chapter highlights how higher education has more and more come to rely on big data in general and, in particular, on the methods and procedures known as learning analytics. The latter comprises a set of practices that relies on big data analysis in order to optimize and evaluate student learning and institutional management of learning environments. Based on a review of the scholarly literature within the field of learning analytics, the chapter examines the values and expectations attached to the use of big data in educational settings as well as its ethical implications. Above all, Cerratto Pargman and McGrath provide a discussion of the context-sensitivity of data-driven practices, the technical and ethical competences demanded to sustain them, and the risk of mistaking "data-driven" for "evidence-based."

New objects—new approaches

The third and last part comprises three chapters that demonstrate different approaches to new media objects: personal computers in the first chapter, YouTube shows in the second, and digital scholarship—or, rather, analog, printed, scholarship disseminated in digital form—in the third. In line with the last point, Johan Jarlbrink's and Christer Johansson's first two chapters also demonstrate the need for distinctions within the category of "new media" itself. Jarlbrink shows both how the life of a personal computer includes stages of "new" and "old," and how new hardware can very well include old software, whereas Johansson investigates how certain YouTube shows are both aligning themselves to and reforming old media such as radio and television.

More specifically, Jarlbrink's chapter examines personal computer biographies in order to trace not the content produced and saved by the previous owners but the micro history of log files, updates, and viruses that constitute an archive of the computers' own histories. Thereby, the author demonstrates the methods of digital forensics and media ethnography as the combinatory approach by which computer histories can be traced and contextualized. Of particular concern is the relation between micro and macro history. The forensic mapping of the time and space of computer use reveals, for instance, which owner saved most content on the space of the hard drive and how this changed over time. Jarlbrink points out that such results can, in the macro context of cultural heritage institutions, broadly serve to illuminate media use in the digital age.

Johansson examines the contemporary phenomenon of dialogue podcasts streaming live on YouTube, or, put differently, auditory media in combination with visual media. The chief examples are the politically controversial shows *The Joe Rogan Experience* and its Swedish counterpart *Hur kan vi?* Both are the targets of a far-reaching investigation of their medial nature and sociopolitical implications or questions that revolve around representation and communication. Method-wise, Johansson elaborates a tripartite approach that interconnects a semiotically based intermedial analysis, examining interactions between different types of modalities and signs, with a media-historical analysis that puts the YouTube shows in relation to the "orality" of earlier media. Lastly, the author intertwines these two threads with a communicational-theoretical analysis that attends to the digital platform as an infrastructure channeling speech into sociopolitical activity.

In the last chapter, Amanda Wasielewski and Anna Dahlgren develop a procedure for text mining art historical journals that displays problems of nonstandard formatting, notes instead of bibliographies, and layouts that change over time. In detailed steps, the authors provide a hands-on demonstration of the advantages and pitfalls of particular text mining tools. The authors make clear that the big issue in text mining a corpus of academic articles is how it can reassess historiographical presuppositions about, for instance, patterns of (supposedly) paradigmatic discourses. One way is to account for frequency of citations, which, the authors emphasize, requires both quantitative (total amount) and qualitative (distribution across different articles) considerations. Although the authors' case is confined to art historical journals, it is well illustrative of both practical problems and historiographical insights the procedure of text mining academic articles can encounter and benefit from in other branches of the humanities and the social sciences.

References

Bal, Mieke. *Travelling Concepts in the Humanities: A Rough Guide*. Toronto: University of Toronto Press, 2002.

Bartscherer, Thomas, and Roderick Coover, eds. *Switching Codes: Thinking Through Digital Technology in the Humanities and the Arts*. Chicago, IL: University of Chicago Press, 2011.

Benardou, Agiati, Erik Champion, Costis Dallas, and Lorna M. Hughes, eds. "Introduction: A Critique of Digital Practices and Research Infrastructures." In *Cultural Heritage Infrastructures in Digital Humanities*, 1–14. Abingdon: Routledge, 2018.

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

Bordalejo, Barbara, and Roopika Risam, eds. "Introduction." In *Intersectionality in Digital Humanities*, 1–8. Leeds: Arc Humanities Press, 2019.

Borin, Lars. "About Språkbanken." Språkbanken website. https://spraakbanken.gu.se/eng/about-us/about-spr%C3%A5kbanken.

Brown, Kathryn, ed. *The Routledge Companion to Digital Humanities and Art History*. New York: Routledge, 2020.

Chicago University website. "The Digital Studies MA." https://digitalstudies.uchicago.edu/about/digital-studies-ma.

Digital Human Sciences website. https://dhv.blogs.dsv.su.se.

Digital Studies/Le Champ numérique website. "About this Journal." https://www.digitalstudies.org.

Drucker, Johanna. *Graphesis: Visual Forms of Knowledge Production*. Cambridge, MA: Harvard University Press, 2014.

Dunn, Stuart, and Kristen Schuster, eds. "Research Methods in the Digital Humanities: General Introduction." In *Routledge International Handbook of Research Methods in Digital Humanities*, 1–9. London: Routledge, 2020.

Fuller, Matthew, ed. *Software Studies: A Lexicon*. Cambridge, MA: MIT Press, 2008.

Fuller, Matthew, and Andrew Goffey. *Evil Media*. Cambridge, MA: MIT Press, 2012.

Gardiner, Eileen, and Ronald G. Musto. *The Digital Humanities: A Primer for Students and Scholars*. Cambridge: Cambridge University Press, 2015.

Gelfgren, Stefan, and Julia Pennlert. "En (digital) humaniora? Potential, dilemma och kritik." In *Digital humaniora – Humaniora i en digital tid*, edited by Per-Olof Erixon and Julia Pennlert, 11–26. Gothenburg: Daidalos, 2017.

Golub, Koraljka, Elisabet Göransson, Anna Foka, and Isto Huvila. "Digital Humanities in Sweden and Its Infrastructure: *Status Quo* and the *Sine Qua Non*." *Digital Scholarship in the Humanities* 35, no. 3 (September 2020): 547–556.

Grau, Oliver, ed. *Imagery in the 21st Century*. Cambridge, MA: MIT Press, 2013.

Grusin, Richard. "The Dark Side of the Digital Humanities: Dispatches from Two Recent MLA Conventions." In *differences: A Journal of Feminist Cultural Studies* 25, no. 1 (2014): 79–92.

Hansen, Mark B. N. *Feed-Forward: On the Future of Twenty-First-Century Media*. Chicago, IL: University of Chicago Press, 2015.

Hayles, Katherine N. *How We Think: Digital Media and Contemporary Technogenisis*. Chicago, IL: University of Chicago Press, 2012.

Huhtamo, Erkki, and Jussi Parikka, eds. *Media Archaeology: Approaches, Applications, and Implications*. Berkeley, CA: University of California Press, 2011.

Kirschenbaum, Matthew. "What Is Digital Humanities and What's It Doing in English Departments?" In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 3–11. Minneapolis, MN: University of Minnesota Press, 2012.

Klein, Julie Thompson. *Interdisciplining Digital Humanities: Boundary Work in an Emerging Field*. Ann Arbor, MI: University of Michigan Press, 2015.

Kroker, Arthur, and Marilouise Kroker, eds. "Introduction." In *Critical Digital Studies*, 3–35. Toronto: University of Toronto Press, 2013 [2008].

Loukissas, Yanni Alexander. *All Data Are Local: Thinking Critically in a Data-Driven Society*. Cambridge, MA: MIT Press, 2019.

Northwestern University website. "Digital Studies Certificate Program." https://sps.northwestern.edu/advanced-graduate-certificate/digital-studies.

Parikka, Jussi. *What Is Media Archaeology?* Cambridge: Polity, 2012.

Poole, Alex H., and Deborah A. Garwood. "Digging into Data Management in Public-Funded, International Research in Digital Humanities." *Journal of the Association for Information Science and Technology* 71, no. 1 (January 2020): 84–97.

Roth, Camille. "Digital, Digitized, and Numerical Humanities." *Digital Scholarship in the Humanities* 34, no. 3 (2019): 617–619.

Schriebman, Susan, Ray Siemens, and John Unsworth, eds. "Preface." In *A New Companion to Digital Humanities*. Chichester: John Wiley and Sons, 2016 [2004].

Stiegler, Bernard. "Call for Digital Studies." Paper composed in 2012. Digital Studies Network website. https://digital-studies.org/wp/call-for-digital-studies.

Svensson, Patrik. *Big Digital Humanities: Imagining a Meeting Place for the Humanities and the Digital*. Ann Arbor, MI: University of Michigan Press, 2016.

Svensson, Patrik, and David Theo Goldberg, eds. "Introduction" and "The Field of Digital Humanities." In *Between Humanities and the Digital*, 1–16. Cambridge, MA: MIT Press, 2015.

University of Maryland website. "Digital Studies in the Arts and Humanities." https://dsah.umd.edu.

University of Pittsburgh website. "Digital Studies." https://www.greensburg.pitt.edu/academics/majors-minors/digital-studies.

University of Wisconsin-Madison website. "Digital Studies." https://digitalstudies.wisc.edu.

Whalen, Ryan, ed. *Computational Legal Studies: The Promise and Challenge of Data-Driven Research*. Cheltenham: Edward Elgar Publishing, 2020.

Wildfeuer, Janina, Jana Pflaeging, John A. Bateman, Ognyan Seizov, and Chiao-I Tseng, eds. "Multimodality: Transdisciplinary Thoughts and the Challenge of Diversity – Introduction." In *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*, 3–38. Berlin: De Gruyter, 2019.

PART ONE: EPISTEMOLOGIES OF OBJECTS, METHODS, AND RESEARCH PRACTICES

Not a Mirror, but an Engine: Digital Methods for Contextual Analysis of "Social Big Data"

Jonas Andersson Schwarz

Introduction

Unlike several of the subsequent chapters in this anthology, I will not focus on audiovisual media, nor new media objects such as Instagram posts (Pennlert et al.), podcasts (Johansson), data visualizations (Uggla) or web art (Wasielewski) for that matter.[1] The medium that I will write about is a much more familiar one for humanists and social scientists: plaintext.

Like Wasielewski, I too would note that the methods for computational text analysis are in many ways less complicated than those for digital image analysis, for example, and that there are several already-established methodological pathways for text-based corpus archiving, distribution, retrieval, and analysis. But the research projects that I will refer to are very novel, however, in that their research objects are what other authors in this volume refer to as "born digital": The initial publication and distribution of the text data in question was made by native internet users in online social media forums. Moreover, as soon as the researcher uses computers to access and select samples from such forums, and then computationally process and analyze his or her findings, we are dealing with methods that are in many ways *very* different from predigital methods for text analysis. In other words, relating

[1] For all of these authors, see this volume.

How to cite this book chapter:
Andersson Schwarz, Jonas. "Not a Mirror, but an Engine: Digital Methods for Contextual Analysis of 'Social Big Data.'" In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 23–47. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.b.

back to the editors' introduction,[2] my own interest in "the digital" is equally *as a tool* and *as an object* of research.

As a scholar of media and communications, my own frame of analysis borrows from traditional sociology and public opinion research, where the question of *representativity* is always of high interest, and medium theory, where the question of the *ontological properties* of particular media forms is similarly important—in this case, the properties of online, user-generated and user-distributed text. I am interested in using digital corpora to be able to say something about "the public," but I am equally interested in noting what possibilities and limitations are afforded by online-mediated social discourse—an artifact that is digital from its very inception, as users type their affective expressions in interfaces that come with certain preconditions to begin with.

Initial Definitions

This chapter aims to discuss the epistemology and normativity of data produced and shared on social media platforms, and the attendant challenges for research: problems of access and representativity, and the need for contextualization and, consequentially, for combinatory methods. My work is situated at the intersection of social science and humanities, in the sense that I have experience from engaging in highly empirical endeavors aiming to capture digital and communicational mechanisms of contemporary society, while at the same time agonizing over the issues of social philosophy that arise from such endeavors, including the abovementioned epistemic and normative tangles.[3] Lastly, I will close the chapter by relating back to the initial empirical and epistemological challenges, pointing out some key pitfalls concerning the validity, reliability, and representativity of such data, and some steps toward improving contextualization, especially regarding semantic text data.

2 This volume.

3 Due to brevity, I will not focus legal aspects, for example those that address consent or copyright, although these are highly important aspects for the generation and collection of this type of data. It is crucial to maintain an acquaintance with the various administrative and governmental applications of data analysis, since many are both politically and commercially expedient.

What do I mean by *epistemology*? Reflecting upon two recent research projects that I have been part of,[4] this chapter is a meditation on the multifaceted methodological and epistemological challenges that mount when researchers face contemporary social media platform architectures as research objects and data sources. Like other critical theorists of the so-called "big data era," I would not declare that the new data landscape heralds an "end of theory." On the contrary, one needs to contemplate the epistemological implications of the very "data revolution" at hand.[5] While new approaches to data generation, collection, and analyses are enabled that make it possible to ask and answer questions in new ways, it is clear that much of the usefulness of relational, real-time datasets has to be complemented by more established, conventional research methods. Dominique Boullier has shown how the novel methods that contemporary data structures enable—for example, tracing trends and inferring relational patterns in (near) real time—often need to be put in context,[6] for example by being complemented by more established methods of assessing representativity (late 20th-century quantitative social science) and validity (qualitative assessment). Moreover, methods and instruments tend to be suffused with epistemological implications, and so too are the research sites and objects to begin with. Data always has prerequisites for its generation in the first place, even putting to one side the politics of access to it: its alleged validity, reliability, salience, and indeed importance; the various idealistic claims that are made for it; and, lastly, the naturalization throughout key societal sectors (academic, administrative, business, policy) of such viewpoints. Data tends to have a normative function, in that it represents the social relations that it registers as valid in and of themselves. If we are to make a comprehensive overview of

4 Jonas Andersson Schwarz et al., *Opinioner och offentligheter online* (Stockholm: IIS, 2015); Stefan Dahlberg, *Linguistic Explorations of Society* (Swedish Research Council, 2017).

5 Rob Kitchin, "Big Data, New Epistemologies and Paradigm Shifts," *Big Data & Society* 1, no. 2 (April–June 2014): 1.

6 Dominique Boullier, "Big Data Challenges for the Social Sciences and Market Research: From Society and Opinion to Replications," in *Digitalizing Consumption: Tracing How Devices Shape Consumer Culture*, eds. Franck Cochoy et al. (London, New York: Routledge, 2017).

challenges, we find that they are indeed philosophical, psychological, cybernetic, *and* political-economic. Formalized methods of data analysis tend to inductively probe data that is abundant with internal relations in order to identify latent structures and patterns in such data. I will take natural language processing (NLP) as an example, but there are many more machine learning (ML) approaches, and methods for formalized social network analysis (SNA). Researchers could utilize combinations of these novel methods, in addition to more traditional content analysis (CA) where samples are taken and manually decoded.

What is *data* in this particular context? The research object I have in mind is *user-generated plaintext generated on web-based internet platforms*.[7] As data sources, large volumes of plaintext are suitable for computer-aided digital human sciences tools such as "information retrieval, text analytics, data mining, visualization, and geographic information systems."[8] More specific methods—like the NLP that I will discuss in this chapter—lend themselves to *large quantities* of plaintext, where corpus-based, statistical approaches can be employed. Here, the general item of analysis would be morphemes (minimal meaningful word segments), where NLP can, for example, successfully parse the syntactical position of individual morphemes, given their position relative to other morphemes in the corpus.[9] More specifically, the models hinge upon the so-called *type-token distinction*.[10] There are functional differences between identifying a *class* (type) of objects and naming the individual *instances* (tokens) of that class, and computers are used to statistically identify such types and tokens.

---

7 These could be highly publicly known platforms like Twitter, but what is more specifically implied here are message boards and pages on the common web, since these are more easily accessible for scraping than many proprietary platforms, such as those owned by Facebook.

8 Michael Piotrowski, *Natural Language Processing for Historical Texts* (London, Williston, VT: Morgan & Claypool, 2012), 6.

9 Lenhart Schubert, "Computational Linguistics," in *The Stanford Encyclopedia of Philosophy* (Spring 2020 ed.), ed. Edward N. Zalta (Stanford, CA: Stanford University, 2020).

10 Linda Wetzel, "Types and Tokens," in *The Stanford Encyclopedia of Philosophy* (Fall 2018 ed.), ed. Edward N. Zalta (Stanford, CA: Stanford University, 2018).

Tokenization is the task of chopping up a document into constituent pieces.[11] By using so-called *fuzzy search* heuristics, misspelled and ambiguous tokens can also be typified together with native, unambiguous tokens. In exploratory data analysis, token occurrences in an unknown language can be explored, in order to improve language models by creating particular “watermarks” of the ways in which word embeddings are distributed in different languages. *Word embeddings* is a way of statistically representing the ways in which words have similar meanings, given their contexts in specific languages. Linguist John Firth expressed it in an often-invoked quip: “You shall know a word by the company it keeps!”[12] As Piotrowski has pointed out, computational linguistics should be considered an “auxiliary science” to digital human sciences, which can aid researchers with *formal modeling* of scholarly knowledge and insights in machine-processable form.[13] There are numerous various approaches in this rapidly developing field, and the practical applications are many: machine translation, document retrieval, knowledge extraction (by way, e.g., of recognition of patterns and/or clusters), sentiment analysis, natural language user interfaces, and so on.

More specifically, what do I mean by *social* data? For the purposes of this chapter, I will focus on the societal contingencies of found web-mediated texts, as those texts are produced by social agents, for specific purposes, and mediated by specific technologies with specific affordances. Despite their profusion, the texts collected should not be expected to mirror the social world in a 1:1 fashion. Rather, they are produced under specific circumstances, having specific properties. Hence, I argue that the mathematical rigor and competence of the NLP scientist have to be combined with a sociological sensibility in order not to adjust the inferences

[11] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *An Introduction to Information Retrieval* (Online ed.) (Cambridge: Cambridge University Press, 2009), 22.

[12] John Firth, “A Synopsis of Linguistic Theory 1930–1955,” in *Studies in Linguistic Analysis 1930–1955*, ed. John Firth (Oxford: Basil Blackwell, 1962), 11.

[13] Michael Piotrowski, “Digital Humanities, Computational Linguistics, and Natural Language Processing,” presentation, *Lectures on Language Technology and History*, Uppsala University (March 4, 2016), 21.

made and put the texts into societal context. As *research objects*, digital media platforms are "moving targets" in the sense that the sociocultural enactments taking place are ever-changing, making replication very hard. If researchers are to allow future researchers to replicate their analyses, they either have to duplicate the data used and be able to hand this to adjacent researchers in some way or form, or provide clear step-by-step heuristics as regards the data collection,[14] with the proviso that, despite following these very same steps, future data harvests might look entirely different. Moreover, while the platforms in question could be seen as "material substrates" to the sociocultural enactments taking place "on top of them," as it were, it is important to understand them, more aptly, as *enablers of agency* that structure, delimit, and harness social action. The platforms are, in other words, *firm ground* that enable and record social action and, at the same time, *structuring agents* that contain certain affordances and, thus, orchestrate user agency. It is advisable to seek out literature on political-economic and material platform architectures and business models, since this literature can help the analyst identify features that might be critical for understanding what users can and cannot do ("hard" governance; code-as-law), and the various ways in which they are expected to act ("soft" governance; user norms).

What is the *bigness* of "big data"? As I have argued elsewhere,[15] the big data signifier should be understood epistemologically rather than out of mere quantitative assessment. The datasets involved *need* not be terabyte-sized; they can actually be much smaller, yet still fit under the header of big data methodology, if the methods employed address and make available the *multidimensional relationality* of the data involved. In SNA, for example, the set of social relations is always *at least* two-dimensional[16] and is usually represented in the form of a *node table* (listing all the nodes

[14] There are, e.g., open science projects like CommonCrawl that routinely scrape the publicly available web, making available enormous quantities of multilingual corpus text.

[15] Göran Bolin and Jonas Andersson Schwarz, "Heuristics of the Algorithm: Big Data, User Interpretation and Institutional Translation," *Big Data & Society* 2, no. 2 (December 2015): 1–12.

[16] John Scott, "Networks & Relations," in *Social Network Analysis*, ed. John Scott (London: Sage, 2013), 1–9.

involved) and an *edge table* (listing the relations between them). The operational term in this chapter will therefore be "social big data"—as a term for the forms of data employed *and* its attendant methods and approaches. Arguably, in this case, the two are hard to separate; the data generated becomes generated through analytical operations. Method and data co-constitute each other, as it were. Ultimately, there is an epistemological case to be made for seeing "data" as not only a research object alone but also a *research tool*. I would like to stress the *boundedness* and *ordinariness* of algorithmic infrastructures—that is, their capacity as bureaucratic systems, and co-constitutive of a resurgence of the technocratic "administered society"[17] as a governmental trope. I argue that methodological reductionism should be understood as a prime cause for this resurgent technocracy. When institutional actors understand social reality through a lens of "data idealism," the resultant policies are likely to be mechanistic in nature, as many other scholars have pointed out.[18] In order not to simplify the world views on which important decisions are taken, empirical and theoretical rigor begs a constructionist understanding of social big data. As Evgeny Morozov once pointed out, data-driven companies like Google operate less as mirrors of social reality and more as engines of social change.[19]

Social Big Data: Perspectives

Given the material-semiotic concerns noted above, I will now list a range of particular contexts that are constituent of the generation of social big data. I will begin by noting the important distinction between numerical (quantitative) data, often accrued through binary sensors, and semantic (qualitative) data, comprised of meaning-bearing units.

First, from a *semiotic* perspective, the theory of Charles Sanders Peirce would imply that what most digital sensor signals do is

---

17 Herbert Marcuse, *One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society* (Boston, MA: Beacon Press, 1964), 9–20.

18 Notable names are Mireille Hildebrandt, Antoinette Rouvroy, Erich Hörl, Nick Couldry, Evgeny Morozov, John Cheney-Lippold, Ted Striphas.

19 Evgeny Morozov, *To Save Everything, Click Here: The Folly of Technological Solutionism* (New York: PublicAffairs, 2013), 145.

to *index* physical reality—as smoke to a fire, a pedometer tick would index a walking step. Similarly, a ticker would count binary "likes" in social media.[20] While this might appear straightforward, indications are in actual real life rarely that straightforward or clear. In practice, a single sensor can indicate many things. This suggests that one should consider data epistemology from the ground up: *signals* are transformed into *data*, which in turn can signify *information*, which in turn requires *interpretation* so as to be generative of *knowledge*. There are many links in this chain, where degrees of arbitrariness might seep in; it is rarely a foolproof chain of evidence. In other words, as even binary sensor data might contain significant ambiguity, it is easy to see how semantic (in the Peircian parlance, we would call it *symbolic*) data is likely to raise even more complex concerns regarding context complexity.

Second, from a *material* perspective, the way interfaces are constituted is also of vital importance. Different interfaces and design choices make for different affordances,[21] including so-called "dark patterns" deliberately designed to steer the user in various directions. The generation of data, as an ontological and epistemological object, is also directly resultant from such interface design choices. As Lev Manovich has argued, data (archives) and interfaces (algorithms) are co-constitutive of each other: "The two goals of *information access* and *psychological engagement* compete within the same new media object."[22] The means of interacting with a digital archive always takes place through an algorithmic interface, and is thus determined by algorithms. Vice versa: what ends up in a database is an outcome of the interface's ability to record its surrounding social world.

Third, *situational/phenomenological* perspectives are deeply entangled in interpretation of data. Pennlert et al.[23] use

[20] Dawn Nafus, *Quantified: Biosensing Technologies in Everyday Life* (Cambridge, MA: MIT Press, 2016), xx.

[21] James J. Gibson, *The Ecological Approach to Visual Perception* (New York, London: Psychology Press, 1979).

[22] Lev Manovich, *The Language of New Media* (Cambridge, MA: MIT Press, 2001), 216, emphasis added.

[23] This volume.

optical metaphors in order to account for methodological operations—"cuts" or "folds" through the corpus, as it were—and similar optical metaphors are also indispensable for understanding relational arrangements of users of interfaces versus databases being examined through interfaces. As Lorraine Daston and Peter Galison have shown, the very notion of objectivity is deeply intertwined with vision. Interestingly, the old meaning of "'objective' referred to things as they are presented to consciousness, whereas 'subjective' referred to things in themselves."[24] That is, things would have "'objective reality' […] by virtue of their clarity and distinctness, regardless of whether they [would] exist in material form."[25] Premoderns thought of things as objective merely by being "objects of the mind and not of the world," so to speak. This, however, is not too far removed from the modern notion of what Walter Lippmann coined "phantom publics." "Public opinion" is something that can never be physically experienced; it only exists as a figment of the imagination. But, to a modern human, the *objectivity* of a poll would be a matter not of its abstractness but of its veracity, trustworthiness, or probability, and the decisive factor would be how statistically representative the underlying samples or polls would be. However, there is a blind spot here: the tendency to think of population samples in terms of representativity means that modern subjects constantly risk making the mistake of seeing the observed sample as somehow representative of the whole. This tendency can of course be exploited; I leave it to the reader to think of such examples.

Consequently, the complex production of data on social platforms cannot be examined without methodologically taking into account several converging factors. If we continue to employ optical metaphors, we find that the "ways of seeing" into digital infrastructure is premised on some rather different mechanics compared to the ways in which physical social reality allows itself to be observed.

For example, digital archives do not lend the observer any vantage point from which (s)he can attain a full overview of it as a

24 Lorraine Daston and Peter Galison, *Objectivity* (New York: Zone Books, 2007), 29.

25 Daston and Galison, *Objectivity*, 29.

"social totality." Observing a town square or, say, a prison yard at a distance, one can (at least theoretically) gain a *panoptic* overview.[26] With digital infrastructures, the mode of vision is instead near-sighted, *oligoptic*,[27] more akin to traversing a sequestered space, say a house or a labyrinth, where one only sees the particular room or corridor that is at hand, never the whole structure. To continue this optical metaphor, the means of access can conveniently be understood through computer science parlance of *back-end* versus *front-end* architectures,[28] the front end usually being premised on psychological engagement (personalized feeds, little or no overview) and the back end more primarily on information access (access to database, queries with aspirations to be objectivizing). The former precipitates manual local reading, continuously or in (near) real time, while the latter enables automated distant reading and systematic, retroactive overview. The former entails a risk of context collapse or misattribution, while the latter entails a risk of context loss.[29]

Moreover, in online milieus, especially those characterized by advertising and various personalization algorithms,[30] users tend to be enticed to make interactions that will in turn be farmed into interesting content for other users to interact with. Crucially, users will be shown content and advertisements; the platform operators expect to make these so that the user finds them interesting. Arguably, users adapt their behaviors in order to suit the algorithm: in order to maintain peer visibility, users update their own posts according to what the algorithm deems

[26] Michel Foucault, "Panopticism," in *Discipline & Punish: The Birth of the Prison*, trans. Alan Sheridan (London: Allen Lane, 1977).

[27] Bruno Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory* (Oxford: Clarendon, 2005).

[28] Metaphors that could, in turn, be related to Erving Goffman's spatial metaphors—"front stage" and "back stage"—which were applied to broadcast and telecommunications media by Joshua Meyrowitz.

[29] Jonas Andersson Schwarz and Johan Hammarlund, "Kontextförlust och kontextkollaps: Metodproblem vid innehållsanalys av sociala medier," *Nordicom-Information* 38, no. 3 (2016): 41–55.

[30] James Webster, *The Marketplace of Attention: How Audiences Take Shape in a Digital Age* (Cambridge, MA: MIT Press, 2014), 88–89.

popular or recognizable to a large audience.[31] There is a kind of built-in conformism to this, a popularity bias.[32] Individuals seem to follow trends based on machine-based predictions and thereby coproduce popular culture by delegating what content should be distributed and when.

Lastly, online accumulations rarely have *Gaussian distributions* (normal curve); they tend to have *Pareto distributions* (skewed or "long tail" curve). Early observations by Albert-László Barabási[33] showed that, although anyone can publish content on an internet website, there is no guarantee that the content will be read by any major audience or at all. This insight was expanded upon by Yochai Benkler[34] and, later, Matthew Hindman.[35] Ten years later, Hindman expanded his critique of the systemic obstacles to growing digital audiences; he attributed it to a combination of minuscule design factors and sheer mathematical odds.[36] He showed how the so-called "long-tail" distributions online[37] are too steep for small sites to add up to substantial collective market share. Such distributions, found everywhere in networked spaces online, result in "starkly inegalitarian outcomes."[38] This means, as we will see below, that conventional tests of sample representativity—which generally require that the total distributions, from which random samples are made, are following Gaussian

[31] Tarleton Gillespie, "The Relevance of Algorithms," in *Media Technologies: Essays on Communication, Materiality, and Society*, ed. Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot (Cambridge, MA: MIT Press, 2014), 183–188.

[32] Webster, *The Marketplace*, 89–91.

[33] Albert-László Barabási, *Linked: How Everything Is Connected to Everything Else and What It Means* (New York: Perseus, 2002).

[34] Yochai Benkler, *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (New Haven, CT: Yale University Press, 2006).

[35] Matthew Hindman, *The Myth of Digital Democracy* (Princeton, NJ: Princeton University Press, 2009).

[36] Matthew Hindman, *The Internet Trap: How the Digital Economy Builds Monopolies and Undermines Democracy* (Princeton, NJ: Princeton University Press, 2018).

[37] More correctly, these should be labelled "lognormal," "power-law," or Pareto distributions.

[38] Hindman, *Internet Trap*, 41.

distributions—are very hard, or even impossible for these types of online corpora.

For all of the above factors, actual understanding of the media and constitutive milieus in question is premised on tacit experiential knowledge of these media and milieus in question. Researchers operate with situational knowledge—much like archivists, librarians, and historians operate with implicit assumptions about the calligraphic- or type-based universes that they traverse.[39] In other words, the particular types of visibility, homosociality, and accumulation engendered online are hard to fully appreciate without actual experience of the applications and sites in question. Ultimately, for researchers of corpuses of socially mediated text, it is indispensable to have an awareness of the situational contexts for the generation of the data in the first place, before one even begins to ask formal questions of provenance, representability, and so on.[40]

Social Big Data: Contexts

Generating observational knowledge about social media platforms requires taking these optical and material-semiotic concerns into account. But it does not stop here; to conclude this section, we will have to consider a range of further structural and global contexts in addition to the concerns listed above.

Microstructural context is crucial, as local network topologies and communities make for a range of sociological factors to consider. Primarily, one should heed the analytically useful differentiation between the "subjective" surrounding that the user finds herself immersed in (phenomenological *Umwelt*, or *constitutive milieu*),[41] and the more "objective" notion of a surrounding, general ecology (*Umgebung*), as it would constitute itself for an external

39 Marshall McLuhan, *The Gutenberg Galaxy: The Making of Typographic Man* (Toronto: University of Toronto Press, 1962).

40 Derek Ruths and Jürgen Pfeffer, "Social Media for Large Studies of Behavior," *Science* 346, no. 6213 (2014): 1063–1064.

41 Jonas Andersson Schwarz, "Umwelt and Individuation: Digital Signals and Technical Being," in *Digital Existence*, ed. Amanda Lagerkvist (London, New York: Routledge, 2018).

observer.[42] However, since any act of observation is premised on the observer's own local faculties and schemata, social reality is never "purely" observed. In other words, those claiming access to a "more objective" overview would in fact be doing so through *their own* constitutive milieu. I have argued that this is particularly decisive when considering social media platforms as interactive design objects, where intelligence-based design decisions intend to help analysts tracing user preferences through various dashboards (i.e., intelligence-compiling interfaces). Such dashboards capture discrete signals given off by users—while the users, on the other hand, are entangled in a mode of usage that is always and forever based on interpretations of the digital interfaces made by the designers. Here, too, we see a co-constitutive "dance" of interpretive agency: everyday users try to draw conclusions from signals provided through interfaces, while the designers of these interfaces try to draw conclusions from intelligence on the signals generated.[43] To really grasp the complexity at hand, it is instructive to peruse the literature on first- and second-order cybernetics and its debates on "observer problems."[44]

Macrostructural context would address the ways in which internet-based social media platforms are shaped by their various biases pertaining to ownership and the attendant economic incentives motivating their owners. This has to be understood in combination with fluctuations of world affairs, as these form a political-historical context. In conventional mass media system theory, macrostructural concepts—that is, the notion of (national) "public opinion"—as well as functional entities—that is, various notions of mass media systems—have been established epistemological tools for the last hundred years.[45] In recent years, this functional theory has been complemented by notions of "social media logic," which act to explain the workings of social media systems by reference to structural incentives. Importantly,

[42] Andersson Schwarz, "Umwelt."
[43] Andersson Schwarz, "Umwelt."
[44] See, e.g., Heinz von Foerster, Niklas Luhmann, Siegfried Schmidt, and Gregory Bateson.
[45] See, e.g., Walter Lippmann, Wilbur Schramm, and, later, Jürgen Habermas, David Altheide, Daniel Hallin, and Paolo Mancini.

the hybrid entanglement of mass media and social media has become something that is increasingly emphasized.[46] Mass media and social media interact, primarily in their mutual interplay of references. "Mass media logic and social media logic get incrementally entangled in defining the popularity of issues and the influence of people."[47]

Lastly, there is a larger cultural *global context*: behavioral, linguistic, narratological, and rhetorical modes of habit. These are patterns more stable over time than the geopolitical and events-based fluctuations in the above category. Habit and cultural norms rarely shift rapidly, and tend to follow generational patterns. It would be spurious to attribute agency to macro factors such as "globalization," "imperialism" *et cetera* without considering the intermediaries or mediators of such forces.[48] The study of so-called "bots" or "trolls" in social media settings would be an example of this; they can be regarded as concrete specimens whose very occurrence can be (albeit speculatively) traced to particular politico-historical conditions. The explanatory power of abductive theory can, if we are to apply this Latourian perspective, never be as certain as the original descriptive observations of these actual mediators. More on such abduction below.

Epistemological Challenges of Social Big Data for Scholarly Research

I will now turn to two research projects that I have had experience from, illustrating the challenges at hand. The first one addresses indexical, numerical data, and the second one symbolic, semantic data.

The first one is a study on *The Transformation of a Swedish Twittersphere* (#svpol, 2014 to 2018), targeting the logics of sharing and retweeting in the Swedish Twittersphere, and, in particular, the interactive logics between social and editorial media.[49]

[46] See e.g., Andrew Chadwick, Tarleton Gillespie, David Beer.

[47] José van Dijck and Thomas Poell, "Understanding Social Media Logic," *Media & Communication* 1, no. 1 (2013): 8.

[48] Latour, *Reassembling*.

[49] Andersson Schwarz et al., *Opinioner*.

Through a combination of methods, mainly SNA and quantitative content analysis, we were able to find patterns and could categorize the shape of affinities among people using the hashtag #svpol on Swedish-language Twitter, during the election year 2014. We could also trace the recycling of tweets in political debate and opinion formation, as well as illustrating some important aspects of the interaction between Twitter and editorial media.

Our study was based on an original data harvest of approximately 109,000 tweets covering the distribution of Sweden's most popular hashtag, #svpol, on 25 different days during the election year of 2014.[50] We combined SNA (14,412 nodes, 37,959 edges[51]) that showed network topology (cluster identification) with a rather conventional content analysis (n=500), in which a systematic selection of tweets was interpreted manually in order to assess semantic meaning. Moreover, our study was a typical example of collaboration between industry and academia. We reflected on the nature of such collaboration, and chose to emphasize the challenges—especially for established organizations within competitive intelligence and editorial media—inherent to the interpretation of social media flows.[52]

Four years later, I was able to make a repeat sample, representative of the same 25 days (i.e., election year 2018) from the same data provider (approx. 125,000 tweets; 42,645 nodes; 77,810 edges). By letting the procedural collection remain the same, the potential for comparability over time increases. While the analysis of this more recent data is still under way, I would like to stress the point that such comparability provides an important aspect of contextualization: regardless of how exogenously representative the data samples in question would be (and I maintain to exhaustively list the various ways in which they would be lacking in that respect), there would be endogenous consistency, making it revealing so as to see how the shape of this particular social space changes over time.

---

50 Data was provided by intelligence company Retriever, which in turn fetched Twitter data from data provider Sysomos.

51 Each edge is one instantiation of a tweet being retweeted.

52 Andersson Schwarz and Hammarlund, "Kontextförlust."

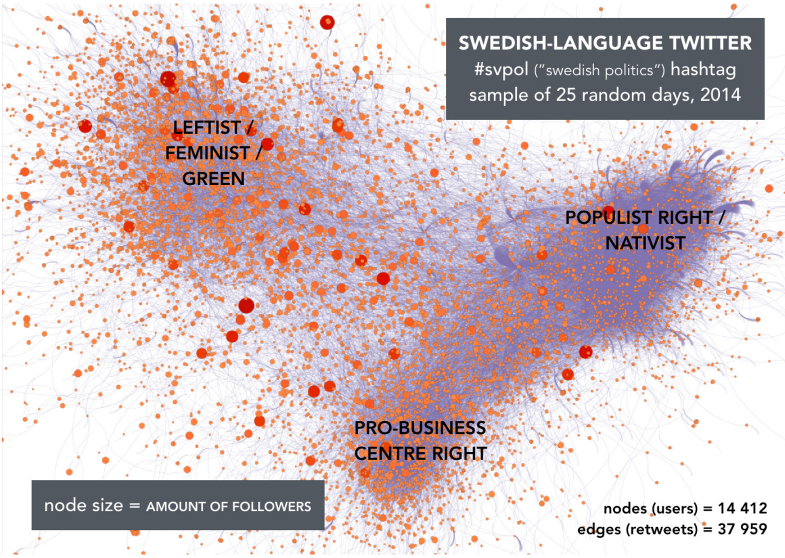


Figure 1: Political clusters in the Swedish Twittersphere, election year 2014. Copyright: Author. License: CC BY-SA.

One interim conclusion from these different analyses is that, while the first dataset (Figure 1) clearly showed three political clusters (leftist/feminist/green; pro-business/center-right; and populist-right/nativist), the later dataset (Figure 2) indicates that the "middle" center-right cluster has all but disappeared, while the populist-right/nativist cluster appears to have intensified in activity. This should not be understood to be showing a popularization of populist-right/nativist sentiment, as the dataset only captures tweets being deliberately tagged #svpol, and it remains unclear how much of the change is due to changing propensities among Swedish-language Twitter users to use that hashtag. However, it should be seen as indication that, for those who choose to use this hashtag, there seems to be an intensification and radicalization of use.

By making a strategic sample, interesting questions about representativity can be answered. The relatively small sample was big enough to generate a reliable visualization of the ways in which this Twitter hashtag propagated, simply by tracing the dissemination of a simple parameter: occurrences of retweets (disregarding the content of each tweet). While the study only observes

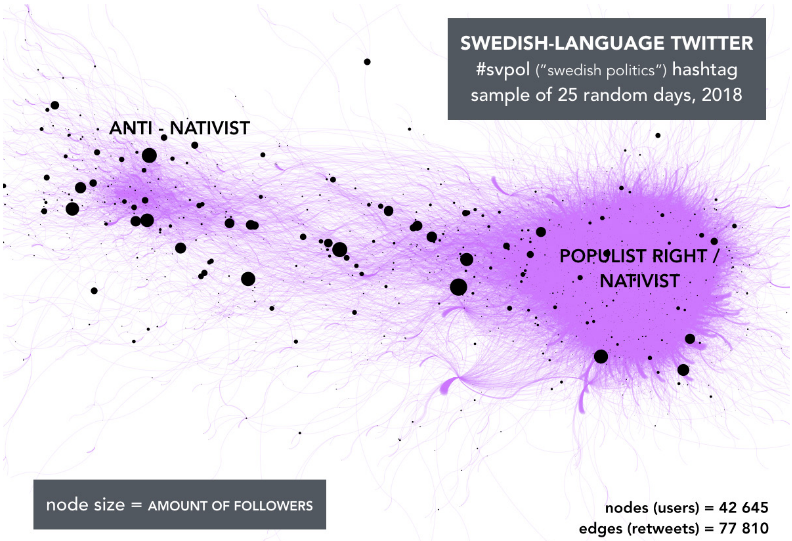


Figure 2: Political clusters in the Swedish Twittersphere, election year 2018. Copyright: Author. License: CC BY-SA.

the formations of a particular hashtag during particular time periods, some of the more principal logics observed are likely to be apparent also in social media in general. The repetition of the exact same sampling over time strengthens internal reliability, recording different historical states of this propagation.

The second project, *Linguistic Explorations of Society*, is transdisciplinary in nature and aims to investigate the very notion of representative text data for populations in various countries, primarily in order to lay the groundwork for better transnational comparability between open-text answers from official national surveys and found text data from online (editorial and/or social) sources.[53] The purpose is to extract large-scale language collections, to make possible purely statistical, computer-driven methods like NLP in order to find intralinguistic dimensions (e.g., entity recognition and linking, relationship extraction, sentiment analysis, topic detection, and author identification). In a preceding project,[54] the same researchers had employed similar methods in

53 Dahlberg, *Linguistic*.
54 Stefan Dahlberg, *Language Effects in Surveys* (Swedish Research Council, 2014).

order to find patterns in corpuses of accumulated responses to open-ended questions in formalized surveys on political opinion, around the globe. The purpose was to be able to more accurately estimate the inherent biases that would skew the comparability due to, for example, Spanish-language speakers, Russian ones, and Swedish ones meaning different things when they respond to words like "democracy."

Currently, I am in the process of surveying available online text data sources across countries. In addition to surveying things like internet penetration, media use, available social media platforms, modes of censorship, freedom of speech indices and so on, one of the key tasks is to unravel and critically describe the complex ecosystem of commercial vendors/providers of user-generated text data, for different languages and in different countries. Access to user-generated online text data is determined by the largely commercial nature of not only the sources, but also the redistributors of such data; complex arrangements of interlinking commercial vendors, each providing different modes of access and collections of sources (generally by way of different APIs[55] and/or visual dashboards). Each vendor offers a particular selection of sources, and moreover remains largely opaque as regards the provenance of these sources. The data in question is not Facebook data (which is, nowadays, practically unavailable, especially for commercial purposes) and rarely even Twitter data (due to arrangements to do with the terms of use of the Twitter API). Rather, the data in question is often scraped from wikis and other online corpora, blogs, and message boards on the open web, alongside various web-based editorial news sources. Moreover, vendors normally only make available the most recent 30 days of data, shelving older data on magnetic tape (thus making it significantly less accessible). Each vendor also presents its own, institutional terms of use, by default often making reuse of the data in question impossible. For these reasons, and several more, it really hard to employ traditional standards of statistical representativity and replicability. Thus, certain types of data are generally very hard to get hold of:

[55] Application programming interfaces.

- Chat app data from, for example, WhatsApp, WeChat, Snapchat, Kik, and Facebook Messenger.
- Individual posts (private and public) from Facebook, Instagram, and LinkedIn.
- Geo-tagged data.
- Demographic variables (e.g., gender, age, and income) for the sources in question.
- Historical data (older than 30 days).

Some interim conclusions can also be noted here. To begin with, few vendors provide adequate service-level agreements concerning data coverage and/or latency. There are generally no formal guarantees whatsoever concerning text quality. Vendors tend to guarantee only certain quantities of text and/or frequencies of specimens.

Moreover, legal conditions—such as copyright, data protection (e.g., the EU's GDPR), and terms of service—dictate a lot of archival uses of web-mediated data. These conditions are stacked on top of each other: the original platform (e.g., Twitter) would stipulate certain terms of use, but also the attendant commercial retailers stipulate their own, extraneous terms of use, often explicitly prohibiting any sharing of the data, due to its business value. All of these factors are inimical not only to the provenance, reliability, and representativity of the data—but also to the reproducibility of scientific results. Despite searching for the same time period and language, it is not hard to imagine one vendor providing a slightly different set of data than another vendor, depending on how they handle things like geo-tagging and/or language detection, original API access, search query designs, and so on. Even the point in time when the query is made would have effect, since the original platform might have removed old content at any time. There are many potential points of failure and/or bias, and providers are unlikely to give any formal guarantees regarding data completeness.

In order to enable at least formal reproducibility/replicability of scientific results, one can employ large-scale web scraping services, for example CommonCrawl, or commercial tools for scraping (so-called "crawler-as-a-service" tools), as alternative means to gather data for the purpose of large-scale processing and pattern recognition. However, manual web scraping is nontrivial: Data is always noisy, and requires considerable processing before use. For any scientific research project, this suggests that considerable resources should be considered for the purpose of scraping, cleaning, and preparing data alone.

Conclusions

Digitization has precipitated an ever-growing, immensely huge glut of behavioral data that I propose to label social big data, consisting not only of skeletal *indices* of human activity but also of *semantic* data user-initiated circulations of user-created bulletins in real time, rapidly shifting in terms both of volume and speed. This data can now be fetched in enormous volume, in near real time, and combinatory approaches can reveal patterns in this data glut. It is becoming increasingly necessary to critically interpret these data streams, while at the same time *not* seeing automatically generated tallies and anecdotal examples as automatically self-explanatory and representative.

Classic criteria for validation remain. *Validity:* does it measure the right thing? *Reliability:* does it measure it reliably, carefully, and comprehensively? *Representativity:* does it accurately correspond with that which is signified? *Replicability:* does the empirical endeavor allow for others to repeat it? *Salience:* does it have features which skew the observer's attention? Out of these important criteria, the three last ones appear to be the most troubling ones, when it comes to found specimens of online discourse. Overstating the certainty, veracity, and quality of social big data leads to four distinct problems.

First, in terms of claims to *objective realism* or representativity, what imaginaries of "the public" are generated? Social media are sometimes alleged to be "closer" to the public, a more "direct" mediation of public sentiment. The ontological observation that social media would "mirror" populations inevitably leads to the epistemological position that social media would be *better positioned* than, for example, editorial media to represent reality. Not only are the arguments presented in this chapter a rebuff to such claims; recent events such as the Cambridge Analytica debacle seem to have led many people also outside of academia to understand that such a position is no longer tenable.

Second, the *persuasive capacity* of numbers becomes urgent. This applies to discourses appearing in social media, as well as to references to social media appearing in mass media. Both tend to have heavily metrological characteristics, in that they divulge numerical tallies of different kinds. Numbers can in themselves have normative effects. The invocations to alleged quantitative impacts that onlookers would be able to make can often be rash and uncorrected misinformation can spread like wildfire. In order to be better at identifying possible misaligned claims to representativity, not only do we need better literacy on behalf of citizens; we also need to see improved transparency on behalf of platform providers.

Third, there are numerous *power dimensions* at hand. End users (not only ordinary citizens but researchers, planners, and journalists as well) are only allowed a filtered front-end access to the data glut to which platform providers like Google and Facebook have unfiltered back-end access. Hence, as end users, we are demoted to a secondary subject position in terms of knowledge generation. In that capacity, we are often forced to merely second-guess actual distributions. Lack of access thus enforces a form of *abductive* reasoning (qualified "guesstimation"). The inherent possibilities that this reinforces tendencies pertaining to the shaping of opinion and knowledge in society is a highly urgent sociological question.

Lastly, *contextualization* is unavoidable. Human beings cannot *not* interpret, and any data that is scrutinized by humans will (unwittingly or not) be categorized, valued, and even perhaps judged—even before formalized explications for such endeavors are formulated. However, it would be valuable to make continuous assessments also of the limits for *overcontextualization*. For the purpose of generating new, repurposed archives, the question of how one should relay archival contexts without overcontextualizing the archive is central.[56] It is also important for singular analytical projects, when abductive conclusions are drawn from data, so as to avoid qualitative overinterpretation of one's findings.

[56] There are many important differences between creating archives for posterity, and creating temporary datasets intended to be used only as interim working tools.

Ultimately, I advocate a pragmatic approach, combining data science with more conventionally interpretative methods. By carefully combining methods,[57] it is possible to maintain the benefits of conventional content analysis (systematic stringency, contextual sensitivity, hermeneutic depth) while providing effective capabilities for overview and/or enumeration through computational methods based on automation and/or algorithms (e.g., visualizing relationships and covariations, as well as detecting the relative sizes of different aggregations).

References

Andersson Schwarz, Jonas, and Johan Hammarlund. "Kontextförlust och kontextkollaps: Metodproblem vid innehållsanalys av sociala medier." *Nordicom-Information* 38, no. 3 (2016): 41–55.

Andersson Schwarz, Jonas, Johan Hammarlund, Stefan di Grado, and Magnus Kjellberg. *Opinioner och offentligheter online: Slutrapport för projektet Vad gör en politisk utsaga framgångsrik? Den användardrivna kommunikationens villkor*. Research report. Stockholm: IIS, 2015.

Andersson Schwarz, Jonas. "Umwelt and Individuation: Digital Signals and Technical Being." In *Digital Existence*, edited by Amanda Lagerkvist, 61–80. London, New York: Routledge, 2018.

Andreotta, Matthew, Robertus Nugroho, Mark J. Hurlstone, Fabio Boschetti, Simon Farrell, Iain Walker, and Cecile Paris. "Analyzing Social Media Data: A Mixed-Methods Framework Combining Computational and Qualitative Text Analysis." *Behavior Research Methods*, no. 51 (2019): 1766–1781.

Barabási, Albert-László. *Linked: How Everything Is Connected to Everything Else and What It Means*. New York: Perseus, 2002.

57 Seth C. Lewis, Rodrigo Zamith, and Alfred Hermida, "Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods," *Journal of Broadcasting & Electronic Media* 57, no. 1 (2013); Matthew Andreotta et al., "Analyzing Social Media Data: A Mixed-Methods Framework Combining Computational and Qualitative Text Analysis," *Behavior Research Methods*, no. 51 (2019).

Benkler, Yochai. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven, CT: Yale University Press, 2006.

Bolin, Göran, and Jonas Andersson Schwarz. "Heuristics of the Algorithm: Big Data, User Interpretation and Institutional Translation." *Big Data & Society* 2, no. 2 (December 2015): 1–12. DOI: https://doi.org/10.1177/2053951715608406

Boullier, Dominique. "Big Data Challenges for the Social Sciences and Market Research: From Society and Opinion to Replications." In *Digitalizing Consumption: Tracing How Devices Shape Consumer Culture*, edited by Franck Cochoy, Johan Hagberg, Magdalena Petersson McIntyre, and Niklas Sörum, translated by Jim O'Hagan, 20–40. London, New York: Routledge, 2017.

Dahlberg, Stefan. *Language Effects in Surveys*. Research project "The Advantage of Country Comparisons: Towards a New Method for Estimating Language Effects in Cross-Cultural Surveys." Swedish Research Council/Vetenskapsrådet, 2014.

Dahlberg, Stefan. *Linguistic Explorations of Society*. Research project "Studying Opinions and Societies Through Communicative Behavior Online: Assessing the Validity, Reliability, and Representativity of Online Text Data" (ongoing). Swedish Research Council/Vetenskapsrådet, 2017.

Daston, Lorraine, and Peter Galison. *Objectivity*. New York: Zone Books, 2007.

Firth, John. "A Synopsis of Linguistic Theory 1930–1955." In *Studies in Linguistic Analysis 1930–1955*, edited by John Firth, 1–32. Oxford: Basil Blackwell, 1962.

Foucault, Michel. "Panopticism." In *Discipline & Punish: The Birth of the Prison*, translated by Alan Sheridan, 195–228. London: Allen Lane, 1977.

Gibson, James J. *The Ecological Approach to Visual Perception*. New York, London: Psychology Press, 1979.

Gillespie, Tarleton. "The Relevance of Algorithms." In *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, 167–194. Cambridge, MA: MIT Press, 2014.

Hindman, Matthew. *The Internet Trap: How the Digital Economy Builds Monopolies and Undermines Democracy*. Princeton, NJ: Princeton University Press, 2018.

Hindman, Matthew. *The Myth of Digital Democracy*. Princeton, NJ: Princeton University Press, 2009.

Kitchin, Rob. “Big Data, New Epistemologies and Paradigm Shifts.” *Big Data & Society* 1, no. 2 (April–June 2014): 1–12. DOI: https://doi.org/10.1177/2053951714528481

Latour, Bruno. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Clarendon, 2005.

Lewis, Seth C., Rodrigo Zamith, and Alfred Hermida. “Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods.” *Journal of Broadcasting & Electronic Media*, 57 no. 1 (2013): 34–52. DOI: https://doi.org/10.1080/08838151.2012.761702

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *An Introduction to Information Retrieval*. Online ed. Cambridge: Cambridge University Press, 2009. https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.

Manovich, Lev. *The Language of New Media*. Cambridge, MA: MIT Press, 2001.

Marcuse, Herbert. *One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society*. Boston, MA: Beacon Press, 1964.

McLuhan, Marshall. *The Gutenberg Galaxy: The Making of Typographic Man*. Toronto: University of Toronto Press, 1962.

Morozov, Evgeny. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: PublicAffairs, 2013.

Nafus, Dawn. *Quantified: Biosensing Technologies in Everyday Life*. Cambridge, MA: MIT Press, 2016.

Piotrowski, Michael. “Digital Humanities, Computational Linguistics, and Natural Language Processing.” Presentation, *Lectures on Language Technology and History*, Uppsala University (March 4, 2016).

Piotrowski, Michael. *Natural Language Processing for Historical Texts*. London, Williston, VT: Morgan & Claypool, 2012.

Ruths, Derek, and Jürgen Pfeffer. "Social Media for Large Studies of Behavior." *Science* 346, no. 6213 (2014): 1063–1064.

Schubert, Lenhart. "Computational Linguistics." In *The Stanford Encyclopedia of Philosophy*. Spring 2020 ed., edited by Edward N. Zalta. Stanford, CA: Stanford University, 2020. https://plato.stanford.edu/archives/spr2020/entries/computational-linguistics.

Scott, John. "Networks & Relations." In *Social Network Analysis*, edited by John Scott, 1–9. London: Sage, 2013.

van Dijck, José, and Thomas Poell. "Understanding Social Media Logic." *Media & Communication* 1, no. 1 (2013): 2–14.

Webster, James. *The Marketplace of Attention: How Audiences Take Shape in a Digital Age*. Cambridge, MA: MIT Press, 2014.

Wetzel, Linda. "Types and Tokens." In *The Stanford Encyclopedia of Philosophy*. Fall 2018 ed., edited by Edward N. Zalta. Stanford, CA: Stanford University, 2018. https://plato.stanford.edu/archives/fall2018/entries/types-tokens.

Revisiting the Human–Society–Technology Nexus: Intercultural Communication Studies as a Looking Glass for Scientific Self-Scrutiny in the Digital Human Sciences

Jonas Stier

Introduction

> We—each one of us—live now, intermittently though quite often simultaneously, in two universes: online and offline. The second of the two is frequently dubbed "the real world," though the question whether such a label fits it better than it does the first turns more debatable by the day.[1] (Zygmunt Bauman)

More than 20 years ago, Manuel Castells[2] was already referring to the information or network society as a key feature of late or liquid modernity, to use Zygmunt Bauman's term.[3] Today, two

[1] Zygmunt Bauman, "Preface," in *Digital Diversities*, eds. Garry Robson and Malgorzata Zachara (Cambridge: Cambridge Scholars Publishing, 2014), ix.

[2] Manuel Castells, *The Information Age: Economy, Society and Culture. Vol. I: The Rise of the Network Society* (Oxford: Blackwell, 1996); Manuel Castells, *The Information Age: Economy, Society and Culture. Vol. II: The Power of Identity* (Oxford: Blackwell, 1997); Manuel Castells, *The Information Age: Economy, Society and Culture. Vol. III: End of Millennium* (Oxford: Blackwell, 1998).

[3] Zygmunt Bauman, *Globalization: The Human Consequences* (New York: Columbia University Press, 1998); Zygmunt Bauman, *Liquid Modernity* (Cambridge: Polity Press, 2000).

How to cite this book chapter:
Stier, Jonas. "Revisiting the Human–Society–Technology Nexus: Intercultural Communication Studies as a Looking Glass for Scientific Self-Scrutiny in the Digital Human Sciences." In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 49–74. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.c.

decades later, digitalization (or digitization) is a frequent and popular word in public discourse.[4] Policy-makers, "influencers," the "tech-industry," innovators, interest groups, researchers, and people in general, equally stress the importance and potential of digital technologies (DTs).[5] It is often argued that DTs are preconditional to innovation and long-term development. Yet, they are both products of and vehicles of globalization. Such technologies are commonly put forward as the solution *par excellence* when it comes to ensuring and facilitating people's learning and self-realization, stimulating economic growth, or making production more efficient. Thus, DTs are preferred weapons to combat the "grand challenges" facing humanity—be those an aging population, finding cures for diseases, combating intolerance, or facilitating sustainable development.[6]

Parallel to this, there is a grand narrative of an ongoing, inevitable, and irreversible digital transformation impacting all areas of society. This said, DTs are disruptive to social order.[7] For example, the American sociologist Richard Sennett argued as early as 20 years ago that the labor market has become increasingly flexible where limitlessness and insecurity are characteristics of work and employment.[8] On the labor market, the current digitalization pertains to an overarching societal process where work previously performed by people becomes automated—that is, at an accelerating pace tasks are taken over by computers, robots, automated

[4] Similarly, as Andersson Schwarz puts in elsewhere in this volume, "As *research objects*, digital media platforms are 'moving targets' in the sense that the sociocultural enactments taking place are ever-changing, making replication very hard."

[5] For practical reasons, I discuss digital media, social media, and new media as parts of digital technologies.

[6] At the same time, Viktor Arvidsson and Anna Foka argue: "Despite the World Wide Web's (WWW) profound societal effects over the past two decades, social science and humanities research appear strangely transfixed by pre-Internet media, information, and communication theories" (no page). "Digital Gender: Perspective, Phenomena, Practice," *First Monday* 20, no. 4 (2015), https://doi.org/10.5210/fm.v20i4.5930.

[7] Guo-Ming Chen, "The Impact of New Media on Intercultural Communication in Global Context," *China Media Research* 8, no. 2 (2012): 5.

[8] Richard Sennett, *The Corrosion of Character: The Personal Consequences of Work in the New Capitalism* (New York: Norton, 1998).

systems, or algorithms. This means that jobs previously done by humans are taken over by machines, at the same time as new jobs are created. It also means that enormous volumes of data can be stored, processed, made accessible and used for different purposes and interests.

Furthermore, there is a steadily growing number of what Nikolaos Mavridis calls "disembodied or even physically embodied intelligent software agents" on social network sites—producing and making use of data.[9] By the same token, digital products and services make up a constantly growing segment of global business and are today significant drivers of worldwide consumerism—once again using a term by Bauman.[10] Against this background, the shift from a late modern information or network society to a digital society has steadily accentuated the last decade.

Three decades ago, it was already being claimed that globalization and late modernity had left people's identities increasingly problematic.[11] It became more and more difficult for individuals to position themselves against traditional social categories such as class, culture, ethnicity, nationality, religion, or a given space or place. Bauman clearly envisioned that a characteristic of late modernity is constant movement—that is, even if we are physically still, we are in constant movement, traveling in physical space or cyberspace at an extreme pace. This movement leaves us nowhere for an extended period of time and makes us no more than *visitors*.[12]

Many claim that this social transformation has been accentuated in the new millennium.[13] Bauman argues that continuity and coherence have been replaced by discontinuity and lack of coherence, where time has fallen apart into events, happenings, adventures,

---

[9] Nikolaos Mavridis, "Artificial Agents Entering Social Networks," in *A Networked Self: Identity, Community, and Culture on Social Network Sites*, ed. Zizi Papacharissi (New York, London: Routledge, 2011), 291.

[10] Zygmunt Bauman, *Consuming Life* (Cambridge: Polity Press, 2007).

[11] Anthony Giddens, *Modernity and Self-Identity: Self and Society in the Late Modern Age* (Cambridge: Polity Press, 1991); Castells, *The Rise of the Network Society*; Bauman, *Liquid Modernity*.

[12] Bauman, *Globalization*.

[13] Jonas Stier, *(Van)modernitet och identitet* (Lund: Studentlitteratur, 2012).

and episodes.[14] Similarly, the German sociologist Hartmut Rosa argues that society of today is characterized by an accelerating transformation pace.[15] Rosa describes this transformation by distinguishing between three types of acceleration reciprocally affecting each other: "technical acceleration," "acceleration of social change," and "acceleration of 'the pace of life.'" Such accelerations impact both people's sense of time and their relationships to DTs. In fact, these accelerations are made possible *via* DTs, and are at the same time accentuated *by* such technologies.

Undoubtedly, this temporal and spatial transformation and recent "digital turn" have implications for people's sense of identity microlevel human interaction as well as for culture and society as such. An increasing presence in cyberspace has, for instance, led us to introduce terms such "in real life" (IRL)—almost implying that direct verbal interaction is an everyday life anomaly. Thus, a growing number of media through which we can communicate with one another has replaced interaction where interlocutors are physically present in the same place. It can be, therefore, argued that life online is no longer separate from life offline. For instance, learning is more and more taking place via DTs—such "teachnology" is common practice for new generations in schools all around the world, even prior to the Covid-19 pandemic. Moreover, DTs offer new arena for identity and *Gemeinschaft*. Guo-Ming Chen writes: "The time and space compression caused by the convergence of new media and globalization creates a universal cyberspace in which new cultural identity is merging in virtual communities."[16]

For the above reasons, it is often claimed that the overall digitalization of society, as well as the use of DTs for interaction purposes, constitutes one of the most profound transformations of human life. This said, there is a steadily growing volume of research and research methods addressing this change—focusing on a wide spectrum of aspects—innovation, design, digital readiness, digital

[14] Bauman, *Consuming Life*, 41.
[15] Hartmut Rosa, *Social Acceleration: A New Theory of Modernity* (New York: Columbia University Press, 2013).
[16] Chen, "The Impact of New Media," 5.

literacy, security, integrity, or effects on human interaction and identity, to mention only a few. Particularly preoccupied with the dialectical relations between DTs, humans, culture, and society is what is called the human–society–technology–paradigm, and science, technology, and society studies, STS.[17]

This research notwithstanding, online life is still often understood as distinct and separated from the "ordinary aspects of the human experience."[18] It is also striking how often such research and research methods are uninterested or insensitive to or altogether disregard the variation, diversity, and uniqueness of the cultural and societal backscene in the design of DTs, policies and overall discourses on digitalization, innovation, and DT development. Similarly, in public discourse DTs are often depicted as "objective" and void of cultural contents and underpinnings. As a friend of mine exclaimed, "In a nutshell, DTs are products of young, middle class Western males." This is likely to have social implications:

> The new culture hatched from new media creates a continuity gap between traditions and innovations within a culture. Before the emergence of new media […] traditions and innovations in human society co-existed in a dynamically synchronized way, but the speed and impact of the new media resulted in the inability of traditional values to keep pace with the new cultural values produced by new media. This cultural gap has caused difficulty in understanding or communication between generations and among people in the same culture.[19]

Although these matters may be addressed by some researchers, typically within the humanities and social sciences, more elaborate analyses of implications, such as the ones above, are often left outside the overall discussion on digitalization and DTs. Similarly, there is still limited scientific literature and empirical research focusing on DTs, diversity, and intercultural communication—its dynamics, dilemmas, *and* inherent potential—making

---

[17] Worth reading is Melvin Kranzberg, "Science-Technology-Society: It's as Simple as XYZ!" *Theory Into Practice* 30, no. 4 (1991): 234–241.
[18] Arvidsson and Foka, "Digital Gender," no page.
[19] Chen, "The Impact of New Media," 4.

use of theories, concepts, and research from intercultural communication studies (ICCS).[20]

Against this background, the aim of this chapter is to discuss DTs in the light of the cultural and societal backscene—including what is referred to as the discursive blind spots—in the design of DTs, policies, and overall discourses on digitalization, innovation, and DT development. This is done in the light of theories, concepts, and research from ICCS, research methodology, interdisciplinarity, and a variety of examples.[21] Based on these discussions, the case is made that the digital human sciences (DHV) offer a valuable contribution to the scientific understanding of the manifestations and consequences of digitalization. In particular, this chapter argues for the usefulness of "intermethodological," interdisciplinary, intercultural, and integrative approaches in DHV.

Discursive Blind Spots in DTs

In the design of DTs, policies, and overall discourses on digitalization, innovation, and DT development, a number of discursive blind spots can be discerned. Drawing upon my own previous research, I propose a typology of blind spots. These are referred to as "technocentrism," "normativism," "homo- and heterocentrism," "human egocentrism and ethnocentrism," and "the reversed problem imperative."[22]

Technocentrism and normativism

Over the years, there have been multiple and slightly different understandings of the concept of technocentrism.[23] Technocentrism

---

[20] As exceptions to the claim made, see Garry Robson and Malgorzata Zachara, eds., "Introduction," in *Digital Diversities* (Cambridge: Cambridge Scholars Publishing, 2014), 1–9, or Margaret D'Silva and Ahmet Atay, *Intercultural Communication, Identity, and Social Movements in the Digital Age* (New York: Routledge, 2020).

[21] I will use the term culture broadly, however, without elaborating on it.

[22] Jonas Stier, "The Blind Spots and Biases of Intercultural Communication Studies: A Discussion on Episteme and Doxa in a Field," *Journal of Intercultural Communication,* issue 24 (October 2010): 1–11.

[23] See for instance Samuel Hays, *Conservation and the Gospel of Efficiency* (Cambridge, MA: Harvard University Press, 1959); Seymour Papert, "A

is seen here as the unreflected notion of DTs as something innately positive, with the capacity to be the tool *par excellence* to solve the grand challenges of humanity. Such a technocentric attitude, in addition to the recent elevation of digitalization, permeates much of public discourse, rendering it an almost hegemonic position, where the consideration to culture is largely absent or placed in the periphery of debate.

In the light of technology in general and DTs more specifically, technocentrism departs from several seldom explicated or scrutinized normative assumptions. One is that innovation and technological advancements (including the wider notion of digitalization) are often and unreflectively equalized with (or, some would argue, mistaken for) social development and economic growth, and are thereby viewed as a necessity for humanity. In the light of our discussion, DTs are viewed as the primary weapons to combat the "grand challenges" facing humanity.

In addition, there is also a sense of technological *inevitability*—DTs are here to stay. While declining the prospect that DTs are conceived of as a nonoption, Unicorn HRO, a company offering digital technology solutions in the human resources sector, states:

> In today's corporate world, businesses must always be willing to adapt to new developments and change. This is certainly true with regards to the rapid emergence of new technological innovations. While new technology can be utilized to help businesses run more efficiently and productively, employers may find that their employees are not initially willing to accept new technologies in the workplace.[24]

Thus, there are divergent views on the usefulness and value of DTs among individuals, but at the collective level it is frequently posited that Swedish people hold relatively positive attitudes toward and have trust in DTs, and are relatively prone to adopt new DTs—compared to many other countries. Despite such differences, on a global scale, DTs are becoming increasingly intertwined with not

Critique of Technocentrism in Thinking about the School of the Future," paper composed in 1987; Stier, "The Blind Spots and Biases."

[24] Unicorn HRO website, "Handling Resistance to Technological Change in the Workforce," https://unicornhro.com/blog/handling-resistance-to-technological-change-in-the-workforce.

only with people's lifestyle but with their identity. Using the ideas of Erving Goffman,[25] more and more of people's "face-work" takes place via Facebook, Instagram, Twitter, Tinder, or LinkedIn, rendering DTs an increasing importance for their self-definition.[26] Or, citing Grant Bollmer:

> Our identities and bodies are quite literally shaped by the material and technical means we have for recording information and communicating. This does not mean that we are nothing other than our data, or that we do not exist outside the technologies we use. But it does mean that our sense of self and our physical sense of embodiment are both shaped by the media we use. Our identities are reimagined according to how media permit our bodies to interact.[27]

In the discourse, the absence of change and the presence of stability (i.e., the lack of innovation and technological advancements) are typically described in terms of unproductive stagnation or even disintegration, which in turn is conceived of as a threat to social development and long-term human prosperity. Both the precedence of development and change presupposes a strong sense of individual agency and a cultural preference for "doing" rather than "being" or "being-in-becoming"—something that in ICCS is referred to as activity orientation.[28] I found an example of this preoccupation with change and doing in a blog by Torben Rick, a senior executive with experience from management consulting and from the telecom sector: "[In digital transformation] [b]uilding a culture of constant change—a state of constant revolution is key."[29]

---

25 Erving Goffman, *The Presentation of Self in Everyday Life* (New York: Doubleday, 1959).

26 Jonas Stier, *(Van)modernitet och identitet* (Lund: Studentlitteratur: 2012).

27 Grant Bollmer, *Theorizing Digital Cultures* (Thousand Oaks, CA: SAGE, 2018), 116.

28 Florence Kluckhohn and Fred Strodtbeck, *Variations in Value Orientations* (New York: Row and Peterson, 1961); William Gudykunst and Young Yun Kim, *Communicating with Strangers: An Approach to Intercultural Communication*, 4th ed. (Boston, MA: McGraw-Hill, 2003).

29 Torben Rick, "Culture Change Is Key in Digital Transformation," blog, September 9, 2017, https://www.torbenrick.eu/blog/culture/culture-change-is-key-in-digital-transformation.

This means that, in doing-oriented cultures (Sweden or the US), activities focus on results that are "external to the individual, which can be measured by someone else. Activities must be tangible."[30] In such societies, social status and individual self-worth is to a higher extent derived from accomplishments and achievements, whereas in more being-oriented cultures (e.g., parts of the Middle East), more emphasis is on "a spontaneous expression of what is conceived to be 'given' in the human personality."[31] One's background and who one is have a greater significance for status and self-worth. In being-in-becoming societies (for instance, where Buddhism is predominant) concern is less on what people achieve and more on who they are and how they can become "an integrated whole in the development of the self."[32]

With this threefold distinction of activity orientation in mind, innovation and digitalization as they are often portrayed in public discourse draw from individual agency and proactivity, and are presumably more consistent with doing-cultures and being-in-becoming cultures than pronounced being-cultures. Although this largely remains an empirical question, one might think that DTs could serve to fulfill the ends of the other activity orientations as well.

Moreover, in discourse digitalization is depicted as a double-edged sword. On the one hand, DTs are often viewed as the paramount vehicles to change—be it for society, an organization, or the individual. DTs can and *should*, for instance, help us to protect, manage, and control nature/the environment. DTs are to facilitate sustainable development and assist us in combating poverty, criminality, or extremism, or reducing carbon dioxide emissions or making welfare systems more resilient. Or, with reference to common concerns of ICCS, DTs may provide means to counteract prejudice, ethnocentrism, cultural clashes, racism, and hatred with roots in dysfunctional communication across cultural boundaries.[33] Yet another example is how the educational sector

---

30 Gudykunst and Kim, *Communicating with Strangers*, 84.
31 Kluckhohn and Strodtbeck, *Variations in Value Orientations*, 16.
32 Gudykunst and Kim, *Communicating with Strangers*, 84.
33 Jonas Stier, "The Blind Spots and Biases."

in Sweden is working hard to implement digital "teachnology," which is said to optimize learning and knowledge production:

> Edmodo is an educational tool that connects teachers and students, and is assimilated into a social network. In this one, teachers can create online collaborative groups, administer and provide educational materials, measure student performance, and communicate with parents, among other functions. Edmodo has more than 34 million users who connect to create a learning process that is more enriching, personalized, and aligned with the opportunities brought by technology and the digital environment.[34]

Advertisements typically portray DTs as good, time-efficient, and status-providing—for example, successful people are busy, want to make a difference, and therefore early adopters of new DTs offered to them at a "good" price. Such representations convey the idea that new DTs are good *per se*.

The underlying reasons for "selling" DTs are, of course, both commercial and political. Today, DTs are seen as instrumental drivers of capitalism and for what Bauman calls consumerism.[35] Therefore, potential downsides or challenges following digitalization (e.g., risks with self-driving cars, personal data leakage, facial recognition, or autonomous weapon systems) are discursively constructed as either acceptable in the light of their foreseeable (positive) effects, manageable or transitory. Solutions or the failure of finding such solutions are typically attributed to the attitude or capacities of the individuals involved, be they citizens, consumers, politicians, students, or researchers in the human, legal or social sciences. Thus, the minority who do not adopt to the new technologies are criticized or dismissed as backward skeptics or "refusers," whereas the vast majority eventually will see the light.

On the one hand, digitalization *per se* is described as a challenge that must be managed—including for commercial and political reasons. In the light of their mission to facilitate Sweden's economic competitiveness, the Swedish Agency for Economic and

34 eLearning Industry website, "11 Digital Education Tools for Teachers and Students," https://elearningindustry.com/digital-education-tools-teachers-students.

35 Bauman, *Consuming Life*.

Regional Growth (Tillväxtverket) stresses the need for all sectors in society to master and take advantage of digitalization to ensure Sweden's competitiveness on the global market.

Homo- versus heterocentrism

In ICCS, the preoccupation with difference and variation is almost axiomatic. A bedrock assumption is that people, cultures, ethnicities, religion, and languages are dissimilar and that such dissimilarities are the roots of misunderstandings, clashes, or conflicts. To address such cultural variation, the works of Edward Hall,[36] Geert Hofstede,[37] and the World Values Survey are commonly referred to. Such variation has implications for social interaction. When cultural differences are at play, individuals engage themselves in intercultural communication. Everett Rogers and Thomas Steinfatt define intercultural communication as:

> the exchange of information between individuals who are unalike culturally. This broad definition implies that two or more individuals may be unalike in their national culture, ethnicity, age, gender, or in other ways that affect their interaction. Their dissimilarity means that effective communication between them is particularly difficult. *The cultural unalikeness of the individuals who interact is the unique aspect of intercultural communication*.[38]

Intercultural communication also pertains to interplay (voluntary or involuntary, intentional or unintentional), where the interlocutors *construct* one another as strangers or different in some respects.[39] In the humanities and social sciences, processes of "othering" and "difference" are focal concerns and there is an ample literature on what Bauman refers to as heterophily,[40] the degree to

---

[36] Edward Hall, *The Silent Language* (Greenwich: Fawcett, 1959); Edward Hall, *The Hidden Dimension* (New York: Doubleday, 1966); Edward Hall, *Beyond Culture* (Garden City: Anchor Press, 1976).

[37] Geert Hofstede, *Culture's Consequences* (Beverly Hills, CA: Sage, 1979); Geert Hofstede, *Culture and Organizations* (London: McGraw-Hill, 1991).

[38] Original emphasis. Everett Rogers and Thomas Steinfatt, *Intercultural Communication* (Long Grove, IL: Waveland Press, 1999), 79.

[39] Gudykunst and Kim, *Communicating with Strangers*.

[40] Zygmunt Bauman, *Modernity and the Holocaust* (Cambridge: Polity Press, 1989).

which people are (perceiving one another as) unlike,[41] and notions of difference, deviance, and outsiders.[42] These socially constructed notion of difference, deviance or outsider-ness are embedded in language—for example, in dichotomies such as: Swedes/non-Swedes, similar/dissimilar, different/not different, like/unlike—and fuel the idea of humanity as one of mainly multiplicity and diversity, not of community and universal qualities.

However, in the field of digitalization, distinctions and differences yield to unreflected universalism—at least at the collective level, where people are viewed as an almost monocultural collective, albeit different in their capacities, literacy, or susceptibility to new DTs. Thus, by contrast to the assumptions of ICCS, techno-centrism and normativism are commonly anchored in ideas of a homogenous world, where DT design is featured by both cultural universalism ("one size fits all") and individualism (DTs help me to find my personal lifestyle, define me as a person, or enable self-actualization). Considering globalization and generational differences in the uptake of DTs, such a universalism may create boundaries between people, both within and between countries and cultures. Chen writes:

> New media also extrinsically breeds communication gaps between different cultural and ethnic groups. The fragmented nature of new media has switched traditional cultural grammar, cultural themes, or cultural maps to a new pattern, resulting in the loss of traditional cultural logic.[43]

Even if there are analyses and theorizations, such as the one above, on innovation, technological development, and DTs, which account for human diversity, much of the discourse on the human–society–DTs nexus fail to consider cultural diversity or intercultural communication—and the potential consequences of digitalization on intercultural communication. When diversity and cultural variations are recognized, it is often more at a rhetorical level than genuinely and elaborately accommodating for such variation.

---

41 See, e.g., Rogers and Steinfatt, *Intercultural Communication*.

42 See Howard Becker, *Outsiders: Studies in the Sociology of Deviance* (New York: The Free Press, 1963).

43 Chen, "The Impact of New Media," 4.

Another aspect of heterocentrism is the culturally constructed distinctions between humans and machines, or between society and DTs. These distinctions are ontological, epistemological, and axiological in nature. Concretely, this means that reality is seen as consisting of the qualitatively distinct entities we refer to as humans and machines. Expressed in the language of homocentrism, humans and machines are ascribed different ethics (in most cases, the value of a person supersedes that of a machine). When these epistemologies are established, and in Peter Berger and Thomas Luckmann's terminology are "objectified,"[44] they underscore their own ontology. But previous objectivation also changes. For example, the dissolution of boundaries between man and machine has led to new words such as "cyborgs" or "bots" and may eventually lead us to change our use of pronouns, or introduce a new noun for an entity "in-between" a human and a machine. Over time, this can presumably make us less prone to think of and maintain the distinction between people and machines, though this calls for empirical research in the future.

Human egocentrism and ethnocentrism

Human beings are egocentric:[45] we are the subjects of our own experience of the world and the world gravitates around us as experiencing subjects. This has at least two implications. The first is that it appears as if we largely fall into seeing the world from our own perspective.[46] The second is that we cannot observe ourselves as observers and co-constructors of reality. So, as we use DTs, it provides us with a *personal*, yet culturally contingent and ethnocentric, window to the world. This means that our experience is formed experience against the backscene of society and culture. Society and culture serve as lenses of the culture and society we have internalized—for example, norms, values, meanings, world views, ideologies, and language. Once internalized, we take it for

---

[44] Peter Berger and Thomas Luckmann, *The Social Construction of Reality: A Treatise in the Sociology of Knowledge* (London: Penguin Books, 1966).

[45] In the sense here, egocentrism should not be mistaken for *selfishness*.

[46] To compensate for this, the importance of empathy, perspective-alteration, and mindfulness is often stressed.

granted, at the same time as we tend to favor our culture more than others. This is usually referred to as ethnocentrism.[47] For these reasons, our experience of the world provides merely an ethnocentric (and personalized) version of the world.[48]

As an example, we can look at the notion of privacy. Cultural codes stipulate whether a space is private or public or, in some instances, semiprivate.[49] As such, we see it both as taken for granted and "right" that our home is a private *place* where nonstrangers need permission to enter. There are variations in how accessible our home is to nonfamily members. Neighbors, acquaintances, or friends may be free to enter some areas (e.g., the living room), whereas this is not norm-consistent in others (e.g., the bedroom).

In many ways, DTs dissolve the notion of privacy, or at least redefine it. Today, you may be physically isolated or secluded in physical space, yet by being online we are open to the world and an indefinite number of places. Or perhaps we are members of a virtual place (e.g., a digital community). Thus, with DTs, new ways of maintaining and constructing the notion of privacy develop. The greater openness and access to the world, the more permeable one's personal boundaries to other people become—which, paradoxically enough, the less control and self-autonomy one may feel.

However, privacy extends beyond space and place, physical or virtual, and ties into personal integrity. In our part of the world—that is, a highly individualistic culture[50]—such integrity is associated (and even regulated) in relation to the *person*. In collectivistic cultures, integrity has a wider meaning, including not merely immediate family but extended family or the local community. The point is that when discussing integrity with regard to DTs an individualistic notion of integrity is *implied* or presupposed. Little

---

47 Marilynn Brewer and Donald Campbell, *Ethnocentrism and Intergroup Attitudes* (New York: Wiley, 1976).

48 See, e.g., William Sumner, *Folkways: A Study of the Sociological Importance of Usages, Manners, Customs, Mores, and Morals* (Boston, MA: Ginn & Co, 1906).

49 Hall, *The Hidden Dimension*; Gudykunst and Kim, *Communicating with Strangers*.

50 For an elaboration of the meaning of the term, see Gudykunst and Kim, *Communicating with Strangers*.

consideration is taken of the more collectivistic understanding of privacy.

The notion of time serves as another example. DTs both presuppose a certain notion of time *and* impact on people's sense of time. Yet, in this domain of society there are also cross-cultural variations. Using the concepts of monochronic and polychronic time cultures, cultures and societies differ in the relation to time.[51] In the Nordic countries, a monochronic time conception is common. It means that time is understood as linear—it is important, bought, lost, or killed. Time is managed and related to efficiency (and thereby the prospect of saving—or wasting—money!) as we plan, follow through, and follow up. Social expectations prescribe promptness and punctuality. By contrast, polychronic time cultures often adhere to a more cyclical view of time and is less preoccupied with time management and punctuality. Bringing DTs into the equation, common practice is to market new digital tools as "time-saving." In polychronic time cultures, the selling argument is more likely to be "status."

Drawing upon Bauman's and Rosa's writings,[52] digitalization fuels the transformation of time—a subsequent question is whether or not polychronic time conception will eventually have to yield to monochronic ones. With this in mind, it seems reasonable that a monochronic notion of time is more consistent with DTs than polychronic ones. Yet, regardless of collective notions of time, *individuals* may be more or less susceptible to the predominant time conceptions in the society in which they live, and as a consequence to DTs. Also, in this area, more humanistic and social scientific research is called for.

Just like time conceptions, DTs are extensions of culture; they mirror the culture in which they are designed.[53] Thus, embedded in DTs are symbols, words, images, colors, meanings, and communicative codes. Similarly, culture is embedded in discourses (e.g., policies, curricula, and mission statements) on DTs. For

[51] Hall, *The Silent Language*; Gudykunst and Kim, *Communicating with Strangers;* Jonas Stier, *Kulturmöten – En introduktion till interkulturella studier* (Lund: Studentlitteratur, 2019).

[52] Bauman, *Consuming Life*; Rosa, *Social Acceleration*.

[53] Hall, *Beyond Culture*.

these reasons, any producer or consumer of such technologies must possess a certain level of cultural competence, since the digital domain is largely permeated by a "Western model of thinking" and typically only marginal adjustments are made for "cultural reasons." But DTs also contribute to the formation of cultural forms. As people adopt DTs, new digital cultures are formed, with unique norms, terminologies, world views, and communities, which in turn affect people's identities and social relations.[54]

The reversed problem imperative

In ICCS, the problem imperative denotes a discursive tendency to "culturalize"—that is, to assume that culture, more than anything else, *determines* people's actions, or causes interpersonal conflicts, clashes, and communication problems.[55] Overall, cultural differences and multicultural contexts are presumed to lead to more problems than cultural homogeneity. Similarly, culture serves as a "garbage can explanation" for communication breakdowns, conflict, management styles, views on gender equality, criminality, poor health, marginality, etc., whereas intersectional analyses still are more uncommon.[56]

When it comes to cross-cultural variations leading to misunderstandings and problems, language and symbols are perhaps most obvious. Languages provide insights into cultures and thereby influence the way we see the world. Thus, *different* languages provide *different* world views. In the case of DTs, it is fair to claim that English is not merely the lingua franca, but the entire discussion around digitalization or digital design takes an Anglo-Saxon world view as the point of take-off (reading from left to right, that green means "ok" and red "no"—to mention just two examples).

For this reason, when discussing DTs, the commonly stressed distinction between high- and low-context cultures in ICCS needs to be accounted for.[57] In high-context cultures, a greater portion of the communicative message is located in the communicative situation (and less in the verbal) message, whereas the

[54] See Bollmer, *Theorizing Digital Cultures*.
[55] Stier, "The Blind Spots and Biases."
[56] Stier, *Kulturmöten*.
[57] Hall, *Beyond Culture*.

verbal message carries more weight in low-context communication. Similarly, codes surrounding communication vary between different parts of the world: how formal or informal one is when communicating with a stranger, power distance and status, prosody and nonverbal language, and assertiveness.[58] Many of these aspects may be less salient or a lost when communicating via DTs. Yet, if we refrain from recognizing their importance, they may cause misunderstandings and problems.

It can be assumed that digitally mediated communication exhibits more similarities to low-context communication than high-context communication. Thus, both the design of DTs rests on and better suits low-context communication, thereby suggesting the existence of fewer "problems" in digitally mediated communication in low contexts cultures. However, in low-context communication the absence of contextual cues may also be a challenge. Thus, to denote contextual information new symbolic systems evolve. Emojis, smileys, and abbreviations such as IRL or LOL are used on a global scale nowadays. In addition, there are idiosyncratic digitally related symbols in different countries. For the Swedes, "att sms:a" denotes "to text" with, for instance, your smartphone and for the German a "handy" is a cellular or mobile phone.

Despite what has been said, in discussions on DTs, culture and cultural variations are largely tucked away. Culture is simply not present or accounted for. Such a *reversed* problem imperative means that DTs are placed outside and above the realms of the cultural and social. They are given a status of being value-free, objective, and void of ideological and political underpinnings and consequences. Although questions of personal integrity and security have rendered some attention, foreseeable implications for democracy, citizenship, and interpersonal interaction still need more debate and analyses. This has drawn increasing attention, in the field of education, to what commonly is referred to as digital literacy. On the online learning platform Study.com, one can read:

[58] Hall, *The Silent Language*; Judee Burgoon, David Buller, and William Woodall, *Non-Verbal Communication: The Unspoken Dialogue (*New York: McGraw-Hill, 1996).

> The internet, search engines, email programs, blogs, and online videos have all contributed to our expanding knowledge and capabilities. Proper selection, use and understanding of these tools is a capability known as digital literacy […]. *Digital literacy* is the ability to navigate various digital platforms and understand, assess and communicate through them.[59]

Nevertheless, in an increasing global and multicultural world digital literacy needs to be accompanied by intercultural competence—not merely linguistic skills but also those pertaining to cultural codes, norms, values, proxemics, and cultural self-awareness,[60] to mention only a few. In providing the scientific basis for this, ICCS has a key role to play.

DHV as a Way of Revisiting the Human–Society–Technology Nexus

In one way or another, the overarching endeavor in science and research is to deconstruct complexity and then reassemble it to make it comprehendible, predictable, or meaningful. For much of the humanities and legal and social sciences, this endeavor entails the tripodal study of the relationship between individual and society, human variation and similarities, and social change. Grappling with this tripodal concern in relation to the complexity and multiplicity of digitalization and DTs is a key endeavor for DHV.

A vital DHV does not merely shed light upon or problematize the technological, societal, or human side of things, but on the dialectical nexus of the three. By being interdisciplinary and intermethodological, DHV offer valuable theoretical and methodological tools for understanding and problematizing how digitalization and DTs transform culture, society, the world, and the notion of humans.[61]

---

59 Study.com website, "What Is Digital Literacy? – Definition & Example," https://study.com/academy/lesson/what-is-digital-literacy-definition-example.html.

60 Jonas Stier, "Internationalisation, Intercultural Communication and Intercultural Competence," *Journal of Intercultural Communication*, issue 11 (April 2006): 1–11.

61 Interdisciplinarity as a matter of "researcher disciplinary representation" needs to yield to authentic researcher collaboration, where colleagues

Yet, all too often, analyses and theorizations of the human–society–technology nexus and on how DTs and overall digitalization, man, and society affect one another do not account for cultural diversity or digitally mediated intercultural communication. Rather, in DT design and in the discourse on digitalization there are several blind spots in ontology and epistemology and in research approaches targeting the human–society–DTs nexus. To study and problematize these blind spots and their implications is a matter for DHV.

Another case for DHV is found in recent innovations where technology is increasingly inserted into the human body to compensate for physical or psychological shortcomings or to cure medical conditions or simply to improve personal characteristics that individuals are dissatisfied with, which redefines the very old boundaries between human and nonhuman agents and qualities.[62] This, as well as DTs and digitalization as such, brings about a wide array of potential consequences for individual integrity, the manifestations and global distribution of high- and low-context communication or mono- and polychronic time conceptions. Will DTs eventually render low-context communication and monochronic time conceptions hegemonic positions in the world? To understand and problematize these questions, DHV need to a higher extent to integrate theories, empirical findings, and not the least, deconstructive perspectives from ICCS.

Furthermore, DTs are likely to continue transforming the form and execution of research. Whereas there is a proxy between quantitative research and big data, it is less so when it comes to qualitative methodology. However, DTs have made their way into qualitative research methodology and will most likely change the ways of conducting conversation analysis, discourse analysis, ethnography observations, interviews, etc., in the human, legal,

---

from different disciplines or fields are jointly and truly involved in the research design, data collection, and data analysis. If such research collaboration happens, the chances of unraveling, problematizing, and accounting for the complexity in the society–human–DTs nexus are increased.

62 Rosi Braidotti, *The Posthuman* (Cambridge: Polity Press, 2013); Francesca Ferrando, "Posthumanism, Transhumanism, Antihumanism, Metahumanism, and New Materialisms: Differences and Relations," *Existenz* 8, no. 2 (2013): 26–32.

and social sciences. Classic close reading of data (e.g., texts or artifacts) or analyses of multimodal interaction and representation are already complemented with new forms of ethnography and new DTs. For instance, sophisticated machine learning and artificial intelligence can be useful for qualitative data compilation and analysis, with regard to a variety of empirical contexts, including social media or the internet as such, will benefit DHV. Similarly, DTs play an increasingly important role in the cultural heritage sector, museum, libraries, and galleries.[63] Yet, digital research tools—as well as the methodology itself—need to be more sensitive to cross-cultural variations (e.g., communicative codes, stipulating for example prosody, pauses, directness, and frankness).

DHV have much to gain from being *integrative*, that is, to a larger extent join theory and practice, research and DTs use, or research and technologically oriented education. DHV can bridge scientific approaches and research findings with the professional (or commercial) realms, as research-informed design, good practice examples or methodologies. Reversibly, the professional domain may supply DHV with urgent research topics and serve as a benchmark to assess the adequacy and relevance of research. Such an integrative approach is enriching for all actors involved and may stimulate co-creative, interprofessional, and action-emancipatory approaches. This is consistent with the ambitions of open science.[64]

A Way Forward

It seems almost like a truism to claim that digitalization and DTs cut at across and have collective level homogenizing effects in all domains of everyday life and culture, or that digitalization

63 See Fiona Cameron and Sarah Kenderine, *Theorizing Cultural Heritage* (Cambridge, MA: MIT Press, 2010)

64 Increasingly, it is stressed that higher education and research needs to play a key role in development, innovation, and in the accommodation of the great challenges facing mankind. It is in this light the former EU-Commissioner Carlos Moedas' "three o-strategy" is to be seen: "Open Innovation, Open Science and Open to the World." European Commission website, "Open Innovation, Open Science, Open to the World – A Vision for Europe," https://ec.europa.eu/digital-single-market/en/news/open-innovation-open-science-open-world-vision-europe.

fuels the transformation of time and space. Consistent with the human–society–technology paradigm and science, technology, and society studies, the thesis put forward here has been that culture, technology, and individuals exist in a dialectical or, some would argue, symbiotic relationship. DTs are products of societies and culture where *people* design new technologies and use such technologies for self-presentation and self-representation.[65] Concurrently, these technologies have intended or unintended consequences for societies, culture and people—their everyday life, modes of communication, world views, time conceptions, and, at a deeper level, how they view themselves as individuals, groups, or humans.[66] But unintended consequences may not always be negative. Looking at social media, for instance, with reference to Agnieszka Stasiewicz-Bienkowska, Garry Robson and Malgorzata Zachara write:

> there is the potential of Facebook, Skype and the rest to aid sojourners in their attempt to manage feelings of vulnerability and dislocation early in their stay; as a set of tools and practices with which to convert the vulnerabilities and insecurities of depersonalized space into places which can be experienced as safe, solid and restful.[67]

It has been argued that cultural diversity and human variation often are downplayed or overlooked both in public discourse and in scientific approaches to the human–society–DTs nexus. Similarly, the cultural underpinnings of digitalization and DTs and cross-cultural variations with potential impact on people, society, and DTs need to be accounted for, with regard to high and low-context cultures, "being" and "doing" cultures, notions of time, privacy and integrity, to mention a few. Thus, discursive blind spots—that is, normativism and technocentrism, homo- and heterocentrism, human egocentrism and ethnocentrism, and the reversed problem imperative—may entrap us in an epistemological tunnel vision.

Today it seems fair to say that there is a growing realization that digitalization and DTs are a much too complex and voluminous

[65] See Nancy Thumim, *Self-Representation and Digital Culture* (London: Palgrave Macmillan, 2012).
[66] Kranzberg, "Science-Technology-Society."
[67] Robson and Zachara, "Introduction," 3.

matter for any sector in society or academic discipline to manage. Instead, in an increasingly digitalized, diverse, and dynamic world, traditional monodisciplinary approaches risk falling short, calling for multifactorial and interdisciplinary approaches when unraveling the human–society–DTs nexus—in computer science[68] and DTs design too. This demands that the human, legal, and social sciences are proactive and willing to renew themselves. Being interdisciplinary and "intermethodological," DHV are well equipped to explore and understand the human–society–DTs nexus, digitalization, and their implications for humans, culture, and society. To be even better equipped in this endeavor, it has been argued that DHV have much to gain from using of ICCS theories—as well as from intersectional, posthumanistic, and postcolonial perspectives.[69]

Although DHV need to be critical toward DTs, they have much to gain by being open to them as research tools for data collection and data analysis and may, over time, be more efficient in accounting for and enabling cross-cultural and comparative analyses, for example when it comes to communicative codes, symbols, the use of rhetorical resources, metaphors, or prosody.

Let me close with three statements from the online description of Rongie Bolanos's book *Digital Communications: Fundamentals and Applications* (2016),[70] where the first is almost apocalyptic: "Modern day life is consumed by technology. Without it, life would not be nearly the same. The digital world is growing bigger and more powerful." Statements like this one tell us why DHV are crucial. Perhaps the most important role of DHV is its deconstructive and critical views, for instance targeting the presupposition of cultural, ideological, or social homogeneity or epistemological dichotomies anchored in cultural ontologies—for example, that there is a qualitative and ethically motivated distinction between humans and machines or between society and DTs. Also, DHV

68 See, e.g., Brian Christian and Tom Griffith, *Algorithms to Live By* (London: William Collins, 2017).

69 Braidotti, *The Posthuman*; Ferrando, "Posthumanism, Transhumanism, Antihumanism, Metahumanism."

70 Rongie Bolanos, *Digital Communications: Fundamentals and Applications* (Burlington: Delve Publishing, 2016).

offer both retrospective and future-oriented analyses of the outcomes and consequences of digitalization and the use of DTs—on humans and society. Because, no matter what, two things ought to be certain—the world is always changing, and so are we, and we are never fully able to see what the future holds for us. Echoing the line of Bob Dylan, "You better start swimmin' or you'll sink like a stone, for the times they are a-changin'."

References

Arvidsson, Viktor, and Anna Foka. "Digital Gender: Perspective, Phenomena, Practice." *First Monday* 20, no. 4 (2015): unpaginated. https://doi.org/10.5210/fm.v20i4.5930

Bauman, Zygmunt. *Globalization: The Human Consequences*. New York: Columbia University Press, 1998.

Bauman, Zygmunt. *Liquid Modernity*. Cambridge: Polity Press, 2000.

Bauman, Zygmunt. *Liquid Fear*. Cambridge: Polity Press, 2006.

Bauman, Zygmunt. *Consuming Life*. Cambridge: Polity Press, 2007.

Becker, Howard. *Outsiders: Studies in the Sociology of Deviance*. New York: The Free Press, 1963.

Berger, Peter, and Thomas Luckmann. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. London: Penguin Books, 1966.

Bolanos, Rongie. *Digital Communications: Fundamentals and Applications*. Burlington: Delve Publishing, 2016.

Bollmer, Grant. *Theorizing Digital Cultures*. Thousand Oaks, CA: SAGE, 2018.

Braidotti, Rosi. *The Posthuman*. Cambridge: Polity Press, 2013.

Brewer, Marilynn, and Donald Campbell. *Ethnocentrism and Intergroup Attitudes*. New York: Wiley, 1976.

Burgoon, Judee, David Buller, and William Woodall. *Non-Verbal Communication: The Unspoken Dialogue*. New York: McGraw-Hill, 1996.

Cameron, Fiona, and Sarah Kenderine. *Theorizing Cultural Heritage*. Cambridge, MA: MIT Press, 2010.

Castells, Manuel. *The Information Age: Economy, Society and Culture. Vol. I: The Rise of the Network Society*. Oxford: Blackwell, 1996.

Castells, Manuel. *The Information Age: Economy, Society and Culture. Vol. II: The Power of Identity*. Oxford: Blackwell, 1997.

Castells, Manuel. *The Information Age: Economy, Society and Culture. Vol. III: End of Millennium*. Oxford: Blackwell, 1998.

Chen, Guo-Ming. "The Impact of New Media on Intercultural Communication in Global Context." *China Media Research* 8, no. 2 (2012): 1–10.

Christian, Brian, and Tom Griffith. *Algorithms to Live By*. London: William Collins, 2017.

D'Silva, Margaret, and Ahmet Atay. *Intercultural Communication, Identity, and Social Movements in the Digital Age*. New York: Routledge, 2020.

Dylan, Bob. "The Times They Are a-Changin'." From the album *The Times They Are a-Changin'*, 1964.

eLearning Industry website. "11 Digital Education Tools for Teachers and Students." https://elearningindustry.com/digital-education-tools-teachers-students.

European Commission website. "Open Innovation, Open Science, Open to the World – A Vision for Europe." https://ec.europa.eu/digital-single-market/en/news/open-innovation-open-science-open-world-vision-europe.

Ferrando, Francesca. "Posthumanism, Transhumanism, Antihumanism, Metahumanism, and New Materialisms: Differences and Relations." *Existenz* 8, no. 2 (2013): 26–32.

Giddens, Anthony. *Modernity and Self-Identity: Self and Society in the Late Modern Age*. Cambridge: Polity Press, 1991.

Goffman, Erving. *The Presentation of Self in Everyday Life*. New York: Doubleday, 1959.

Gudykunst, William, and Young Yun Kim. *Communicating with Strangers: An Approach to Intercultural Communication*. 4th ed. Boston, MA: McGraw-Hill, 2003.

Hall, Edward. *The Silent Language*. Greenwich: Fawcett, 1959.

Hall, Edward. *The Hidden Dimension*. New York: Doubleday, 1966.

Hall, Edward. *Beyond Culture*. Garden City: Anchor Press, 1976.

Hall, Edward. *The Silent Language*. New York: Anchor Books, 1981.

Hays, Samuel. *Conservation and the Gospel of Efficiency*. Cambridge, MA: Harvard University Press, 1959.

Hofstede, Geert. *Culture's Consequences*. Beverly Hills, CA: Sage, 1979.

Hofstede, Geert. *Culture and Organizations*. London: McGraw-Hill, 1991.

Kluckhohn, Florence, and Fred Strodtbeck. *Variations in Value Orientations*. New York: Row and Peterson, 1961.

Kranzberg, Melvin. "Science-Technology-Society: It's as Simple as XYZ!" *Theory Into Practice* 30, no. 4 (1991): 234–241.

Mavridis, Nikolaos. "Artificial Agents Entering Social Networks." In *A Networked Self: Identity, Community, and Culture on Social Network Sites*, edited by Zizi Papacharissi, 291–303. New York, London: Routledge, 2011.

Papert, Seymour. "A Critique of Technocentrism in Thinking about the School of the Future." Paper composed in 1987. http://www.papert.org/articles/ACritiqueofTechnocentrism.html.

Rogers, Everett, and Thomas Steinfatt. *Intercultural Communication*. Long Grove, IL: Waveland Press, 1999.

Robson, Garry, and Malgorzata Zachara, eds. "Introduction." In *Digital Diversities*, 1–9. Cambridge: Cambridge Scholars Publishing, 2014.

Rosa, Hartmut. *Social Acceleration: A New Theory of Modernity*. New York: Columbia University Press, 2013.

Sennett, Richard. *The Corrosion of Character: The Personal Consequences of Work in the New Capitalism*. New York: Norton, 1998.

Stier, Jonas. "Internationalisation, Intercultural Communication and Intercultural Competence." *Journal of Intercultural Communication*, issue 11 (April 2006): 1–11.

Stier, Jonas. "The Blind Spots and Biases of Intercultural Communication Studies: A Discussion on Episteme and Doxa in a Field," *Journal of Intercultural Communication*, 24 (October 2010): 1–11.

Stier, Jonas. *(Van)modernitet och identitet*. Lund: Studentlitteratur, 2012.

Stier, Jonas. *Kulturmöten – En introduktion till interkulturella studier*. Lund: Studentlitteratur, 2019.

Study.com website. "What Is Digital Literacy? – Definition & Example." https://study.com/academy/lesson/what-is-digital-literacy-definition-example.html.

Thumim, Nancy. *Self-Representation and Digital Culture*. London: Palgrave Macmillan, 2012.

Torben Rick blog. "Culture Change Is Key in Digital Transformation." September 9, 2017. https://www.torbenrick.eu/blog/culture/culture-change-is-key-in-digital-transformation.

Unicorn HRO website. "Handling Resistance to Technological Change in the Workforce." https://unicornhro.com/blog/handling-resistance-to-technological-change-in-the-workforce.

World Values Survey website. http://www.worldvaluessurvey.org/wvs.jsp.

Teleoptical Perspectives on Digital Methods: Scientific Claims and Consequences

Julia Pennlert, Björn Ekström, and David Gunnarsson Lorentzen

Introduction

Digital methods, according to Richard Rogers, "strive to make use not only of born-digital data but also the methods that are native to the medium."[1] Following Rogers's description, digitized or born-digital material opens up the possibility for adopting new computer-assisted research methods and techniques that collect, analyze, or describe a specific material. Such tools potentially influence the research process as well as the scientific outcome of a certain investigation or study. The development and use of digital methods within social sciences and the humanities have also had impact on scientific debates and discussions about how these new methods can contribute to the development of knowledge claims and knowledge-making in a contemporary media landscape.[2] As noted by Kenneth M. Price and Ray Siemens, digital technologies affect the ways that texts are available and perceived, which, according to Price and Siemens, leads to a need for the researcher to "adopt altered research methodologies."[3] The notion of change

[1] Richard Rogers, *Doing Digital Methods* (London: Sage, 2019).

[2] E.g., Katherine Hayles, *How We Think: Digital Media and Contemporary Technogenesis* (Chicago, IL: University of Chicago Press, 2012).

[3] Kenneth M. Price and Ray Siemens, eds., "Introduction," in *Literary Studies in the Digital Age* (New York: Modern Language Association of America, 2013), https://dlsanthology.mla.hcommons.org.

How to cite this book chapter:
Pennlert, Julia, Björn Ekström, and David Gunnarsson Lorentzen. "Teleoptical Perspectives on Digital Methods: Scientific Claims and Consequences." In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 75–102. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.d.

that Price and Siemens identify is intertwined with two important aspects, namely: how a specific material is presented and how this material might be studied with the help from digital technologies. Expressed differently, digitized or born-digital material potentially create, as stated by Price and Siemens, "new" research approaches where certain digital traits are addressed and analyzed.

The descriptions of the benefits of using digital methods are often associated with visual metaphors by for example emphasizing the importance of gaze, scale, or scope.[4] The "altered methodologies" that Price and Siemens describe are also framed as presenting the researcher with new perspectives through interfaces, functions, and applications that the digital methodological tools can utilize. In a similar visual description, Ted Underwood underlines the significance of digital methods as a way of striving to overcome the distance between researchers and research material.[5] Important benefits of these methods are that they, according to Underwood, present new ways of viewing research material. Underwood uses the following formulation to describe their potential: "a single pair of eyes at ground level can't grasp the curve of the horizon, and arguments limited by a single reader's memory can't reveal the largest patterns organizing literary history."[6] For Underwood, digital methods facilitate the researcher in presenting an unbiased overview of a vast material that might otherwise be impossible for an individual to get a hold of and see.[7]

The current discussions of new and digital methodologies have common traits across the social sciences and humanities, in terms of how the research process is affected by adopted methods. Digital methods have thus given scholars new vocabularies and new ways to grasp bigger fields of view.[8] While this is a positive

4 E.g., Katherine Hayles, *How We Think*, 27–31.
5 Ted Underwood, *Distant Horizons: Digital Evidence and Literary Change* (Chicago, IL, London: University of Chicago Press, 2019).
6 Underwood, *Distant Horizons*, x.
7 Cf. Alan Liu, "Theses on the Epistemology of the Digital: Advice for the Cambridge Centre for Digital Knowledge," Alan Liu's website, August 19, 2014, https://liu.english.ucsb.edu/theses-on-the-epistemology-of-the-digital-page.
8 Cf. Underwood, *Distant Horizons*.

development for technological advancement in social scientific and humanistic methodology, the use of these methods can also be criticized because some processes of knowledge-making may be hidden or made invisible through the utilization of digital methods. From our point of view, methodical approaches and their impact on the scientific process need further investigation not just by focusing on the descriptions and discussions as such in relation to research processes during the 21st century but rather to explore how the methods in use in different ways can be applied when studying and analyzing digitized or born-digital material.

Our aim with this contribution is therefore to discuss how digital methods can serve to provide the researcher with a *teleoptical perspective* on a certain digital material and explain how this perspective affects the researcher's gaze. In this chapter, we intend to investigate how the selection of certain digital methods and perspectives impact the research process, from data collection to selection, analysis, and interpretation.[9] The investigation is based on three case studies where different types of digital material are studied and analyzed according to the following questions:

- How does a selected digital method impact views of a specific body of material?
- What are the methodological consequences of these views?
- How can the concept of teleoptical perspective be used to understand and interpret the limitations and possibilities of digital methods for data collection and analysis?

A Teleoptical Perspective on Digital Methods

Before applying the analytical term *teleoptical perspective*, a short definition is needed. Following the abovementioned visual traits that can be found in the discussions of digital methods and the consequences for the individual researchers when applying them

[9] Our focus is not to analyze or discuss the research outcomes such as visualizations. For readers interested in an overview of the analytical vocabulary of information visualization, see Karolina Uggla's chapter in this anthology.

in a specific study, the teleoptical perspective that we propose in this text can be seen as an overarching analytical term. By using the term it is possible for the researcher to make visible and underline the potential relations between the researcher and the selected digital methods. Our purpose with this text is also to argue for an approach to that makes the uses of digital methods more transparent for the research community.[10]

The *teleoptical perspective* can also be seen to work on a more metaphorical level, as a way to formulate and grasp different processes and perspectives attached to various digital tools or methods. As noted by George Lakoff and Mark Johnson, a conceptual metaphor allows for understanding a concept through senses and physical experience.[11] Moreover, the use of a metaphor makes it possible to make abstract theorizing more specific and understandable.[12] In this sense it is therefore possible to think of the teleoptical perspective as a mode of analysis that works both on a concrete methodological level, and makes it possible for the researcher to draw theoretical and methodological conclusions of a conducted study.

In this chapter we also present subterms such as the concepts of *instrument*, *observation*, *lens*, and *filter* as an attempt to grasp the teleoptical perspective afforded by specific selected digital methods. The instrument is in this context defined as the tool through which data is collected and/or analyzed; the observation is what the researcher chooses to investigate through the instrument, such as specific sets of data and/or metadata. The lens is defined as a specification of the dataset, which determines whether to add or subtract data and/or metadata. The filter is a final calibration of specific aspects of the collected data. The filter can be utilized in different parts of the research process, initially to specify the data

10 For a discussion of the critique on how digital methods sometimes can be opaque for others, see Katherine Bode, *A World of Fiction: Digital Collections and the Future of Literary History* (Ann Arbor, MI: University of Michigan Press, 2018), 5.

11 George Lakoff and Mark Johnson, *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought* (New York: Basic Books, 2010), 45.

12 Lakoff and Johnson, *Philosophy in the Flesh*, 128.

collection (e.g., through search terms, file types, hashtags), and later on to further specify which data to focus on (e.g., searching within the collected data). Important to note is that all the steps depend on what the researcher aims to do with the data, as well as what is possible to do within the platform studied.[13] The different functions of the concepts, as well as what dimensions they cover, are illustrated by Table 1 below.

The chapter is based on three empirical case studies, which together investigate different aspects of our research questions. The first example discusses how automatic collection of data and metadata from Instagram can be understood using a teleoptical perspective. The second example shows how Twitter conversations concerning vaccination can be viewed from different perspectives, focusing on what aspects can be hidden by the application programming interface (API) that is used to collect data, and how these issues can be solved.[14] The third example deals with what kind of themes and terms are used in published articles from the scientific journal *Digital Humanities Quarterly* during the years 2007–19. In the two first examples, data is collected from social media platforms, hence born-digital material, while the third focuses on digital material published as open-access scientific journal articles. These three cases together demonstrate how digital technology can be viewed as both user-generated content

Table 1: The teleoptical metaphor as a tabulated model.

Instrument	Observation	Lens	Filter
Tool for data collection and/or analysis	Data and/or metadata observed	Command/ option for specifying observed data/ metadata	Command/ option for filtering specified, observed data/ metadata

[13] What kind of research endeavours the platform affords might be different from time to time; cf. Tommaso Venturini et al., “A Reality Check(List) for Digital Methods,” *New Media & Society* 20, no. 11 (November 1, 2018): 4195–4217, https://doi.org/10.1177/1461444818769236.

[14] Essentially, an API for this end is a set of rules and access points to different types of platform data, such as tweets, users, trends, etc.

and editorial-controlled texts that are selected and curated by platform owners and editors, respectively.

Instagram through the Lens of a Scraper

Studies of activity on Instagram can benefit from a critical perspective on automatic data collection to illuminate an otherwise black boxed procedure of collecting empirical data. This process of bulk data collection is often called web scraping. Web scraping should here be understood as "systematic collection of a specified type of data [from] websites[; t]o use a scraper means to start a little robot looking for a specific piece of code or a specific form of content, automatically archiving it in a dataset."[15] There are several methods for conducting this type of data collection. In this section, we will examine the web scraping method with a command-line application Instagram Scraper.[16] As a case example, the literary access and preservation project the Swedish Literature Bank (Swedish: Litteraturbanken) is investigated. The Swedish Literature Bank is a cooperative project by a number of literary and cultural heritage institutions—among them the Swedish Academy, the National Library of Sweden and the Society of Swedish Literature in Finland—providing digital editions of literary classics in the Swedish language as e-texts, facsimiles, and PDF or EPUB files.[17] For outreach purposes, the material provided by the Swedish Literature Bank is promoted through the project's Instagram account.[18] Using Instagram Scraper, the Swedish Literature Bank's outreach activity is examined and discussed in line with a critical perspective to investigate possibilities and limitations for web scraping social media metadata.

Instagram Scraper allows the automatic data collection of media files and metadata of single Instagram users, a list of such

15 Andreas Birkbak and Anders Kristian Munk, *Digitale Metoder* (Copenhagen: Hans Reitzels Forlag, 2017), 93 (our translation).

16 Richard Arcega, "Instagram-Scraper: Scrapes an Instagram User's Photos and Videos," GitHub website, https://github.com/rarcega/instagram-scraper.

17 Litteraturbanken website, "Om Litteraturbanken," https://litteraturbanken.se/om/english.html.

18 Litteraturbanken on Instagram, https://www.instagram.com/litteraturbanken.

users, or one or several hashtags. Instagram Scraper does not have a graphical user interface but rather relies on the user's familiarity with a command prompt or a terminal window. The options at hand within the command-line interface of Instagram Scraper can be understood as lenses or filters, since they allow the user to select which data to access. Instagram Scraper consists of a total of 25 flags, which can be understood as subcommands.[19] These can be used to specify, for example, the Instagram user through which the scraping operation will be processed, the media types (images, videos, stories) to scrape, whether or not profile metadata should be collected, and whether or not media metadata should be scraped, in JavaScript Object Notation (JSON) format.[20]

The following terminal command (Figure 1) searches images uploaded by the Swedish Literature Bank. It also generates a separate file with post metadata. At the time of scraping, the total number of Instagram posts by the Swedish Literature Bank adds up to 558.[21] The command returns 553 posts—including only image uploads—as well as associated metadata.

The terminal command can be broken down into the following concepts using the teleoptical metaphor (Table 2).

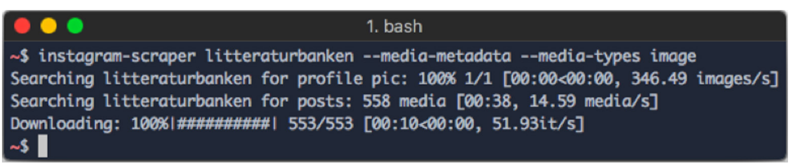


Figure 1: A terminal command for scraping media and metadata from the Swedish Literature Bank's Instagram account through Instagram Scraper. Copyright: Authors. License: CC BY 4.0.

Table 2: The terminal command described using the teleoptical metaphor.

Instrument	Observation	Lens	Filter
Instagram scraper	Litteraturbanken	Media/ metadata	Media type: image

[19] Arcega, "Instagram-Scraper."
[20] A format commonly used to serialize data, for instance in an API.
[21] November 14, 2019.

In the example command above, Instagram Scraper is the teleoptical instrument, while the following string—in this case the handle *litteraturbanken*—is the observation. The scholar's attention is thus pointed toward the digital surrogates of media uploaded to the Swedish Literature Bank's Instagram account. Additional flags added to this command can be understood as including or excluding data in terms of lens and filter perspectives. One lens is in use, namely "—media-metadata." This indicates that the lens is set to collect, apart from the media files, additional metadata associated with all of the Swedish Literature Bank's Instagram posts. These metadata is saved to a file titled "Litteraturbanken.json" within an automatically created folder titled "Litteraturbanken." The metadata retrieved includes information related to, for example, the types of media, the dimensions of media files, web links to the media file on Instagram's servers, the number of likes, the text written as a description for the media, any hashtags used for a post, timestamps, and view counts (if the media file is a video). The flag "—media-types image" command can be seen as a restrictive filter that tells Instagram Scraper to only search for media in image format. Additional lenses and filters can be combined and added to the data collection command through the use of flags and further zooming in or out on the observations.[22]

An excerpt of the metadata collected from the larger number of retrieved Instagram posts uploaded by the Swedish Literature Bank can be seen in the JSON file in Figure 2 below. Orange-colored text are keys and purple-colored text are values.

Figure 2 provides the metadata for a certain Instagram image post through key-value pairs. For example, line 272 describes the type of media of the post, namely a graphic image. Lines 275–276 depict the dimensions of the image uploaded in the post, namely 937 × 750 pixels. Line 278 returns the display URL (uniform resource locator, i.e., the web address) for the image uploaded. Line 280 returns the amount of likes for the post; in this case, the post has generated 53 likes. Line 286 provides a representation of the full text description accompanying the image upload. Line 292 displays the number of comments this particular post has

[22] Arcega, "Instagram-Scraper."

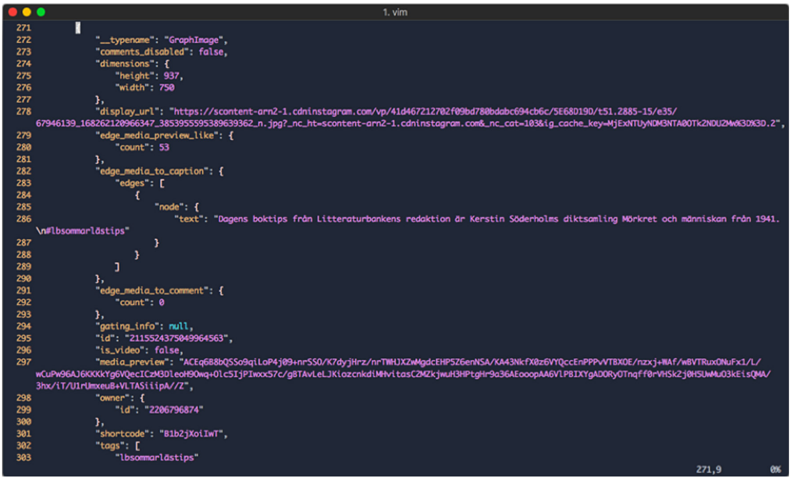


Figure 2: Metadata excerpt in JSON from a single Instagram post uploaded by the Swedish Literature Bank. Copyright: Authors. License: CC BY 4.0.

received, in this case none. Line 303 returns the hashtag that have been used in the upload for this post: "#lbsommarlästips."[23] This metadata has been collected based on the lens in place, namely media/metadata, and the filter, the image media type. This in turn denotes that all video media are excluded from the data collection, as the teleoptical filter is not focused on this particular media format. This extracted metadata can in turn be used as a basis for data analysis.

Thus, Instagram as a platform provides access to users' media uploads as well as metadata, collected through Instagram Scraper using a series of commands with additional flags specifying which data to obtain. Altering the data and metadata itself is not possible for the researcher. However, the options through which additional metadata and information can be chosen, added, ordered, combined, and filtered does indeed provide these types of choices through which the scope of the data and the collection procedure can be controlled and refined. Some of these flags are very much like tick boxes in a graphical user interface. By providing the flag

[23] Translation from Swedish: Summer reading tips from the Swedish Literature Bank.

"—media-metadata," the user metaphorically *ticks* a box telling the program to include media metadata and save it to a designated file when performing the scraping process.

Instagram Scraper hence places a perspective on the user's uploaded media files as well as the metadata assigned to either these media files or the user's profile as a whole. The inclusion of specific lenses enhances certain observations on a more granular (i.e., metadata through the command "—media-metadata"), while other lenses enhance other details. Certain lenses and filters enable the collection of certain data but limit or obscure other kinds of data. Acquiring datasets from this perspective facilitates qualitative analyses through investigations of the content of visual data as media files and textual data as media file descriptions, user biographies, hashtags, and comments, as well as a focus on material features.[24] Quantitative analyses can, for instance, include natural language processing techniques for topic modeling post descriptions in order to understand the text that accompanies image or video uploads.[25] Statistical analyses are also made possible through counting and visualizing the number of likes or view counts. Furthermore, computer vision techniques can be applied to the media files, which also provides a way to automatically identify objects in the Instagram media uploads.

Limitations to data collection occur as Instagram Scraper does not collect data on single users' practices like clicking, following, searching, scrolling, liking, commenting, previewing, sharing, bookmarking, tagging, reporting, searching, or discovering content. The tool also does not provide information about other users, tags, followers, or other forms of interaction. The teleoptical view is thus limited to providing data and metadata on users' media, rather than a full perspective on practices within Instagram. Practices related to the reporting of media deemed

---

[24] E.g., Janet Vertesi and David Ribes, eds., *Digital STS: A Field Guide for Science & Technology Studies* (Princeton, NJ: Princeton University Press, 2019); Paul Dourish, *The Stuff of Bits: An Essay on the Materialities of Information* (Cambridge, MA: MIT Press, 2017).

[25] E.g., Ted Underwood, "A Genealogy of Distant Reading," *Digital Humanities Quarterly* 11, no. 2 (June 27, 2017).

inappropriate fall out of the context as well, as the practices are hard to trace, and reported media deemed inappropriate by Instagram are supposedly blocked and removed from any web servers. In what follows, the teleoptical perspective will be further discussed in terms of Twitter conversation and retweet networks.

Filtering Twitter for Conversations and Patterns

Whereas Instagram is centered on visual media, their metadata, and image-based social interaction, Twitter emphasizes publicly visible individual messages and the conversation that these generate. When investigating Twitter, collecting data is arguably the most challenging part of the research process.[26] This undoubtedly makes choices of instruments and selection of suitable lenses and filters crucial. Technically speaking, the Twitter Streaming API pushes a stream of tweets for a client to filter in real time,[27] where a filter specifies what content to collect, for example through keywords or hashtags.[28] The chosen keywords can be seen as representing the observation, for example a topic. First, there are semantic challenges associated with the choice of keywords, because words and hashtags might represent different topics or aspects than those the researcher intended to study. Second, different qualities of the topic can be studied through the selection and combination of filters. Such qualities may include communication patterns, trends, and conversations. If the aim is to study conversations, a challenging issue arises as a consequence of the affordances of the API.

Following a conversational thread as it evolves is challenging because not all tweets in a thread include the selected keyword. A relevant option to consider is the user filter, which collects

---

[26] David Gunnarsson Lorentzen and Jan Nolin, "Approaching Completeness: Capturing a Hashtagged Twitter Conversation and Its Follow-On Conversation," *Social Science Computer Review* 35, no. 2 (April 1, 2017): 277–286, https://doi.org/10.1177/0894439315607018.

[27] The Streaming API does not offer an archive to search in, but through the Search API it is possible to retrieve tweets that are a few days old.

[28] The *track* parameter in the API, which can contain up to 400 keywords. The tweet is collected if it includes either of the selected keywords.

tweets if a selected user is either the sender or recipient.[29] However, not all tweets in the thread are posted by or to a user in the filter. Hence, choosing one filter leads to incomplete datasets, as only a certain set of objects are visible to the researcher (e.g., keyword-matching tweets). A proposed solution to this problem is to combine two filters, starting with a static set of hashtags and—as tweets are collected—build a dynamic list of the most active users in the dataset.[30] By matching tweet content with the list of keywords and the "in_reply_to_status_id_str" metadata field with tweet IDs in the database, both keyword-matching tweets and replies (with or without keywords) can be collected using continuous scripts.[31]

In what follows, data collection and selection will be illustrated through conversations around the topic of vaccination (Table 3). The first step is data collection through the combination of a keyword and a user filter.[32] Thirty-six keywords related to vaccination were selected as the first filter and, as data was collected, the 2,000 most active users over the last 48 hours were selected as the second filter.[33] The sizes and contents of the filters decide to what extent it is possible to study the intended object of research, in this case the conversations around the topic. If too much content has been produced, the user filter might not cover all users in the dataset that have been replied to. If the user filter contains too many highly active or popular users, there is a risk that an API cap will be reached or the computer may not be able to keep up with the stream.[34] The lenses represent what is focused on within the

[29] The *follow* parameter in the API, which can contain up to 5,000 user IDs. The tweet is collected if the selected user is the sender or receiver.

[30] Lorentzen and Nolin, "Approaching Completeness: Capturing a Hashtagged Twitter Conversation and Its Follow-On Conversation."

[31] Note that the Twitter API is currently under development. One consequence is that conversations will be easier to collect as it is possible to search with a conversation ID. However, this does not mean that conversations are automatically collected through the API. The researcher still needs a filter for keyword matching and a filter for the conversation.

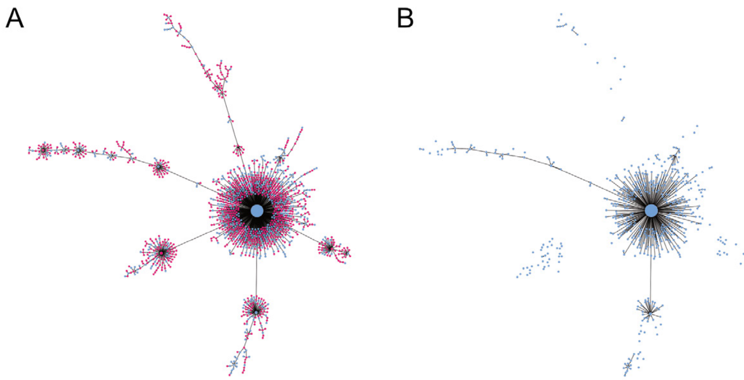
[32] Data collected November 19–26, 2018.

[33] The user filter was updated programmatically every 120 seconds.

[34] The API cap is set at 1% of all tweets produced at any given time.

Table 3: The demarcations and filters for conversation-oriented data collection.

Instrument	Observation	Lens	Lens	Filter	Filter
Custom made Python application based on Tweepy	Vaccination as a topic on Twitter, November 19–26, 2018	Micro level (conversations)	High detail level (metadata captured)	#antivaccination, #vaccine, #vaccination[35] (static)	List of 2,000 user IDs (dynamic)



Figures 3A–B: Two conversation networks where each dot is a tweet and the connection between two dots represents a reply. A: The full conversation. B: The part of the conversation where all tweets include either of the selected keywords. Replies without keywords in magenta, tweets with keywords in blue. Copyright: Authors. License: CC BY 4.0.

observation, in this case conversations about vaccination, guiding which metadata tied to tweets and users is captured.

The differences between the keyword and conversation-based approaches are illustrated in Figures 3A and 3B. In each of the figures, one tweet is represented by a node and edges are drawn between one tweet that is replying to another tweet. Figure 3A shows the structure of one conversation including all collected tweets in the thread, that is, what is collected using both

[35] Only a subset of the used keywords is included in the table.

filters combined. Figure 3B represents the same conversation but includes only tweets matching keywords. The fully connected conversation stretches longer and includes far more tweets, 1,416 compared to 509 (395 of them connected) in the second example. In the total dataset, 43% of the 302,328 tweets did not include a selected keyword, and 97 conversations with at least 100 tweets were found. The implications of these data qualities are that, when the contents and dynamics of the conversations are studied, the keyword-based approach would omit large parts of the discussions, including reactions to the collected tweets, and that the analysis of the topic would be biased toward the selected keywords. Additionally, the contexts of the disconnected tweets are unknown without access to the rest of the conversation.

As a sample, the 172,327 tweets collected with the keyword filter might constitute a substantial body of data, and the structure in Figure 3B represents a conversational thread as there are tweets connected as replies. When the contents of the tweets are analyzed, the researcher might be given the false impression of completeness, that is, that the data represented in the study is all data connected to the topic of vaccination. However, it is important to note that, due to the abovementioned issues regarding filter options, incomplete retrieval of tweets during data collection is still a risk, even though two filters have been used.

To highlight the necessity of using the two filters when studying conversations on Twitter, an analogy can be drawn between Twitter and a discussion forum. In the latter, the context of the posts is clearly visible, making it possible to analyze what the discussion is about. The affordances of the Twitter API make it challenging to collect tweets in their contexts. The example illustrates that, when viewing activity through different filters, a more complete picture of the conversational structure appears.

The conversation-based analysis enables a microlevel focus. Another way to analyze Twitter data is to focus on macrolevel patterns. In the following example, a dataset collected for the purpose of outlining communication patterns over time was utilized. This observation was made over eight weeks, when the instrument was programmed to search for tweets matching a set of keywords

Table 4: Demarcations and filters for macrolevel data collection.

Instrument	Observation	Lens	Filter
Custom Python application based on Tweepy	Vaccination as a topic on Twitter, eight weeks starting at January 14, 2019	Macro level, retweet patterns	Keywords: vaccine, vaccination, protective immunity, immune responses, immunogenicity

(Table 4).[36] Here, the first step is to create a zoomed-out overview of the communication patterns.

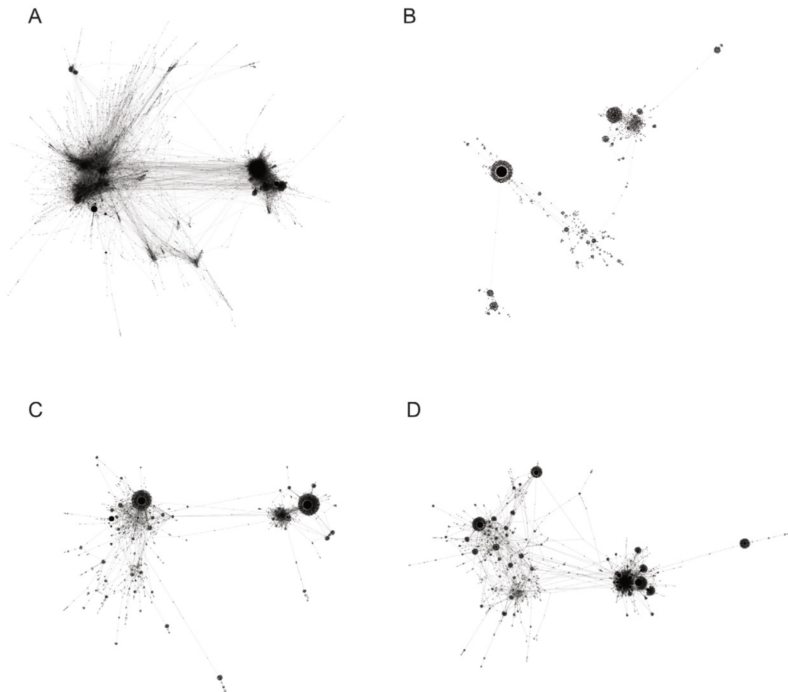
Here, the topic is represented through a series of retweet networks, in which each node is a user. A connection is drawn between two users if one redistributes the other user's content. Retweet networks tend to show clustering tendencies in topics that are polarized.[37] This makes them suitable for analysis of potential controversies, as users with similar viewpoints are likely to be grouped in clusters.[38] Through the use of the visualization application Gephi, networks were created using the force-directed layout algorithm *Force Atlas 2*. Figure 4 shows focused views of the observation where the largest connected part of users with at least two connections are kept.

---

36 Only a subset of the keywords used is included here.

37 Michael D. Conover et al., "Political Polarization on Twitter," in *Proceedings of the Fifth International Conference on Weblogs and Social Media* (Menlo Park, CA: AAAI Press, 2011), 89–96; David Gunnarsson Lorentzen, "Polarisation in Political Twitter Conversations," eds. Axel Bruns and Katrin Weller, *Aslib Journal of Information Management* 66, no. 3 (January 1, 2014): 329–341; Leo G. Stewart, Ahmer Arif, and Kate Starbird, "Examining Trolls and Polarization with a Retweet Network," in *Proceedings of WSDM Workshop on Misinformation and Misbehaviour Mining on the Web (MIS 2)* (New York: ACM, 2018), unpaginated, https://faculty.washington.edu/kstarbi/examining-trolls-polarization.pdf.

38 Noortje Marres, "Why Map Issues? On Controversy Analysis as a Digital Method," *Science, Technology & Human Values* 40, no. 5 (2015): 655–686.



Figures 4A–D: Four filtered views within the vaccination retweet observation, each representing a time-delimited slice of the retweet connections among users. A: aggregated network representing all eight weeks. B: week two. C: week five. D: week eight. Copyright: Authors. License: CC BY 4.0.

Figure 4A represents an aggregation of all eight weeks. It includes 25,157 users sharing 46,336 connections and has two large clusters separated by what appears to be a wide bridge. At the fringe areas, some medium-sized clusters appear. Focusing on weeks two, five, and eight (Figures 4B, 4C, and 4D), various patterns appear. Whereas the networks from weeks five and eight have similar polarized structures, week two does not. It consists of a few loosely connected islands, one of which includes a heavily retweeted user that is barely involved in the topic during the rest of the period. This network is also the smallest of the three, with its 2,414 nodes and 2,982 connections (week five: 4,745 and 6,314; week eight: 5,982 and 8,682). While the aggregated network shows clear clustering tendencies indicating that vaccination is a polarized topic, a one-week sample of Twitter activity within a given topic does not capture the characteristics of the topic on its own. The researcher thus needs to collect data for a longer time period and analyze the aggregated view as well as the individual views to capture polarizing tendencies, if any, as well as structural dynamics of the topic. Differences in characteristics regarding for example the number of participants, communication patterns, and network structures illustrate the problem of analyzing a snapshot, which in itself might not represent the activity particularly well.

Summarizing this section, it is important to address the issue of what these kinds of figures represent, when each figure represents a different perspective, rather than a whole. Through digital methods and visualization tools, different views can be presented, each of them giving a picture of a specific set of activities. By compiling several views side by side, a more representative picture of the discussions appears. Seemingly, the two large clusters linked with a wide bridge (Figure 4A) indicate different viewpoints colliding, which is a setting in which one could expect conversations to evolve. By following up the structural analysis with a qualitative reading of the tweets, representations of viewpoints can be identified. However, learning about the actual viewpoints requires different methods, such as interviews with the users. Content analysis of threads, both quantitative and qualitative, are recommended to understand the conversation dynamics. For large volumes of tweets, other relevant methods to consider are identification of linguistic patterns and topic modeling, which will be discussed in the next section.

Observing and Analyzing Patterns in *Digital Humanities Quarterly*

In this section, an investigation of the articles published in *Digital Humanities Quarterly* (*DHQ*) will be presented. *DHQ* is not a social media platform like Twitter and Instagram but rather a scientific online journal where the published articles have undergone a peer-review process and after that been published by the editors. By scraping, collecting, compiling, and visualizing data using a variety of digital methods, it is possible to investigate what themes, topics, or specific terms are covered in the articles. This third case study will explore the extent to which digital methods can be helpful in finding and charting patterns in all 308 articles published in *DHQ* during the years 2007–19.[39] *DHQ* is an "an open-access, peer-reviewed, digital journal covering all aspects of digital media in the humanities."[40] The journal is also related to and associated with the field of digital humanities and, in the section *DHQ on Digital Humanities*, it is stated that digital humanities can be defined as:

> a diverse and still emerging field that encompasses the practices of humanities research in and through information technology, and the exploration of how the humanities may evolve through their engagement with technology, media and computational methods.[41]

This last example will focus on how themes, terms, and topics found in published articles in *DHQ* can be said to relate to the journal's own definition of digital humanities, as an intersection of technology, computer-assisted methods, and/or media.[42] This is done in two steps, the first being to choose a suitable instrument for collecting the chosen material for observation, in this case the XML editions of the published articles. The second step is to compile the material into one text corpus and then to choose a digital method for charting and visualizing the themes and topics present in the material. These steps of the research processes and which instruments, lens, and filters have been used are presented in the table below (Table 5).

When the material is collected and compiled, it is possible to use various computer-assisted tools to visualize themes, subjects, frequencies, and relations between terms by looking at specific words or concepts. In the figure below (Figure 5), the material has been visualized through Voyant Tools, "a web-based text reading and analysis environment" suitable for working with text collections in different formats.[43] As for processing the material, it is

[39] The investigated texts are published under the section "Articles," and hence reviews and editorial texts are not included in the selection.
[40] *Digital Humanities Quarterly* website, "About DHQ," January 6, 2020, http://www.digitalhumanities.org/dhq/about/about.html.
[41] *Digital Humanities Quarterly* website, "About DHQ."
[42] *Digital Humanities Quarterly* website, "About DHQ."
[43] Voyant Tools website, "Getting Started," https://voyant-tools.org/docs/#!/guide/start.

Table 5: Description of the collection, selection, and specification of the investigated dataset.

Instrument	Observation	Lens	Filter
Python-Library Beautiful Soup	XML editions of *Digital Humanities Quarterly* journal articles	Journal issues between 2007, vol. 1, no. 1 and 2019, vol. 13, no. 2	<p> elements within <body> element

also possible to adjust the stopwords through a stopword list, which will exclude common words. In Figure 5, the words *digital* and *humanities* were added to the stopword list, among other common words, and are thus not included.

The visualization below presents an overview of the most frequently occurring words in the articles, such as "text," "new," "work," "research," "data," and "information." The fact that these words can be found in the material is no surprise, since the scientific discussion of the digital humanities as a discipline is often described as an intersection between humanities and digital technology, and hence a combination of text, data, and information.[44] As pointed out by Todd Presner et al., digital humanities can be seen as a universe where print-based media are no longer "the normative medium"[45] for knowledge production or for the object of study. In the figure above, words related to print-based media formats such as "book," "codex" and "print" do not appear at all. Instead words like "text," "information," and "work" appear. By zooming in on the smaller words, for example "model," "project," "process," and "tools," the compiled material also presents a picture of what Matthew G. Kirschenbaum identifies as a "common methodological outlook" that characterizes research within the

44 E.g., Patrik Svensson, "The Landscape of Digital Humanities," *Digital Humanities* 4, no. 1 (2010): http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html.

45 Todd Presner, Jeffrey Schnapp, and Peter Lunenfeld, "The Digital Humanities Manifesto 2.0," (2009) 391.Org blog, June 22, 2009, https://391.org/manifestos/2009-the-digital-humanities-manifesto-2-0-presner-schnapp-lunenfeld.

Figure 5: Word cloud showing the most significant words in articles published in *Digital Humanities Quarterly* 2007–19. Copyright: Authors. License: CC BY 4.0.

digital humanities.[46] These words can also be said to underline the research process in academic investigations that have an affinity with the digital humanities and can also contribute to a larger discussion about how studies in the digital humanities should be conducted and understood.

It is, however, possible to problematize the view of the material given above. Figure 5 can be used as an illustration of significant words in the material, but the visualization does not reveal potential relations between the words. To be able to grasp the connections between the terms, topic modeling can be useful. A topic can be described as “a collection of words that have different probabilities of appearance in passages discussing the topic.”[47] By choosing 100 topics and the 10 most significant words in every topic, it is possible to chart and explore the potential connections between the words. To summarize the result of the topic

46 Matthew G. Kirschenbaum, “What Is Digital Humanities and What’s It Doing in English Departments?” *ADE Bulletin* 150 (2010): 55–61, 56, https://doi.org/10.1632/ade.150.55.

47 Ted Underwood, “Topic Modeling Made Just Simple Enough,” The Stone and the Shell blog, April 7, 2012, https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough.

modeling accuracy, we have selected a couple examples of the topics revealed below:

- Topic 4: http, archive, search, library, users, resources, scholarly, collection, metadata, user.
- Topic 26: technology, users, word, tool, figure, user, objects, textual, http, visual.
- Topic 46: visualization, design, literary, visual, users, space, user, image, images, narrative.
- Topic 61: projects, search, archive, corpus, users, literature, literary, national, google, objects.
- Topic 89: space, infrastructure, landscape, humlab, visualization, cyberinfrastructure, technology, science, past, design.

To be able to explore the relations between the topics and the words in each topic, visualizing using Gephi can be useful. This step can be described with a table that in detail specifies the teleoptical perspective of the digital method in use (Table 6).

The result of these steps is illustrated by the network graph below (Figure 6). The figure shows the relations and positions between words/terms/concepts accessed through the compiled material of the scientific journal *DHQ*.

Through the color coding of word co-occurrences, it is possible to get a chart of connections between the identified topics. In that

Table 6: Description of the visualization, selection, and specification of the investigated dataset.

Instrument	Observation	Lens	Lens	Filter
Gephi	Topic modeled *Digital Humanities Quarterly* journal articles	Force Atlas 2 visualization algorithm	Word co-occurrence clustering through modularity function	10 most significant words from 100 topics. Words occurring in at least five topics are included in the visualization.

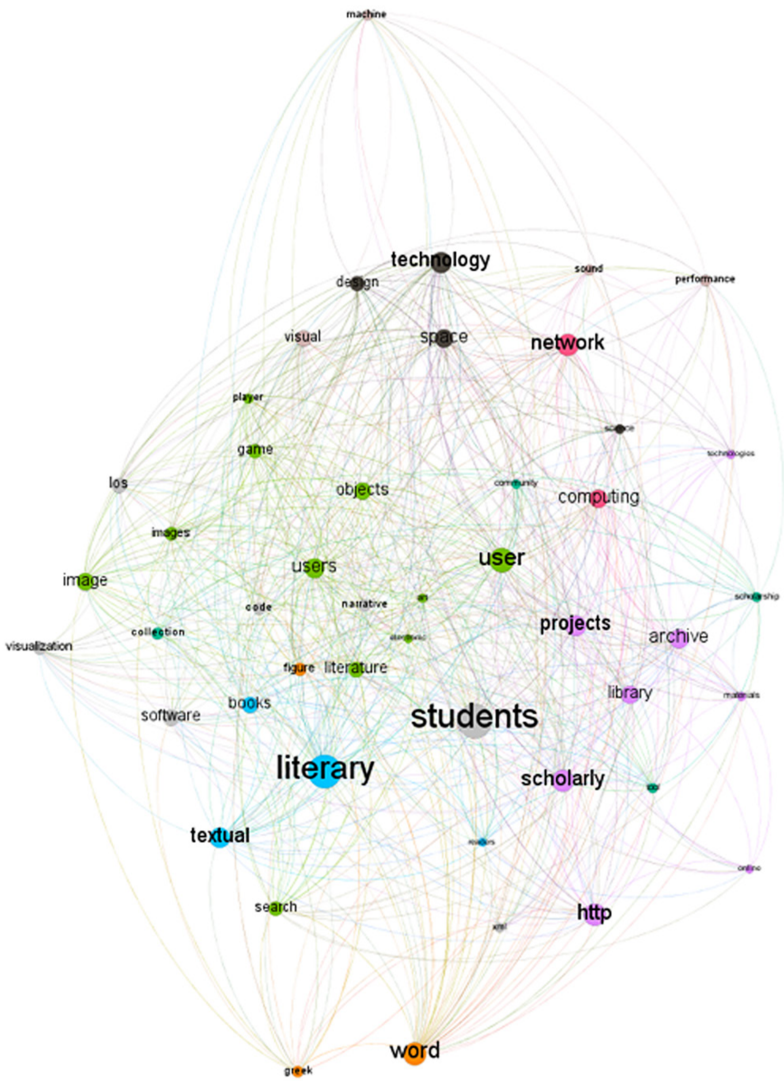


Figure 6: Relations and positions between words in published articles in *DHQ* 2007–19. Copyright: Authors. License: CC BY 4.0.

sense, the visualization in Figure 6 presents an overview of proximity conditions between words by, for example, locating words like “literary” (in green) and “students” (in gray) in relation to one another. However, the network graph also shows where there is no connection between the different topics and where the topics remain intact, as for example with the black topic covering words like: "machine," "technology," "design" and, "space." As a result, the network graph gives the researcher significant clues about the different relations between topics in the compiled material.

In a more general perspective, the investigated articles from *DHQ* are aligned with the journal's own definition of the digital humanities as an "engagement with technology, media and computational methods."[48] However, the visualizations above also make it possible to observe that the words, themes, and topics that are covered in the compiled material also show how "digital humanities hardly make up an uncontested or well-defined landscape."[49]

By investigating *DHQ* through different digital tools, it is possible to collect, compile, and visualize different aspects of the themes, terms, and words used in the scientific articles published in the journal. This can be done through various different digital methods and visualization tools that illustrate different aspects of the material. What these figures, however, do not reveal is how a specific word in a topic can be interpreted and analyzed more closely, within its full textual context. The teleoptical view as an observation across a given timespan is, similarly to the investigations of Instagram and Twitter, limited, because it fails to provide a close analysis of how words are used within their specific contexts, within individual articles.

Discussion and Conclusions

In this chapter, we explored if and how a teleoptical metaphor can be used as an analytical term for understanding and illuminating how a certain digital method can affect the steps of a research process as well as the results of a certain study. This was done by three case examples where concepts such as *observation*, *instrument*, *lens*, and *filter* were used in the processes of collecting, visualizing, and analyzing born-digital or digitized material. Our investigations were conducted through scraping the Instagram account of the Swedish Literature Bank, visualizing vaccine

48 *Digital Humanities Quarterly* website, "About DHQ," January 6, 2020, http://www.digitalhumanities.org/dhq/about/about.html.

49 Patrik Svensson, "The Landscape of Digital Humanities."

discussions on Twitter, and topic modeling and visualizing journal articles in *Digital Humanities Quarterly*. In what follows, these investigations are discussed further, in a wider context, with regard to the initial research questions of *how* digital methods alter scientific research practices and *what* the consequences are for employing certain digital methods.

The teleoptical view of scraping data and metadata from Instagram can be understood in operative terms through Instagram Scraper.[50] The scope is aligned, lenses and filters are fixed, and observations can be sighted depending on how these lenses and filters have been determined. Understanding web scraping as a digital method in teleoptical terms makes automatic data collection understandable, but also problematized with regard to which phenomena can be sighted. This method enables, for instance, distant reading of Instagram content, including descriptions, comments, or hashtags, which can help a researcher to understand word choices, grammatical patterns, or common topics between one or several Instagram accounts or hashtags. However, practices of—for instance—clicking, following, searching, previewing, and liking content cannot be understood solely through the utilization of Instagram Scraper, and therefore reveal a limitation of the web scraping method. The Twitter example shows that one filter gives a partial view of the context of the topic. The combination of two filters gives access to discussions connected to a certain topic. Another aspect relates to what the studied view actually represents. As with the Instagram example, it is important to acknowledge the limitations of the methods, and to bear in mind that the contents are representations or manifestations of opinions rather than actual opinions. Asking participants about their opinions and intentions would be one way to ground the results in other data sources.[51] Another issue that the researcher needs to be aware of is that, while large datasets need to be filtered so that they can be studied, the choice of filters influences the findings. Also, focusing on a shorter time period within the dataset will not capture a representative view of a topic. The research process involves multiple choices with regard to scope, filters, and

50 Arcega, "Instagram-Scraper."
51 Cf. Venturini et al., "A Reality Check(List) for Digital Methods."

lenses, all of which influence the findings and the scope of those findings. By collecting, organizing, and visualizing the scientific articles published in *DHQ*, it is possible for the researcher to get an overview of topics covered in the journal, as well as how these topics relate to each other. These kinds of overviews can work as a point of departure for a certain study, making it possible to grasp the notion of digital methods as offering a variety of ways to present different views or scopes, compared to closer scrutinization or manual textual analysis. However, in our example, the presented case studies do not cover individual traits in a specific text, nor how a specific article relates to a certain theme or topic.

By applying a teleoptical perspective on the three examples, we have shown how the research process—from selection and collection to analysis of data—can be made transparent. This has been done with the ambition to problematize their application, including how digital methods shape and limit the creation, interpretation, and framing of knowledge. Methodical transparency is key to providing an understanding of the contributions of digital methods as well as to make research processes visible.

Acknowledgments

The Twitter datasets were collected as part of the Data4Impact project. Data4Impact has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 770531. Thanks to Johan Eklund, senior lecturer at the Swedish School of Library and Information Science, University of Borås, Sweden, for web scraping and natural language processing assistance. Thanks to Gustaf Nelhans, senior lecturer at the Swedish School of Library and Information Science, University of Borås, Sweden, for network analysis assistance. Thanks to Litteraturbanken (the Swedish Literature Bank) for approving web scraping of the organizations' Instagram posts and associated metadata.

References

Birkbak, Andreas, and Anders Kristian Munk. *Digitale Metoder*. Copenhagen: Hans Reitzels Forlag, 2017.

Bode, Katherine. *A World of Fiction: Digital Collections and the Future of Literary History*. Ann Arbor, MI: University of Michigan Press, 2018.

Conover, Michael D., Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. "Political Polarization on Twitter." In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 89–96. Menlo Park, CA: AAAI Press, 2011. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847.

Digital Humanities Quarterly website. "About DHQ." January 6, 2020. http://www.digitalhumanities.org/dhq/about/about.html.

Dourish, Paul. *The Stuff of Bits: An Essay on the Materialities of Information*. Cambridge, MA: MIT Press, 2017.

Gunnarsson Lorentzen, David, and Jan Nolin. "Approaching Completeness: Capturing a Hashtagged Twitter Conversation and Its Follow-On Conversation." *Social Science Computer Review* 35, no. 2 (April 1, 2017): 277–286. https://doi.org/10.1177/0894439315607018.

Gunnarsson Lorentzen, David. "Polarisation in Political Twitter Conversations," edited by Axel Bruns and Katrin Weller. *Aslib Journal of Information Management* 66, no. 3 (January 1, 2014): 329–341. https://doi.org/10.1108/AJIM-09-2013-0086.

Hayles, Katherine. *How We Think: Digital Media and Contemporary Technogenesis*. Chicago, IL: University of Chicago Press, 2012.

Kirschenbaum, Matthew G. "What Is Digital Humanities and What's It Doing in English Departments?" *ADE Bulletin* 150 (2010): 55–61. https://doi.org/10.1632/ade.150.55.

Lakoff, George, and Mark Johnson. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books, 2010.

Litteraturbanken on Instagram. "Litteraturbanken.se." https://www.instagram.com/litteraturbanken.

Litteraturbanken website. "Om Litteraturbanken." https://litteraturbanken.se/om/english.html.

Liu, Alan. "Theses on the Epistemology of the Digital: Advice for the Cambridge Centre for Digital Knowledge." Alan Liu's website, August 19, 2014. https://liu.english.ucsb.edu/theses-on-the-epistemology-of-the-digital-page.

Marres, Noortje. "Why Map Issues? On Controversy Analysis as a Digital Method." *Science, Technology & Human Values* 40, no. 5 (2015): 655–686.

Moretti, Franco. *Distant Reading*. London: Verso, 2013.

Presner, Todd, Jeffrey Schnapp, and Peter Lunenfeld. "The Digital Humanities Manifesto 2.0 (2009)." 391.Org blog, June 22, 2009. https://391.org/manifestos/2009-the-digital-humanities-manifesto-2-0-presner-schnapp-lunenfeld.

Price, Kenneth M., and Ray Siemens, eds. "Introduction." In *Literary Studies in the Digital Age*. New York: Modern Language Association of America, 2013. https://dlsanthology.mla.hcommons.org.

Rogers, Richard. *Doing Digital Methods*. London: Sage, 2019.

Stewart, Leo G., Ahmer Arif, and Kate Starbird. 2018. "Examining Trolls and Polarization with a Retweet Network." In *Proceedings of WSDM Workshop on Misinformation and Misbehaviour Mining on the Web (MIS 2)*. New York: ACM, 2018. https://faculty.washington.edu/kstarbi/examining-trolls-polarization.pdf.

Svensson, Patrik. "The Landscape of Digital Humanities." *Digital Humanities* 4, no. 1 (2010). http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html.

Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change*. Chicago, IL, London: University of Chicago Press, 2019.

Underwood, Ted. "A Genealogy of Distant Reading." *Digital Humanities Quarterly* 11, no. 2 (June 27, 2017). http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

Underwood, Ted. "Topic Modeling Made Just Simple Enough." The Stone and the Shell blog, April 7, 2012. https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough.

Vertesi, Janet, and David Ribes, eds. *Digital STS: A Field Guide for Science & Technology Studies*. Princeton, NJ: Princeton University Press, 2019.

Venturini, Tommaso, Liliana Bounegru, Jonathan Gray, and Richard Rogers. "A Reality Check(List) for Digital Methods." *New Media*

& Society 20, no. 11 (November 1, 2018): 4195–4217. https://doi.org/10.1177/1461444818769236.

Voyant Tools website. "Getting Started." https://voyant-tools.org/docs/#!/guide/start.

Interpreting Information Visualization

Karolina Uggla

Information visualization has become a persistent part of the visual culture of the first decades of the 21st century. Such visualizations of various kinds appear in news media, on screens, and in mobile apps with specific functions in industries, organizations, and academic disciplines. Information visualization, as a phenomenon closely related to statistical visualization, can be said to have its origins in the enlightenment era of encyclopedias and instructional images. In the 1960s, Roland Barthes identified and interpreted how the illustrations of Denis Diderot and Jean le Rond D'Alembert's *Encyclopédie* (1751–72) were structured to convey information. Others after him place these illustrations as a novel and groundbreaking way of visually representing information.[1] In contemporary writing on information visualization, there has been increased attention, both in an academic and in a popular scientific discourse, to understanding information visualization by rediscovering and studying the phenomenon and its so-called roots and origins. Some individual predecessors working with statistic visualization in the 19th century are, for instance, Joseph Priestley, William Playfair, Florence Nightingale, and Charles-Joseph Minard.[2] It is clear that some visual forms

[1] John B. Bender and Michael Marrinan, *The Culture of Diagram* (Stanford, CA: Stanford University Press, 2010); Charles Kostelnick and Michael Hassett, *Shaping Information: The Rhetoric of Visual Conventions* (Carbondale, IL: Southern Illinois University Press, 2003).

[2] Sandra Rendgen et al., *History of Information Graphics* (Cologne: Taschen, 2019); Daniel Rosenberg and Anthony Grafton, *Cartographies*

How to cite this book chapter:
Uggla, Karolina. "Interpreting Information Visualization." In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 103–126. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.e.

that appeared then, such as the line chart, the bar chart, and circle diagrams, have become visual conventions that keep reoccurring even at times when information visualization is produced and represented in digital environments.[3]

The aim of this chapter is to map out and exemplify how *information visualization* is defined and analyzed as visual objects. These visual representations, if at all acknowledged as nontextual representations, occupy a position in-between image and text and their interpretations are methodologically hard to capture.[4] Visualizations of many different kinds are being used as tools in the humanities and the social sciences; however, these methods are not the focus of this chapter.[5] The aim of observing types of visual analysis of visualizations originates from my belief that the competence of interpreting visual material in the humanities, like in the subject areas of art history and visual studies, would benefit from more consciousness when it comes to interpretative tools and terminology for information visualization. What is at stake here is not only to be able to read charts and graphs critically but also to make information visualization more transparent as

of Time (New York: Princeton Architectural Press, 2010); Stephen Boyd Davis, "Beholder of All Ages: The History of the World in a French Mappemonde," *Textimage: Revue d'étude du dialogue text-image* 7 (July 2015): 1–6; Johanna Drucker, *Graphesis: Visual Forms of Knowledge Production* (Cambridge, MA: Harvard University Press, 2014); Sandra Rendgen, *Minard System: The Complete Statistical Graphics of Charles-Joseph Minard* (Cologne: Taschen, 2018).

[3] Lev Manovich, "What Is Visualization?" *Visual Studies* 26, no. 1 (March, 2011): 38; Kostelnick and Hassett, *Shaping Information*.

[4] These are noticed and categorized by, for instance: Sybille Krämer, "Trace, Writing, Diagram: Reflections on Spatiality, Intuition, Graphical Practices and Thinking," in *The Power of the Image: Emotion, Expression, Explanation*, eds. András Benedek and János Kristóf Nyíri (Frankfurt am Main, New York: Peter Lang, 2014), 3–22; Sybille Krämer, *Medium, Messenger, Transmission: An Approach to Media Philosophy*, trans. Anthony Enns (Amsterdam: Amsterdam University Press, 2016).

[5] See Anna Foka and Stefan Gelfgren, "Visualisering som verktyg och metod för historieforskning," in *Digital humaniora: Humaniora i en digital tid*, eds. Per-Olof Erixon and Julia Pennlert (Gothenburg: Daidalos, 2017), 147–164, and Andersson Schwarz and Pennlert et al. in this volume for examples of the use of visualizations in the humanities and the social sciences.

subject to analysis. My chapter provides a starting point for this greater goal: It makes a survey of some of the literature on information or data visualization in order to distinguish tools and terminology that has been previously used in the different branches of the field. These, of course, have different purposes, many with the objective of educating designers. Nevertheless, they may be found usable for other ends. After this brief literary overview, focusing on terminology, I exemplify how information visualization can be interpreted as visual objects of study with two different frameworks: one based on semiotics and the other directed at the emotions and ethics of information visualization.

Information Visualization Terminologies

Many different concepts and meanings circulate when addressing the phenomenon of information visualization, data visualization, or visual representations of numerical value. The term *information visualization* will be used here as a family name of a category of visual representations, both historical and more recent examples.[6] Information visualization, or "infoviz," is a research field, referring both to the discipline and the objects of study.[7] To work with information visualization is to work with the structuring, categorizing, and designing of data or information, as well as with technological tools.[8] Parallelly, the name data visualization, or "dataviz," is being used.[9] The subjects to be visualized—the concepts of information or data—are not entirely unproblematic. Information is a curious entity, an abstract concept that can

[6] Stuart K. Card, Jock D. Mackinlay, and Ben Schneiderman, "Information Visualization," in *Readings in Information Visualization: Using Vision to Think* (San Francisco, CA: Kaufmann, 1999), 8. According to this source, the term information visualization was first used in 1989.

[7] Liam McNabb and Robert S. Laramee, "Survey of Surveys (SoS) – Mapping the Landscape of Survey Papers in Information Visualization," *Computer Graphics Forum* 36, no. 3 (July 4, 2017): 589–617, https://doi.org/10.1111/cgf.13212.

[8] Yvonne Eriksson, *Bildens tysta budskap: Interaktion mellan bild och text* (Lund: Studentlitteratur, 2017); Manovich, "What Is Visualization?" 38.

[9] Andy Kirk, *Data Visualisation: A Handbook for Data Driven Design* (Los Angeles, CA: Sage Publications, 2019), 27.

contain almost anything. The established conception of information as an entity or stream with exchangeable content is dependent on information or data translated to numerical value or units, and is often said to be contemporary with the visual communication model of Claude Shannon's "Mathematical Theory of Communication" (1948).[10]

Scientific and medical visualization is sometimes seen as a branch on the tree of information visualization, sometimes separated with its own properties. These orientations can be described as "data that describes a physical phenomenon."[11] Lev Manovich has made the observation that scientific visualization evolved in the 1980s alongside 3D visualization techniques in scientific communities, while information visualization was developed in the design community in the 1990s, mainly using 2D graphic tools and techniques.[12] The input data can be derived from readings of technical equipment and instruments, from entities that are infinitesimally small or otherwise invisible to human senses, and the line between a "physical phenomenon" and information or data in general can be hard to draw. Here, it is also relevant to mention the field of knowledge visualization: emphasis on the use of visualizations as facilitation in collaborative, knowledge producing processes and contexts, for instance in management studies.[13] Visualizations have also been used in scheduling and in industry

---

[10] N. Katherine Hayles, *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics* (Chicago, IL: University of Chicago Press, 2008); James Gleick, *The Information: A History, a Theory, a Flood* (New York: Pantheon Books, 2011); Gunther R. Kress and Theo van Leeuwen, *Reading Images: The Grammar of Visual Design* (London: Routledge, 2006), 50–51; Otfried Czaika, Jonas Nordin, and Pelle Snickars, *Information som problem: Medieanalytiska texter från medeltid till framtid* (Stockholm: Kungliga biblioteket, 2014); John Durham Peters, *Speaking into the Air: A History of the Idea of Communication* (Chicago, IL: University of Chicago Press, 1999); James R. Beniger, *The Control Revolution: Technological and Economic Origins of the Information Society* (Cambridge, MA: Harvard University Press, 1986).

[11] McNabb and Laramee, "Survey of Surveys (SoS)."

[12] Manovich, "What Is Visualization?" 38.

[13] Martin J. Eppler, "What Is an Effective Knowledge Visualization? Insights from a Review of Seminal Concepts," in *Knowledge Visualization Currents*, eds. Francis T. Marchese and Ebad Banissi (London: Springer, 2013), 3–4.

production planning since the early 1900s, for instance with the well-known Gantt chart and its many succeeding adaptations.[14]

Only from the 1970s and onwards did the term "visualization" become synonymous with visual representations on paper or on screens.[15] Information visualization, as forms of nontextual representation, is closely related to "information graphics" or "graphical representations" and the concept of the diagram, referring to a written, drawn, or painted trace on a surface.[16] Certain two-dimensional pictures, like maps, possess the capacity to harbor large amounts of information in one single sheet of paper, or equivalent.[17] The pictorial units can be called "graphisms," inscriptions, or traces—graphic marks on a surface.[18] As I will return to, the *mark* is the established term to describe graphic units in information visualization, in handbooks, in academic discourse, and by practitioners.[19]

Tamara Munzner's handbook and survey *Visualization Analysis and Design* (2014) defines visualization without any information or data prequel. Here, a user- and technology-oriented definition is given: "Computer-based visualization systems provide visual

[14] Karolina Uggla and Yvonne Eriksson, "Visualization of Production Planning," in *Proceedings of the 2019 23rd International Conference Information Visualisation (IV)* (IEEE, 2019), 312–317, https://doi.org/10.1109/IV.2019.00059.

[15] Colin Ware, *Information Visualization: Perception for Design* (Boston, MA: Morgan Kaufmann, 2012), 2.

[16] W. J. T. Mitchell, *Iconology: Image, Text, Ideology* (Chicago, IL: University of Chicago Press, 1986), 29; Christoph Lüthy and Alexis Smets, "Words, Lines, Diagrams, Images: Towards a History of Scientific Imagery," *Early Science and Medicine* 14, no. 1 (April 1, 2009): 401, 409–410; Drucker, *Graphesis*; Klaus Hentschel, *Mapping the Spectrum: Techniques of Visual Representation in Research and Teaching* (Oxford: Oxford University Press, 2002).

[17] Bruno Latour, "Drawing Things Together," in *Representation in Scientific Practice*, eds. Michael Lynch and Steve Woolgar (Cambridge, MA: MIT Press, 1990), 26–28. In this essay, Latour also addresses pictures as "immutable mobiles," as units that can easily be moved, spread without loss of information, thus enhancing knowledge and also power.

[18] Latour, 36; Krämer, *Medium, Messenger, Transmission*, 174–175, 187.

[19] Jacques Bertin, *Semiology of Graphics: Diagrams, Networks, Maps* (Madison, WI: University of Wisconsin Press, 1983), 42; Tamara Munzner, *Visualization Analysis and Design* (Boca Raton, FL: CRC Press, 2014), 114.

representation of datasets designed to help people carry out tasks more effectively."[20] Visualizations are seen as objects that can facilitate users' decision-making processes and thereby function as extensions of human cognitive skills or, in Munzner's words, to "augment human capabilities."[21]

Some of the literature uses "data" instead of information visualization. Generally, data is seen as the raw material of information, which is processed and organized data, stepwise proceeding to a visualization.[22] Data, like information, as a concept, earned extended meaning and use in relation to the first digital revolution.[23] Data as a rhetoric term is derived from the Latin *dare*, something given in an argument.[24] Data is thus termed the units or the building blocks of information, as items that precede information. For instance, Andy Kirk's *Data Visualisation* (2019), a handbook aimed at designers, uses a pragmatic approach to the concept of data. In the introductory glossary, a problematized view of the notion of raw data is acknowledged and appreciated, but is not in focus since it is said not to affect user comprehension of visualizations.[25] Kirk uses a concise definition of data visualization: "The visual representation and presentation of data to facilitate understanding."[26] Apart from the pragmatic approach to data as something collected or given, not yet subject to any intervention, sorting or design choice, there is a distinction between representation and presentation within this definition. Presentation is seen as a way to "portray your data visually" and, for instance, the different options of chart types. Representation is all that falls beyond this: design choices making the visual representation accessible to the intended users.[27]

---

[20] Munzner, *Visualization Analysis and Design*, 1.
[21] Munzner, *Visualization Analysis and Design*, 1.
[22] Alberto Cairo, *The Functional Art: An Introduction to Information Graphics and Visualization* (Berkeley, CA: New Riders, 2013). For a discussion on "Big Data", see Andersson Schwarz in this volume.
[23] Daniel Rosenberg, "Data Before the Fact," in *"Raw Data" Is an Oxymoron*, ed. Lisa Gitelman (Cambridge, MA: MIT Press, 2013), 30.
[24] Rosenberg, "Data Before the Fact," 18.
[25] Kirk, *Data Visualisation*, 11.
[26] Kirk, *Data Visualisation*, 15.
[27] Kirk, *Data Visualisation*, 17–19.

Information into visualization

One theme in the literature when defining information visualization is to identify steps in the process of decoding and coding data or information into visualizations. In a survey of surveys of academic writing on information visualization, "The Information Visualization Pipeline model" is introduced: "The pipeline describes the transition of raw data into a visualization which is visible to a user." This consists of five identified steps: "1. Data Enhancement & Transformation, 2. Visual Mapping & Structure, 3. Exploration & Rendering, 4. Interactive Analysis and 5. Perception."[28] The process is one of transformations: So-called "raw data" has to be "captured" and structured into tables. The second stage is "Visual Mapping & Structure," where choice of visualization technique, and questions of which "visual primitives" such as color, texture, and geometry, are mapped onto chosen data sets. Stage three, "Exploration & Rendering," also takes on the question of visual presentation to the user, but has more to do with viewpoints, scaling, and presentation to the user, how to handle distortions and projections. The fourth stage, "Interactive Analysis," consists of surveys that address the question of how feedback from users is recorded. The level of interactivity can diverge significantly from different information visualization types and projects.[29] The habit of dividing the process of visualizing information into different stages is part of how to define and interpret information visualization. Colin Ware uses it in *Information Visualization: Perception for Design* (2012), starting out with: "1) The collection and storage of data itself, 2) The preprocessing designed to transform the data into something we can understand, 3) The display hardware and the graphics algorithms that produce an image on the screen, 4) The human perceptual and cognitive system (the perceiver)."[30]

Methods and tools that facilitate visualization to generate an overview, to expose trends and patterns in large amounts of data, have now become more available and more user friendly.[31] Johanna

---

[28] McNabb and Laramee, "Survey of Surveys (SoS)," 590–591.
[29] McNabb and Laramee, "Survey of Surveys (SoS)," 590–591.
[30] Ware, *Information Visualization*, 4–5.
[31] Johanna Drucker, "Graphical Approaches to the Digital Humanities," in *A New Companion to Digital Humanities*, eds. Susan Schreibman,

Drucker also has critiqued the use of the word data as something already "given" and introduces the term "capta"—"constructed and not given."[32] She stresses that humanities scholars, who are used to subjects laden with "uncertainty and ambiguity," should be more aware of problems that can occur when transferring qualitative information into numerical data to be visualized. Such transfer risks creating unnecessary dichotomies. Some variables are disputed and problematic and would need to be represented as fuzzy and nuanced, not definitive. Drucker exemplifies with the concepts of gender and that of nation, both common dividing categories in demographical statistics, but not static if viewed across time, culture, and who makes the definition. Data or information visualization can, to follow Drucker, risk bringing in simplified formats into disciplines that are used to discussing problems of any given category taken as unproblematic.[33]

Ideally, there is a discrete relation between information, or data, in the form of numerical value, and the visual elements—point, line, or surface—of a visualization. This is also the core argument of Manovich's preliminary definition of information visualization: "a mapping between discrete data and a visual representation."[34] Information visualization, then, forms a specific representational type which deviates from many other visual representations. In the essay that was essential to the development of visual semiotics, "Rhetoric of the Image" (1964), Roland Barthes stated that

---

Raymond George Siemens, and John Unsworth (Chichester: John Wiley & Sons, 2016), 238–250; Drucker, *Graphesis*.

32 Drucker, "Graphical Approaches to the Digital Humanities," 244–245; danah boyd and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication & Society* 15, no. 5 (June 2012): 662–679, http://doi.acm.org/10.1080/1369118X.2012.678878; Katherine Hepworth, "Big Data Visualization: Promises & Pitfalls," *Communication Design Quarterly* 4, no. 4 (March 2017): 7–19, http://doi.acm.org/10.1145/3071088.3071090.

33 Drucker, "Graphical Approaches to the Digital Humanities." See also Catherine D'Ignazio and Lauren F. Klein, *Data Feminism* (Cambridge, MA: MIT Press, 2020), 109, for a discussion on the fluidity of the categories of gender, which should, but rarely does, affect the construction of for instance surveys.

34 Manovich, "What Is Visualization?" 37.

images are "polysemic." Meaning ascribed to images is arbitrary depending on context, time, and place, that is, images can mean different things to different viewers at different occasions.[35]

There is a strong tradition within the design-oriented literature on information visualization of categorizing it as a more monosemic representational type than other visual representations. That means that, as opposed to the polysemic or multiple possible meanings of most visual representation, it is narrowed down to having only one possible meaning. In his major work *Sémiologie graphique: Les diagrammes, les réseaux, les cartes* from 1967, cartographer Jacques Bertin states that graphics—diagrams, networks, and maps—are monosemic sign systems, where each sign represents only one specified meaning. In maps this can be specified in a legend, but Bertin also stresses the agreement on visual meaning within a community.[36] Bertin's writings have been influential in the formation of the theoretical field of information visualization.[37] Thereby, a sometimes simplified view that information visualization forms a pictorial language of truth can be present in handbooks of design of information visualization.[38]

When it comes to the interpretation of information visualization as visual objects, the specificity of the way that these objects represent can provide some difficulties. First, their truth aspect can stand in the way of scrutinizing them. Second, they demand of the viewer to be able to decipher their accuracy in their representation of data. These aspects are not at all insignificant, but they risk keeping scholars from other fields, such as art history and visual studies, away from interpreting their formal visual properties.

---

[35] Roland Barthes, "Rhetoric of the Image," in *Image, Music, Text*, trans. Stephen Heath (London: Fontana, 1977), 32–51.

[36] Bertin, *Semiology of Graphics*; Cairo, *The Functional Art*; Alberto Cairo, *The Truthful Art: Data, Charts, and Maps for Communication* (San Francisco, CA: New Riders, Peachpit, 2016); Karolina Uggla, "Voir ou lire: Maps as Art – Art as Maps," in *Art Theory as Visual Epistemology*, ed. Harald Klinke (Cambridge: Cambridge Scholars Publishing, 2016), 95–110.

[37] Munzner, *Visualization Analysis and Design*, 114, 175.

[38] Per Mollerup, *Data Design: Visualising Quantities, Locations, Connections* (New York: Bloomsbury Visual Arts, 2015).

Designerly approaches to information visualization

A couple of handbooks aimed at designers have already been mentioned, and there are plenty of others ranging between an engineering audience, guides to different tools and software, and those targeting visual designers and artists.[39] The research field of information visualization is expanding, and it can be fairly practical, oriented toward design practice and technologies, when it comes to data collection, enhancement, classification, visualization techniques, rendering, and user feedback. In the different stages of the visualization process, transforming data into a visual form that a user can comprehend, questions of design are central. This has resulted in some very visually compelling textbooks and accounts of what information visualization is and can be, showing significant, beautiful, and fascinating examples. Such examples, which started out as popular blogs or online sites, are designer and journalist David McCandless's blog beautifuldata.com, and, likewise, designer Manuel Lima's blog visualcomplexity.com. The former has written books on data journalism.[40] Lima's beautifully illustrated surveys of historical and recent examples of information visualization draw a line from antiquity to early modern times to contemporary digital information visualizations.[41] There

---

[39] Cairo, *The Functional Art*; Cairo, *The Truthful Art*; Stephen Few, *Now You See It: Simple Visualization Techniques for Quantitative Analysis* (Oakland, CA: Analytics Press, 2009); Stephen Few, *Show Me the Numbers: Designing Tables and Graphs to Enlighten* (Burlingame, CA: Analytics Press, 2012); Kirk, *Data Visualisation*; Mollerup, *Data Design*; Munzner, *Visualization Analysis and Design*; Robert Spence, *Information Visualization: An Introduction*, 3rd ed. (New York: Springer International Publishing, 2014); Matthew Ward, Georges Grinstein, and Daniel Keim, *Interactive Data Visualization: Foundations, Techniques, and Applications* (Boca Raton, FL: CRC Press, Taylor & Francis Group, 2015); Ware, *Information Visualization*. This is by no means a comprehensive list of handbook literature but some often-used examples.

[40] David McCandless, *Information Is Beautiful* (London: Collins, 2012); David McCandless, *Knowledge Is Beautiful* (New York: HarperCollins Publishers, 2014); Anna-Lena Carlsson, "The Aesthetic and the Poietic Elements of Information Design," in *2010 14th International Conference Information Visualization* (Los Almitos, CA: IEEE, 2010), 450–454, https://doi.org/10.1109/IV.2010.69. See Carlsson for a discussion on the separation of "form and function" in the information design discourse.

[41] Manuel Lima, *Visual Complexity: Mapping Patterns of Information* (New York: Princeton Architectural Press, 2011); Manuel Lima, *The*

are various examples of publications that are richly illustrated. When it comes to historical perspectives Sandra Rendgen and Julius Wiedemann's folios *Information Graphics* (2012) and *Understanding the World* (2014) are prominent examples that also include a historical perspective.[42] There are examples of scholarly exposés nurtured by the rich visual history of information visualization, or its earlier equivalents.[43] The aims of such work is not to provide the reader with an analytical framework but it is always hard to make this kind of visual odyssey transferable. It requires expert knowledge to move around in the history of visualization. As contextualizations, this kind of work is very welcome and elegantly made. Another direction to strengthen the design of information visualization is to focus strictly on human perception—how humans perceive visual information—as in Colin Ware's *Visual Thinking for Design* (2008), and the more accessible *Information Visualization: Perception for Design* (2000).[44]

A predecessor in popularizing information or statistical graphics, not using the term information or data visualization, is designer Edward Tufte. Many of his books compile visually attractive examples of information visualization from different epochs and geographical places. As Alberto Cairo writes in *The Functional Art* (2013), Tufte's writing can be hard to follow, since it forms rather subjective surveys of different examples of excellence in information visualization, juxtaposed with examples of less fortunate ones.[45] His committed and fluid narrative makes it hard for the reader to single out the analytical methods used. Tufte has conveyed opinions on what constitutes excellence in graphics, in the mission of spreading the word of how to envision information

---

Book of Trees: Visualizing Branches of Knowledge (New York: Princeton Architectural Press, 2014); Manuel Lima, *The Book of Circles: Visualizing Spheres of Knowledge* (New York: Princeton Architectural Press, 2017).

42 Sandra Rendgen and Julius Wiedemann, *Information Graphics* (Cologne: Taschen, 2012); Sandra Rendgen and Julius Wiedemann, *Understanding the World: The Atlas of Infographics* (Cologne: Taschen, 2014).

43 Drucker, *Graphesis*; Rosenberg and Grafton, *Cartographies of Time*; Kostelnick and Hassett, *Shaping Information*; Bender and Marrinan, *The Culture of Diagram*; Boyd Davis, "Beholder of All Ages."

44 Ware, *Information Visualization*; Colin Ware, *Visual Thinking for Design* (Burlington, MA: Morgan Kaufmann, 2008).

45 Cairo, *The Functional Art*, 64.

in better ways.[46] Some of his terms have become commonplace in speaking about and assessing information visualization and capture a concentrate of some of his thought. Tufte advocates a minimalist, less-is-more standpoint to the design of statistical graphics. The concept of "chart-junk" denotes all superfluous visual elements a designer uses to embellish statistical graphics, out of which some can be derived from convention and routine.[47] The second widely disseminated concept of Tufte's, which also is a potentially analytical one, is "data–ink ratio." This concept denotes the amount of "ink" necessary to visually represent data, for instance records from measuring equipment, divided by "total ink used to print the graphic." The overall principle is to reduce and erase all unnecessary ink.[48] The impact Tufte has had on general knowledge on information graphics, statistical graphics, and information visualization, to the general public, visual journalism, and designers cannot be overlooked. There is, however, a sometimes overexplicit minimalist incentive of visual reductionism that can overshadow the usefulness of an analytical tool like the "data–ink ratio."

Cairo, who has a background as a visual journalist and designer of information graphics, now in US higher education, is a prominent figure in design of information visualization through *The Functional Art* (2013) and *The Truthful Art* (2016), both mainly aimed at designers with lots of practical advice.[49] His latest book is a contribution to an ongoing debate on how to interpret and validate charts and graphs in news media, *How Charts Lie* (2019), in an era of "fake news."[50] Since the design handbook perspective

---

[46] Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, CT: Graphics Press, 2001), 51. "Principles of graphical excellence" are for instance "that which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space."

[47] Tufte, *The Visual Display*, 107.

[48] Tufte, *The Visual Display*, 93–97. D'Ignazio and Klein, *Data Feminism*, 76–77 discuss Tufte's advocacy of minimalism as representative of a larger problem of a false and also gendered binary between reason and emotion in information graphics and visualization.

[49] Cairo, *The Functional Art*; Cairo, *The Truthful Art*.

[50] Alberto Cairo, *How Charts Lie: Getting Smarter about Visual Information* (New York: W. W. Norton & Company, Inc., 2019).

is in focus in his first two books, there is not much room for presenting tools for analysis. However, there is one such, "the information graphics wheel," which is developed to assess qualities in information graphics, to help the designer during the design process, and for stakeholders to value commissioned work. The wheel model presents ways to find a balance between antipodes like density and lightness, figuration and abstraction, functionality and decoration. Cairo calls it "a visualization for planning visualizations." Hereby, it is clearly stated that it is not meant to be used as an academic analytic tool.[51] However, since analytical tools for evaluating information visualization are so rare, this could also form a first step of developing such analytical categories.

I have singled out terminology in naming different graphical elements in information visualization as an area of interest in analyzing information visualization. These also constitute the building blocks that designers have at their disposal. As earlier mentioned, the conceptual hierarchy and terminology in Jacques Bertin's *Semiology of Graphics* has laid a foundation for the naming of graphic building blocks upon which visualization literature still depends.[52] Bertin's full terminology is too extensive to recapitulate here, but the graphic mark is at the center of his system, as it is in Munzner's clarifying overview. The mark equals one item in visualization, forming its "basic visual element." Marks can, in turn, be defined by the number of dimensions they occupy on a two-dimensional plane: zero dimensions constitutes a point, one dimension is a line, and two is an area.[53] Other factors that determine a mark on the visual plane are called visual variables by Bertin, and channels by Munzner, which is the established term in the visualization literature. Bertin and Munzner refer to eight different channels, basically corresponding but with some variation in terminology. These channels regard the mark's horizontal or vertical placement on the plane, its size, value, texture, color,

[51] Cairo, *The Functional Art*, 50–51.

[52] Bertin, *Semiology of Graphics*; Munzner, *Visualization Analysis and Design*, 114, 175.

[53] Munzner, *Visualization Analysis and Design*, 95–96.

orientation, or shape.[54] In addition, there can be relations between marks, in visualization types like network diagrams. Here, a mark can be either an individual item, a node—a point, or a connector—usually a line.[55] These are only the most basic categorizations of building blocks in information visualization, following the established terminology extracted from Bertin. They are useful when understanding and assessing these visual objects, providing primary insights into the vocabulary of the field.

Interpretative Framework 1: Semiotics and Information Visualization

One way of analyzing information visualization is to establish a grammar of its visual building blocks by using terms and tools from semiotics.[56] The building block, or element, in semiotics is the sign, and its grammar is about "sign-making" and visual entities like direction, color, and placement.[57] Gunter Kress and Theo van Leeuwen do not address information visualization *per se* but "abstract visuals," or "diagrams" in terms of realistic and less realistic representations. I have for the purpose of this text focused on some representational types in the archive of examples, like diagrams, charts, and maps, with elements like geometric volumes and vector lines, which are also intrinsic parts of information visualization. The main example in this genre starts out with previously mentioned Claude Shannon's influential communication model from 1948. The recurrent use of simple geometrical shapes is discussed: boxes and arrows in this visual model

---

54 Bertin, *Semiology of Graphics*, 7; Munzner, *Visualization Analysis and Design*, 96.

55 Munzner, *Visualization Analysis and Design*, 100.

56 Kress and van Leeuwen, *Reading Images*, 1–2, 14. The term grammar is used with some reservations. The aim is to analyze combinations of elements and the meanings that can be ascribed to their aggregations in social interactions. It is less concerned with sets of rules. Illuminating power relations and ideological positions that images convey is also an aim, leading to the method of analysis presented to become a contribution to broadening the scope of critical discourse analysis in a visual direction.

57 Kress and van Leeuwen, *Reading Images*, 6–7.

indicate abstract participants and their conjunctive processes and relations, respectively.[58] According to the authors, in a broader cultural field, the fascination with geometrical shapes draws on modernist, abstract art of the early 20th century, associated with being "pure, quasi-scientific 'atoms' of the visible world, a 'pure manifestation of the elements'."[59] These clean, basic forms are seen as the building blocks of the visual grammar, and are compared to linguistic structures and concepts, such as nouns and verbs.[60] Geometric shapes are also associated with "the world of human construction": buildings in modern cities, machines and roads, as opposed to organic or "natural," irregular shapes.[61] Different geometrical shapes, in semiotic terms, represent different categories of things, and/or actions. A vector line can also indicate direction, or the linkage and some kind of transfer between elements. In short, these geometrical shapes and lines help to form a visual narrative.[62] In laying out the differences between vectors and volumes, Kress and van Leeuwen draw on art historian Rudolf Arnheim, who stated that "motion and change" are usually represented as a sequence in Cartesian space either horizontally or vertically, connecting the function of geometrical shapes to the basic categories of being as static, and acting as motion. Being is thus tied to volumes and acting to vectors.[63] Kress and van Leeuwen also address another diagrammatic feature that is important in information visualization, and that is the capability of basic geometric forms to display taxonomy, that is, how different elements—or, in their terminology, "participants,"—relate to each other, hierarchically, in, for instance, tree diagrams, network diagrams, and flowcharts.[64] The whole idea of a "grammar of visual design" is an interesting attempt to identify and name visual elements, applied to many different examples, but the development of an actual method for analysis with some generality is still to be refined.

[58] Kress and van Leeuwen, *Reading Images*, 49–51.
[59] Kress and van Leeuwen, *Reading Images*, 53–54.
[60] Kress and van Leeuwen, *Reading Images*, 74–77.
[61] Kress and van Leeuwen, *Reading Images*, 54–55.
[62] Kress and van Leeuwen, *Reading Images*, 46, 59–60.
[63] Rudolf Arnheim, *The Power of the Center: A Study of Composition in the Visual Arts* (Berkeley, CA: University of California Press,1982), 230.
[64] Kress and van Leeuwen, *Reading Images*, 82–86.

Linked to social semiotics is Wiebke Weber's framework of analysis of data visualization, with its theoretical base in multimodality and social semiotics. Again, visual variables with a semiotic substance that are identified are, for instance, points, lines, areas, and colors. The framework also takes into account the aesthetic form of the visualization. The analytical framework is tested on examples of information visualization in news media, out of a larger corpus. The analysis is divided into graphical and textual modes and ideational, interpersonal, and compositional metafunctions, that is: the types of messages in the visualization, the relation to the audience, roles and relations within the visualization, and, finally, its overall composition.[65] My own orientation in this present text, especially when categorizing previous research and writing on information visualization, owes a lot to Weber's study. The latter further exposes some of the complexity and arduousness of building an analytical framework for information visualization. In the shorter format, the description of the analytical framework and its terminology risks appearing heavy in comparison the case studies. However, it shows the kind of rigor that is necessary when analyzing information visualization, taking into consideration a multitude of interpretative layers.

Interpretative Framework 2: Emotions and Ethics in Information Visualization

In the late 2010s there appeared signs of a paradigmatic shift in the way information visualization was produced and received. Charles Kostelnick, a professor with a background in comparative literature, is one of the founders of writing about information visualization and other forms of diagrammatic representations, together with Michael Hassett in *Shaping Information* (2003).[66] He observed how "emotional appeal" has returned as a dominant aspect in data visualization (the term used) in the 2010s. The shift contrasts greatly to Tufte's preeminent minimalist ideal,

[65] Wibke Weber, "Towards a Semiotics of Data Visualization – An Inventory of Graphic Resources," in *2019 23rd International Conference Information Visualization (IV)* (IEEE, 2019), 323–328, https://doi.org/10.1109/IV.2019.00061.

[66] Kostelnick and Hassett, *Shaping Information*.

and this ideal is the main antagonistic stance in Kostelnick's exposition. Kostelnick analyzes examples of data visualization using concepts derived from Aristotelian rhetoric such as *pathos* and *enargeia*, but also produces a background in form of a historical excursion, showing that the emotional appeal was vital in the late Victorian era, equivalent to data visualization, and diminished during 20th-century modernism.[67] The "emotional appeal," according to Kostelnick, not only appears in design features such as color and "design novelty"; recent technology has also made it possible to put the user in the center and add interactivity such as personalized displays, such as visualizing real-time data, personal data, and options of selection of which data and range to view.[68]

Following the discourse of emotions in information visualization, both in the ranks of designers and as a response in users, there is an orientation toward ethical information visualization, which came to my knowledge via Katherine Hepworth. In the paper "Racism in the Machine" (2018), Hepworth, together with Christopher Church, analyzes and compares two examples of web-based, interactive mapping projects, both dealing with the atrocity of lynchings in the US in the 19th and 20th centuries. In the analysis, the projects and their visual and interactive characteristics are carefully described.[69] Also, it is quite unusual in the writings on information visualization that colors, contrast, shape, and typography is considered. However, the study also gives at hand some of the difficulties of conducting a formal analysis of interactive information visualization, where "dots," "zoom," and "click" are recurrent. Without being able to go into detail of the analysis, the main question is how these chosen projects manage to expose the victims of these cruel acts without dehumanization. Lives lost risk becoming mere statistics in these kinds of representations if, as Church and Hepworth say, "the

67 Charles Kostelnick, "The Re-Emergence of Emotional Appeals in Interactive Data Visualization," *Technical Communication* 63, no. 2 (May 2016): 116–135.

68 Kostelnick, "The Re-Emergence of Emotional Appeals," 117.

69 Katherine Hepworth and Christopher Church, "Racism in the Machine: Visualization Ethics in Digital Humanities Projects," *Digital Humanities Quarterly* 12, no. 4 (February 4, 2019), unpaginated.

dignity of the represented subjects" is not carefully defended.[70] The large claim in the paper is the proposition of an "ethical visualization workflow." To recall, for instance, the "the Information Visualization Pipeline model," literature on information visualization is preoccupied with describing processes of a visualization project.[71] Contrary to such plans and descriptions, the "ethical visualization workflow" starts on the premises that visualization of data or information is not neutral. Thereby, the work to ensure ethics in a visualization project start "pre-data collection," studying up-to-date academic work on the subject, to avoid bias, and similarly regarding the design choices. For instance, some conventional forms frequently used in information visualization have built-in problems, like the Mercator projection, proportionally favoring the northern hemisphere.[72] Lastly, the expanding field of digital humanities is addressed. The criticality inherent in humanist subjects is needed throughout the whole process.[73] Scholars of the human and social sciences do have a role to play in the understanding and shaping of information visualization.

Conclusion

This chapter was first an attempt to map out some parts of the scholarly fields and literature on information visualization in order to illuminate some of its tools and terminology to scholars outside of those fields. The writings on information visualization are disseminated across different areas of knowledge, aimed at different readers. There are handbooks aimed at designers, and there is the academic field, which can be broken down into converging orientations, for instance concerning the visual design and the technologies of transferring data or information into visualizations. From these fields I highlighted some elemental terminology of entities of information visualization from well-known and widely used sources, like Jacques Bertin and Edward Tufte.

70 Hepworth and Church, "Racism in the Machine."
71 McNabb and Laramee, "Survey of Surveys (SoS)."
72 Hepworth and Church, "Racism in the Machine." D'Ignazio and Klein, *Data Feminism*, 205–207, list other examples of methods and community networks of securing design processes of mapping and visualization.
73 Hepworth and Church, "Racism in the Machine."

Second, to exemplify, I selected two interpretative frameworks or ways to analyze information visualization, a framework inspired by social semiotics represented by Gunther Kress and Theo van Leeuwen and Wiebke Weber, and a framework circling around the emotions and ethics of information visualization, represented by writings by Charles Kostelnick and Katherine Hepworth. Kostelnick uncovers what he sees as a trend in information visualization projects in the 2000s. Hepworth and Church focus on the safeguarding of the design process, making it ethical in every step, most notably in the phase of data or information collection.

As a visual object of study, information visualization becomes even more curious and ungraspable when we are faced with interactive graphics and real-time-generated visualizations, or interactivity graphics that can be altered by the user, and thereby personalize the experience.[74] Real-time-generated visualizations are ephemeral visual objects that challenge methods of capture and documentation, and this discussion has not been included in this chapter. When more decision-making is based on information that is made available to different users as visualizations, these are visual objects that researchers in the human and social sciences need to take into consideration, and to develop methods for analysis. Studying visual representations on screen is nothing new to art historians and scholars of visual culture, but there is a need for a methodological discussion on how to handle study objects that change over time or with the input of users.

Neither of these two identified frameworks is in itself a fully developed methodology for analyzing information visualization as visual objects, or aims to be so. But some components do provide valuable terminology and analytical concepts when working with information visualization in the humanities and the social sciences. They could constitute a welcome start of a broader discussion on how scholars of for instance art history and visual studies can influence these fields in offering more apt and subtle tools for visual analysis of visual representations that are not always seen as such but nonetheless have an impact on how we see and interpret the world around us.

74 Kostelnick, "The Re-Emergence of Emotional Appeals," 117, 125–127.

References

Arnheim, Rudolf. *The Power of the Center: A Study of Composition in the Visual Arts*. Berkeley, CA: University of California Press, 1982.

Barthes, Roland. “Rhetoric of the Image.” In *Image, Music, Text*, translated by Stephen Heath, 32–51. London: Fontana, 1977.

Bender, John B., and Michael Marrinan. *The Culture of Diagram*. Stanford, CA: Stanford University Press, 2010.

Beniger, James R. *The Control Revolution: Technological and Economic Origins of the Information Society*. Cambridge, MA: Harvard University Press, 1986.

Bertin, Jacques. *Semiology of Graphics: Diagrams, Networks, Maps*. Madison, WI: University of Wisconsin Press, 1983.

boyd, danah, and Kate Crawford. “Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon.” *Information, Communication & Society* 15, no. 5 (June 1, 2012): 662–679. https://doi.org/10.1080/1369118X.2012.678878.

Boyd Davis, Stephen. “Beholder of All Ages: The History of the World in a French Mappemonde.” *Textimage: Revue d’étude du dialogue text-image* 7 (July 2015): 1–6.

Cairo, Alberto. “Ethical Infographics.” *IRE Journal* 37, no. 2 (Spring 2014): 25–27.

Cairo, Alberto. *How Charts Lie: Getting Smarter about Visual Information*. New York: W. W. Norton & Company, 2019.

Cairo, Alberto. *The Functional Art: An Introduction to Information Graphics and Visualization*. Berkeley, CA: New Riders, 2013.

Cairo, Alberto. *The Truthful Art: Data, Charts, and Maps for Communication*. San Francisco, CA: New Riders, Peachpit, 2016.

Card, Stuart K., Jock D. Mackinlay, and Ben Shneiderman. “Information Visualization.” In *Readings in Information Visualization: Using Vision to Think*, 1–34. San Francisco, CA: Kaufmann, 1999.

Carlsson, Anna-Lena. “The Aesthetic and the Poietic Elements of Information Design.” In *2010 14th International Conference*

Information Visualization, 450–454. Los Almitos, CA: IEEE, 2010. https://doi.org/10.1109/IV.2010.69.

Czaika, Otfried, Jonas Nordin, and Pelle Snickars. *Information som problem: Medieanalytiska texter från medeltid till framtid*. Stockholm: Kungliga biblioteket, 2014.

D'Ignazio, Catherine, and Lauren F. Klein. *Data Feminism*. Cambridge, MA: MIT Press, 2020.

Drucker, Johanna. *Graphesis: Visual Forms of Knowledge Production*. Cambridge, MA: Harvard University Press, 2014.

Drucker, Johanna. "Graphical Approaches to the Digital Humanities." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Raymond George Siemens, and John Unsworth, 238–250. Chichester: John Wiley & Sons, 2016.

Eppler, Martin J. "What Is an Effective Knowledge Visualization? Insights from a Review of Seminal Concepts." In *Knowledge Visualization Currents*, edited by Francis T. Marchese and Ebad Banissi, 3–12. London: Springer, 2013.

Eriksson, Yvonne. *Bildens tysta budskap: Interaktion mellan bild och text*. Lund: Studentlitteratur, 2017.

Few, Stephen. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Oakland, CA: Analytics Press, 2009.

Few, Stephen. *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Burlingame, CA: Analytics Press, 2012.

Foka, Anna, and Stefan Gelfgren. "Visualisering som verktyg och metod för historieforskning." In *Digital humaniora: Humaniora i en digital tid*, edited by Per-Olof Erixon and Julia Pennlert, 147–164. Gothenburg: Daidalos, 2017.

Gleick, James. *The Information: A History, a Theory, a Flood*. New York: Pantheon Books, 2011.

Hayles, N. Katherine. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago, IL: University of Chicago Press, 2008.

Hentschel, Klaus. *Mapping the Spectrum: Techniques of Visual Representation in Research and Teaching*. Oxford: Oxford University Press, 2002.

Hepworth, Katherine. "Big Data Visualization: Promises & Pitfalls." *Communication Design Quarterly* 4, no. 4 (March 2017): 7–19. https://doi.org/10.1145/3071088.3071090.

Hepworth, Katherine, and Christopher Church. "Racism in the Machine: Visualization Ethics in Digital Humanities Projects." *Digital Humanities Quarterly* 12, no. 4 (February 4, 2019): unpaginated.

Kirk, Andy. *Data Visualisation: A Handbook for Data Driven Design*. Los Angeles, CA: Sage Publications, 2019.

Kostelnick, Charles. "The Re-Emergence of Emotional Appeals in Interactive Data Visualization." *Technical Communication* 63, no. 2 (May 2016): 116–135.

Kostelnick, Charles, and Michael Hassett. *Shaping Information: The Rhetoric of Visual Conventions*. Carbondale, IL: Southern Illinois University Press, 2003.

Krämer, Sybille. "Trace, Writing, Diagram: Reflections on Spatiality, Intuition, Graphical Practices and Thinking." In *The Power of the Image: Emotion, Expression, Explanation*, edited by András Benedek and János Kristóf Nyíri, 3–22. Frankfurt am Main, New York: Peter Lang, 2014.

Krämer, Sybille. *Medium, Messenger, Transmission: An Approach to Media Philosophy*, translated by Anthony Enns. Amsterdam: Amsterdam University Press, 2016.

Kress, Gunther R., and Theo van Leeuwen. *Reading Images: The Grammar of Visual Design*. London: Routledge, 2006.

Latour, Bruno. "Drawing Things Together." In *Representation in Scientific Practice*, edited by Michael Lynch and Steve Woolgar, 19–68. Cambridge, MA: MIT Press, 1990.

Lima, Manuel. *The Book of Circles: Visualizing Spheres of Knowledge*. New York: Princeton Architectural Press, 2017.

Lima, Manuel. *The Book of Trees: Visualizing Branches of Knowledge*. New York: Princeton Architectural Press, 2014.

Lima, Manuel. *Visual Complexity: Mapping Patterns of Information*. New York: Princeton Architectural Press, 2011.

Lüthy, Christoph, and Alexis Smets. "Words, Lines, Diagrams, Images: Towards a History of Scientific Imagery." *Early Science and Medicine* 14, no. 1 (April 1, 2009): 398–439.

Manovich, Lev. "What Is Visualization?" *Visual Studies* 26, no. 1 (March 15, 2011): 36–49. https://doi.org/10.1080/1472586X.2011.548488.

McCandless, David. *Information Is Beautiful*. London: Collins, 2012.

McCandless, David. *Knowledge Is Beautiful*. New York: HarperCollins Publishers, 2014.

McNabb, Liam, and Robert S. Laramee. "Survey of Surveys (SoS) – Mapping The Landscape of Survey Papers in Information Visualization." *Computer Graphics Forum* 36, no. 3 (July 4, 2017): 589–617. https://doi.org/10.1111/cgf.13212.

Mitchell, W. J. T. *Iconology: Image, Text, Ideology*. Chicago, IL: University of Chicago Press, 1986.

Mollerup, Per. *Data Design: Visualising Quantities, Locations, Connections*. New York: Bloomsbury Visual Arts, 2015.

Munzner, Tamara. *Visualization Analysis and Design*. Boca Raton, FL: CRC Press, 2014.

Peters, John Durham. *Speaking into the Air: A History of the Idea of Communication*. Chicago, IL: University of Chicago Press, 1999.

Rendgen, Sandra. *Minard System: The Complete Statistical Graphics of Charles-Joseph Minard*. Cologne: Taschen, 2018.

Rendgen, Sandra, and Julius Wiedemann. *Information Graphics*. Cologne: Taschen, 2012.

Rendgen, Sandra, and Julius Wiedemann. *Understanding the World: The Atlas of Infographics*. Cologne: Taschen, 2014.

Rendgen, Sandra, Julius Wiedemann, David Rumsey, Michael Friendly, Michael Stoll, and Scott Klein. *History of Information Graphics*. Cologne: Taschen, 2019.

Rosenberg, Daniel. "Data Before the Fact." In *"Raw Data" Is an Oxymoron*, edited by Lisa Gitelman, 15–40. Cambridge, MA: MIT Press, 2013.

Rosenberg, Daniel, and Anthony Grafton. *Cartographies of Time*. New York: Princeton Architectural Press, 2010.

Spence, Robert. *Information Visualization: An Introduction*. 3rd ed. New York: Springer International Publishing, 2014.

Tufte, Edward R. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press, 2001.

Uggla, Karolina. "Voir Ou Lire: Maps as Art – Art as Maps." In *Art Theory as Visual Epistemology*, edited by Harald Klinke, 95–110. Cambridge: Cambridge Scholars Publishing, 2016.

Uggla, Karolina, and Yvonne Eriksson. "Visualization of Production Planning." In *Proceedings of the 2019 23rd International Conference Information Visualisation (IV)*, 312–317. IEEE, 2019. https://doi.org/10.1109/IV.2019.00059.

Ward, Matthew, Georges Grinstein, and Daniel Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2015.

Ware, Colin. *Visual Thinking for Design*. Burlington, MA: Morgan Kaufmann, 2008.

Ware, Colin. *Information Visualization: Perception for Design*. Boston, MA: Morgan Kaufmann, 2012.

Weber, Wibke. "Towards a Semiotics of Data Visualization – An Inventory of Graphic Resources." In *2019 23rd International Conference Information Visualization (IV)*, 323–28. IEEE, 2019. https://doi.org/10.1109/IV.2019.00061.

The Growing Pains of Digital Art History: Issues for the Study of Art Using Computational Methods

Amanda Wasielewski

Over the past few decades, a series of "questionnaires" in the art journal *October* has provided a snapshot of new and contentious critical frameworks in art history.[1] While these questionnaires contain their own selection biases,[2] the editorial impulse to take the temperature of trends in both scholarly and artistic circles has nevertheless provided a useful guide to some of the major concerns in the field in recent years. The second questionnaire to be published by the journal—and the first to be called a "questionnaire"—was on visual culture, a term that aims to expand the notion of art history beyond professionalized artistic practice into the realm of visual artifacts produced outside traditional art contexts, for example popular film, television, advertising, and other cultural ephemera that might have a visual component. *October*'s visual culture questionnaire (1996) was preceded by "Questions of Feminism" (1995) and followed by

[1] Previous issues saw "conversations" and "interviews" but the first issue to contain responses to a questionnaire posed by the journal was on the topic of feminism in 1995: "Questions of Feminism: 25 Responses," *October* 71 (1995): 5–48. The first so-called "questionnaire" appeared the following year in 1996: "Visual Culture Questionnaire," *October* 77 (1996): 25–70.

[2] This bias stems from the fact that the journal itself is a historical object and embodies the methodology of a particular period in art history. It was founded in 1976 by Rosalind Krauss and Anette Michaelson as a venue for poststructuralist thought in the English-speaking academe.

How to cite this book chapter:
Wasielewski, Amanda. "The Growing Pains of Digital Art History: Issues for the Study of Art Using Computational Methods." In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 127–151. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.f.

questionnaires on “obsolescence” in artistic practice (2002), the Iraq War (2008), “the Contemporary” (2009), Occupy Wall Street (2012), “materialisms” (2016), monuments (2018), and—most recently—decolonization (2020).[3] This list contains both questionnaires that deal with contemporary political issues as well as issues around theory and methodology.

The impact of digital culture has become an urgent question in the fields of art and art history. In artistic practice, “post-internet” art has gained a foothold in commercial galleries, and there is increased administrative pressure in universities to incorporate digital humanities methodologies into traditional humanities programs.[4] “Digital art history” (DAH) is typically defined as art historical research that uses computational methodologies.[5] Unlike humanities disciplines that deal primarily with analysis of text, art history deals with objects and images. This means that the art historical “data” is more complicated to process and runs aground

[3] George Baker, ed., “Artist Questionnaire: 21 Responses,” *October* 100 (2002): 6–97; Benjamin H. D. Buchloh and Rachel Churner, eds., “Questionnaire: In What Ways Have Artists, Academics, and Cultural Institutions Responded to the U.S.-Led Invasion and Occupation of Iraq?” *October* 123 (2008): 3–184; Hal Foster, ed., “Questionnaire on ‘The Contemporary,” *October* 130 (2009): 3–124; David Joselit and Carrie Lambert-Beatty, eds., “[15 Responses to a Questionnaire on Occupy Wall Street],” *October* 142 (2012): 26–73; David Joselit, Carrie Lambert-Beatty, and Hal Foster, eds., “A Questionnaire on Materialisms,” *October* 155 (Winter 2016): 3–110; Leah Dickerman et al., eds., “A Questionnaire on Monuments,” *October* 165 (August 1, 2018): 3–177; Huey Copeland et al., “A Questionnaire on Decolonization,” *October* 174 (December 1, 2020): 3–125.

[4] It should be noted that an important element of the digital humanities sector is focused on pedagogical tools, which will not be addressed in this essay.

[5] Matthew K. Gold uses the term “algorithmic” instead of computational, while other scholars use the term “computational.” For my purposes, the latter term is more all-encompassing and will be my preferred terminology. For use of these terms see Matthew K. Gold, ed., *Debates in the Digital Humanities* (Minneapolis, MN: University of Minnesota Press, 2012); Susan Schreibman, Raymond George Siemens, and John Unsworth, eds., *A Companion to Digital Humanities* (Malden, MA: Blackwell Publishing, 2004); Anne Burdick et al., *Digital_Humanities* (Cambridge: MIT Press, 2012); Melissa M. Terras, Julianne Nyhan, and Edward Vanhoutte, eds., *Defining Digital Humanities: A Reader* (London: Routledge, 2016).

against complex questions of beauty and aesthetic experience that have haunted the field since its inception. These special issues therefore necessitate a separate discussion of digital humanities methodologies in art history under the rubric of digital art history.

Computational methodologies in art history typically entail quasi-quantitative approaches for studying artists and artwork including image recognition, network analysis, data mining, machine learning, mapping, visualization, and digital reconstruction.[6] In all of these methods, computer software is used to process a set of data, whether that is images, text, dates, locations, or any other metric. In some cases, the term digital art history has also been used as a term for the *study* of digital art (that is, art made in digital media) or a catch all term for both the study of digital art and the use of digital methods.

Given the ongoing debates around digital methodologies and the centrality of text and image databases for contemporary scholarly work, why has *October* not done a questionnaire on "digital art history"? While the "materialisms" questionnaire circled around post-internet art (in the context of a "materialist turn"),[7] questions of methodology and digitization have gone largely unremarked in mainstream art history scholarship.[8] Up until now, it has primarily been left to art institutions (that is,

---

6 Quasi-quantitative methods would perhaps make up the bulk of the data used in digital humanities methodologies. I would define quasi-quantitative data as any numerical data that comes from qualitative interpretation. Humanities scholarship, by definition, is primarily concerned with qualitative data points.

7 "Post-internet" is a term coined by Marisa Olson in 2008 to describe artistic practices that come out of a society where the internet is omnipresent. Olson says, "I think it's important to address the impacts of the internet on culture at large, and this can be done well on networks but can and should also exist offline." The fact that the art world has largely ignored digital, internet, and computer art until it had material manifestations (or implications for new materialisms) is remarkable insofar as dematerialization has long been a key facet of digital art (despite the impossibility of true dematerialization). Régine Debatty, "Interview with Marisa Olson," *We Make Money Not Art* (blog), March 28, 2008, http://we-make-money-not-art.com/how_does_one_become_marisa; Artie Vierkant, "The Image Object Post-Internet," *Jstchillin* (blog), 2010, http://jstchillin.org/artie/pdf/The_Image_Object_Post-Internet_a4.pdf.

8 Joselit, Lambert-Beatty, and Foster, "A Questionnaire on Materialisms."

archivists, librarians, curators, *et cetera*) and non-art historians to address the study of art and visual culture via computational methodologies. Attempts by non-art historians, in particular, to study art in this way has hindered acceptance of these methodologies as legitimate tools for art historical scholarship.[9] While scholars such as Harald Klinke, a founding editor of the *Journal of Digital Art History*, often claims that DAH is transforming the discipline of art history at large, there is scant evidence of this in mainstream art history journals.[10]

Art history is a conservative discipline that, more than most areas of study in the humanities, continues to police its boundaries in both subject matter and methodology.[11] Keeping a critical outlook is certainly important for scholars of new technology, given the pervasiveness of Silicon Valley's techno-utopian ideology, but completely ignoring its political, cultural, and societal impact is an untenable position for art history scholarship.[12] It is certainly time for art history to reckon with "the digital" as more than a subfield or a medium. This chapter aims to interrogate how we might begin to look at "digital art history" from within the field and how digital methods of art history scholarship square with digital (and postdigital) *practices* of artists.

---

[9] The work of a computer scientist at Rutgers University, Dr. Ahmed Elgammal, is an example of one such project, run by non-art historians, that would benefit from a deeper knowledge of the discipline. This is the type of project that Claire Bishop focuses her critique on in "Against Digital Art History." "Digital Humanities Laboratory at Rutgers," https://sites.google.com/site/digihumanlab/home; Claire Bishop, "Against Digital Art History," *International Journal for Digital Art History*, no. 3 (July 27, 2018): 122–131.

[10] See Harald Klinke, "The Digital Transformation of Art History," in *The Routledge Companion to Digital Humanities and Art History*, ed. Kathryn Brown (New York: Routledge, 2020), 32–42; Harald Klinke, "Big Image Data within the Big Picture of Art History," *International Journal for Digital Art History*, no. 2 (October 18, 2016): 14–37; Anna Dahlgren and Amanda Wasielewski, "The Digital U-Turn in Art History." Submitted.

[11] Klinke affirms this widely held view but argues against it in "The Digital Transformation of Art History," 33.

[12] One has only to read mainstream media studies scholarship on internet culture, the gig economy, and other topics to see that interest in new technology by no means signals lack of criticality.

Interest in the digital humanities and digital art history is often confined to nonmodern, noncontemporary, and non-Western specialties. It is somewhat ironic that ancient and medieval art—the oldest art we study—seems to lend itself to digital humanities methodologies, whereas scholars of modern and contemporary art largely ignore such techniques (with the notable exception of scholars of photography and film). There are many possible explanations for this, including the availability of data and the clear delineation of the art world and artistic practice in modern times. However, now that there have been some 50 years of artwork created with the aid of software and computer technology, digital methodologies can also be tested on objects within their own structural paradigm. How might the terms of engagement shift when digital methods are used to study born-digital works of art versus artwork created with predigital materials? Despite the methodological fundamentalism of individual scholarly inquiry, different methodologies—including the computational—have a role to play in understanding both pre- and postdigital artistic practice.

Learning from the Visual Culture Debate

Looking back at *October*'s "visual culture" questionnaire from 1996, it seems that many of the respondents were wringing their hands over nothing—at least, nothing as serious as they make it out to be. The study of visual culture has been folded into art history departments without the destruction and dismantlement that scholars seemed to fear in the mid-'90s. Much like computational methodologies today, visual culture studies are popular with historians of photography, medievalists, and specialists in non-Western art. This is hardly surprising, as these specialties were some of the latest additions to the discipline and, so, perhaps the less attached to canonical/exclusive definitions of art.[13] Despite the alarm raised

[13] Photography, medieval art, and non-Western art were, for example, not considered part of the discipline of art history until the 20th century. Contemporary art, too, was not considered part of art history but, rather, art criticism. The reasons for this are, however, different from the reasons for the exclusion of those other specialties.

in the pages of *October*, scholarship on "visual culture" has not spelled the death of art history, or, indeed, the term "art" itself. Likewise, there is no obsessive preference in art history and visual culture departments for the "visual" over nonvisual artistic practices.[14] Given that the internet had recently arrived for a broader public at the time of writing, it follows that some of the writers in the visual culture questionnaire make reference to digital culture—"cyber" culture, in '90s parlance. However, not nearly as many of the respondents comment on the rise of the internet as one would expect given its relevance for the circulation of digital images and videos.

Despite residing somewhat outside of the discipline of art history herself, Susan Buck-Morss takes a conservative position with regard to the implications of visual culture for the category of "art."[15] Her thoughts on the topic are numerous and complex and are, in fact, inadequately conveyed in the short text for the questionnaire. However, her response is notable in that she is one of the few scholars to refer to digital culture, albeit in a dismissive way. She writes:

> While the Internet is the topic and the medium for new courses in visual culture, it is striking to anyone who has visited the Internet how visually impoverished a home-page can be. Cyberdigits reproduce the moving image haltingly, and the static image unimpressively. The possibility of computer screens replacing television

[14] This was an issue (i.e., visuality) that was raised in the questionnaire by Jonathan Crary, Thomas Crow, and others. "Visual Culture Questionnaire."

[15] Buck-Morss's stance toward visual culture studies was tempered in other writing and statements in the years that followed the questionnaire. In 2004, she published a paper praising the political potential of this methodology, writing, "Visual studies can provide the opportunity to engage in a transformation of thought on a general level. Indeed, the very elusiveness of visual studies gives this endeavor the epistemological resiliency necessary to confront a present transformation in existing structures of knowledge, one that is being played out in institutional venues throughout the globe." Susan Buck-Morss, "Visual Studies and Global Imagination," *Papers of Surrealism*, no. 2 (2004). I would also like to note that in spring 2014 I took a course titled "Software, Globalization and Political Action" at the CUNY Graduate Center cotaught by Buck-Morss and Lev Manovich. Buck-Morss has clearly revised her early dismissal of digital media.

> screens may mean a great deal to stockholders of telephone companies, but it will not shake the world of the visual image. Aesthetic experience (sensory experience) is not reducible to information. Is it old fashioned to say so? Perhaps the era of images that are more than information is already behind us. Perhaps discussions about visual culture as a field have come too late.[16]

The most striking part of this excerpt is the claim that "aesthetic experience (sensory experience) is not reducible to information." Most generous definitions of the term information would allow it its broad etymological origins as idea, concept, or form of knowledge.[17] A strict Kantian reading of aesthetics would indeed separate sensory experience from mere concept, but in the passage above the term "information" is a convenient straw man.[18] Buck-Morss conflates aesthetic experience with images themselves, neither of which she wishes to be merely "information"—but who ever said that they were?[19] Art history sets itself to the task of interpretation and understanding art and so, inevitably, deals with concepts, but no one is arguing that an artwork is reducible merely to the concepts attendant to it.

We need to continually demystify terminology like "information," "data," and "digital" in the humanities. These terms are not obscure loan words from computer science but, rather, concepts that apply to all kind of contemporary scholarship. Although contemporary technological terminology may seem foreign to those steeped in the traditional language of art history, all of the

16 Susan Buck-Morss, "Visual Culture Questionnaire," *October* 77 (1996): 30.

17 "Information, n.," in *OED Online* (Oxford University Press), http://www.oed.com/view/Entry/95568.

18 Immanuel Kant, *Critique of Judgement*, ed. Nicholas Walker, trans. James Creed Meredith (Oxford: Oxford University Press, 2007), 42–43 (§§6–9).

19 Buck-Morss expanded her thoughts on Kantian aesthetics in relation to contemporary art in an interview with Grant Kester from 1997. Following the thinking of Walter Benjamin, she considers "art" a category to be dead. She labels certain contemporary artwork the "Piss and Shit School" and declares identity politics a "cul-de-sac." Grant H. Kester, "Aesthetics after the End of Art: An Interview with Susan Buck-Morss," *Art Journal* 56, no. 1 (1997): 38–45.

concepts floated in digital art history are the same as those that have been floated in art history proper under different auspices.

Some of the most fruitful applications of digital humanities techniques in art history have been in the realm of artists'/artworks' networks and provenance. As Matthew Lincoln argues, an interest in describing or charting networks is already "implicit" within many nondigital research practices across the discipline.[20] However, computational studies of networks are still mostly published in specialist digital art history journals or scientific journals rather than traditional art history journals.[21] One study, by Samuel P. Fraiberger et al. and published in *Science*, looks at quantitative network relationships between museums, exhibitions, auction prices, and artists. They found that, if artists had access to prestigious institutions located at the center of many networks early in their career, they typically had access to those same institutions their entire careers. Artists at the periphery of networks had a higher rate of dropping out.[22] While this may come as no surprise to art historians—that artists with the "best" networks are more likely to succeed—studies like these add nuance and weight to art world truisms regarding artists' relationship to the market, capital, success, and status. This kind of study need not be an end in itself. As Benjamin Zweig argues, this type of study might qualify as "results" but it does not provide "answers" (that is, arguments), and that is where other methodologies and theory can come into play.[23] Writing specifically on methods of digital mapping, Béatrice Joyeaux-Prunel writes, "Mapping is a research process rather than an outcome."[24] DAH critic Claire Bishop writes

[20] Matthew D. Lincoln, "Tangled Metaphors: Network Thinking and Network Analysis in the History of Art," in *The Routledge Companion to Digital Humanities and Art History*, ed. Kathryn Brown (New York: Routledge, 2020), 73.

[21] A number of examples are cited in Lincoln, "Tangled Metaphors."

[22] Samuel P. Fraiberger et al., "Quantifying Reputation and Success in Art," *Science* 362, no. 6416 (November 16, 2018): 825–829.

[23] Benjamin Zweig, "Defining Digital Art History: What It Is, Is Not, and Should Be," paper presented at the Digital Humanities Seminar, Uppsala University (September 25, 2019).

[24] Béatrice Joyeaux-Prunel, "Digital Humanities for a Spatial, Global, and Social History of Art," in *The Routledge Companion to Digital*

that computational methodologies are "boring and formulaic," but the results of computational methodologies need not be an end in themselves.[25]

One of the often-cited examples of a successful application of data analysis in art history is the network diagram MoMA created for their exhibition *Inventing Abstraction 1910–1925* (December 23, 2012–April 15, 2013).[26] This network map was designed with another famous chart in mind. In 1936, the first director of MoMA, Alfred H. Barr Jr., created a diagram to accompany the exhibition *Cubism and Abstract Art*. It depicts the networks of influence that led to the development of "nongeometrical abstract art" and "geometrical abstract art" in Europe and the United States.[27] The neat modernist progression of such a diagram, which purports to trace the evolution of Western art from the late 19th century to the contemporary, is rife with *dei ex machina* in the form of non-Western art ("Japanese prints," "Near Eastern Art," and "Negro Sculpture") and the technological development of the time period ("Machine Esthetic"). These static and isolated "helpers" propel the development of Western art in Barr's diagram.

Humanities and Art History, ed. Kathryn Brown (New York: Routledge, 2020), 94.

[25] Johanna Drucker and Claire Bishop, "A Conversation on Digital Art History," in *Debates in the Digital Humanities 2019* (Minneapolis, MN: University of Minnesota Press, 2019), https://dhdebates.gc.cuny.edu/read/untitled-f2acf72c-a469-49d8-be35-67f9ac1e3a60/section/3aedfd2c-280f-4029-b3f1-3e9a11794c01#ch27.

[26] Nancy Ross, "Teaching Twentieth Century Art History with Gender and Data Visualizations," *The Journal of Interactive Technology and Pedagogy*, no. 4 (December 2, 2013); Mitchell Whitelaw, "Representing Digital Collections," in *Performing Digital: Multiple Perspectives on a Living Archive*, eds. David Carlin and Laurene Vaughan (Farnham: Ashgate Publishing, 2015), 91; Lev Manovich, "Data Science and Digital Art History," *International Journal for Digital Art History*, no. 1 (June 26, 2015): 16; Johanna Drucker et al., "Digital Art History: The American Scene," *Perspective: Actualité en histoire de l'art*, no. 2 (December 5, 2015): 3; Claire L. Kovacs, "Mapping Paris: Social and Artistic Networks, 1855–1889," *Leonardo* 49, no. 5 (March 31, 2016): 446; Zweig, "Defining Digital Art History: What It Is, Is Not, and Should Be."

[27] For further discussion of these two diagrams see Alexander Alberro, "A Messier Coherence: Inventing Abstraction at the Museum of Modern Art," *Modernism/Modernity* 20, no. 2 (June 27, 2013): 371–381.

The connections that mapped in the 2013 diagram, on the other hand, are based on the number of letters exchanged between the figures in the diagram.[28] Key nodes of influence in the network are highlighted in red. This provides some predictable results: Pablo Picasso and Vasily Kandinsky are among those most well connected. It also, however, contains some surprises: historically underappreciated female artists Sonia Delaunay-Terk and Natalia Goncharova are also among those most well connected on the chart. In addition to visual artists, the chart includes the likes of Guillaume Apollinaire and Claude Debussy in the network of influences. Network mapping such as this has the potential to bring figures back into the canon.[29] However, researchers must compile their datasets carefully, as women and minorities may be underrepresented in digitization projects and therefore within existing museum databases.[30] This means that any analysis merely reproduces existing collection or archival biases.

An issue that often arises in the literature on digital humanities is the relationship between the research question and data analysis. Media theorist Lev Manovich is a proponent of "exploratory data analysis" or "unsupervised learning"—terms borrowed from the field of machine learning. Manovich argues that using such methods means that research questions in the humanities will not be predetermined but will be drawn from the data itself. He writes:

> Why should we use computers to classify cultural artifacts, phenomena or activities into a small number of categories? Why not instead use computational methods to question the categories we already have, generate new ones, or create new cultural maps that relate cultural artifacts in original ways?[31]

Although Manovich's stated goal is to do away with existing bias in the categorization of cultural artifacts, this freeform approach

[28] Manovich, "Data Science and Digital Art History," 16.

[29] Kathryn Brown and Elspeth Mitchell, "Feminist Digital Art History," in *The Routledge Companion to Digital Humanities and Art History*, ed. Kathryn Brown (New York: Routledge, 2020), 44.

[30] Brown and Mitchell, "Feminist Digital Art History," 46.

[31] Manovich, "Data Science and Digital Art History," 24.

is built on the assumption that datasets are neutral or comprehensive and that there are objective facts about art held within them. The data itself will reproduce human bias (just as humans would) in the categorization of objects. As recent scholarship on the topic of data bias has shown, this is a huge issue in data collection and analysis, and the belief in the objectivity of big data sets can have dire social consequences.[32]

For humanities scholars who favor a critical theory perspective, the centrality of the argument and authorial voice cannot be discounted.[33] In an article from 2013, Johanna Drucker points particularly to the influence of "[s]emiotics, structuralism, poststructuralism, psychoanalysis, Marxism, cultural and critical studies, and feminist" methodologies as "profound" for the discipline of art history, a seismic change she does not see (yet) in digital art history.[34] Bishop draws from Drucker to argue against digital art history on the basis that what Drucker outlines as the future of digital art history amounts to a "combination of digital technologies, network analysis, and connoisseurship."[35] Bishop takes issue with what she perceives as the "fatuity" of digital art history scholarship, where "[b]asic terms like beauty (and even portraiture) remain uninterrogated."[36] Ultimately Bishop argues that the reliance on and belief in data as revolutionary to the field of art history "perpetuates uncritical assumptions about

[32] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: New York University Press, 2018); Ricardo Baeza-Yates, "Bias on the Web," *Communications of the ACM* 61, no. 6 (June 2018): 54; Shaozeng Zhang, Bo Zhao, and Jennifer Ventrella, "Towards an Archaeological-Ethnographic Approach to Big Data: Rethinking Data Veracity," *Ethnographic Praxis in Industry Conference Proceedings 2018*, no. 1 (2018): 62–85; Andrew G. Ferguson, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement* (New York: New York University Press, 2017); Megan Garcia, "Racist in the Machine: The Disturbing Implications of Algorithmic Bias," *World Policy Journal* 33, no. 4 (2016): 111–117.

[33] Michelle Millar Fisher and Anne Swartz, "Why Digital Art History?" *Visual Resources* 30, no. 2 (April 3, 2014): 131.

[34] Johanna Drucker, "Is There a 'Digital' Art History?" *Visual Resources* 29, no. 1–2 (June 1, 2013): 5.

[35] Claire Bishop, "Against Digital Art History," *International Journal for Digital Art History*, no. 3 (July 27, 2018): 123.

[36] Bishop, "Against Digital Art History," 124.

the intrinsic value of statistics" and feeds a neoliberal system.[37] While this may be true in some cases, computation analysis of networks can also expose some of the more insidious ways markets have driven art's valuation and propped up certain artists. As long as a thoughtful mix of art historical methodologies are utilized, which recognize the politics and problematics of terminology and the discipline's historicity, adding digital tools to the mix need not produce fatuous or neoliberal work.

Drucker and Bishop subsequently debated some of these points directly in an email exchange published in the 2019 edition of *Debates in the Digital Humanities*. Countering the accusation that digital humanities methodologies are somehow more complicit with neoliberalism than traditional methods, Drucker writes, "This is patently false. It also suggests that a 'pure' humanities exists that is untainted: the humanities of work that embraces social good and the highest virtues of humankind without complicity in the institutional frameworks that support it."[38] Bishop's point, however, that the university's focus on digital humanities research as a reflection of neoliberal values imposed on otherwise unquantifiable humanities research still stands.

Some of the most controversial and problematic digital art historical research concerns studies of digital images and image compositing/creation through machine learning. As Bishop points out, analysis of large groups of image data often ignores and erases too much context for it to be meaningful. That is not to say it could not be done with an eye to these issues but that, as of yet, too many examples of image analysis are superficial, and the results of such research risk being inconsequential to art's *history*. For example, a 2019 commercial gallery show in New York was advertised as a "collaboration" between a computer scientist and an AI system to create new art out of a repository of old masters portrait imagery, but, as journalist Ian Bogost points out, the subject

37 Bishop, "Against Digital Art History," 125.

38 Johanna Drucker and Claire Bishop, "A Conversation on Digital Art History," in *Debates in the Digital Humanities 2019* (Minneapolis, MN: University of Minnesota Press, 2019), https://dhdebates.gc.cuny.edu/read/untitled-f2acf72c-a469-49d8-be35-67f9ac1e3a60/section/3aedfd2c-280f-4029-b3f1-3e9a11794c01#ch27.

of the portraits used and the specificity of their time/place is completely erased in such an exercise.[39] The project assumes that the value of old masters' work has nothing to do with its specificity and everything to do with its superficial features. While such visual analyses of digital images might be linked with the dates of creation for the images in question, dates alone are not enough to construct history. Drucker also points to some of the issues surrounding the use of digital images to collect "data" on the object depicted, writing, "[a]s previously noted, digital objects are fully remediated. They exist in the fungible condition of code. The way artifacts are encoded depends on the parameters set for scanning and photography."[40] The layers of mediation in both metadata and encoding practices cannot be discounted in any study that employs digital humanities methodologies.

Studying Internet Art

Like digital art history, digital art (computer art, internet art, *et cetera*) has long been ghettoized under the category of new media art. During the early stages of a recently completed research project, I met with Lev Manovich to discuss how digital methods might be applied to my study of early internet art.[41] He suggested text mining the <nettime> mailing list,[42] one of the key forums for internet art and theory in the '90s. Since <nettime> is archived online, I would be able to do different kinds of textual analysis of the bulk of the content there. In the end, however, I decided to focus my project on an earlier period leading up to the advent of the internet in the Netherlands within a community of urban squatters, hackers, activists, and artists. As such, my methodology turned out to include the decidedly analog techniques of archival research and oral histories.

---

[39] Ian Bogost, "The AI-Art Gold Rush Is Here," *The Atlantic*, March 6, 2019, https://www.theatlantic.com/technology/archive/2019/03/ai-created-art-invades-chelsea-gallery-scene/584134.

[40] Drucker, "Is There a 'Digital' Art History?" 8.

[41] Amanda Wasielewski, *From City Space to Cyberspace: Art, Squatting, and Internet Culture in the Netherlands*. Amsterdam University Press. Forthcoming.

[42] "<nettime> Mailing List Archive," https://nettime.org/.

As with the selection of any methodology, a researcher needs to interrogate what they hope to learn by using it, which does not necessarily have to be a concrete outcome. It goes without saying that, even when a researcher constructs an explicit plan for a piece of humanities research, an argument can turn out slightly or even vastly different from how they thought it would. Any methodological decision will yield a different kind of output and, despite the warring camps within any discipline, this does not mean that one is "right" and another is "wrong." Though it seems simple to point out these things about choosing a methodology, a version of which is taught in art history departments across the world, it is useful to be reminded of it. Even as a senior scholar—perhaps especially as a senior scholar who has become known for working in a certain way within a certain framework of analysis—it is useful to think of methodologies as techniques that can be combined and mixed together rather than camps ready to wage war on one another. This also applies to political and ideological methodologies: intersectionality would never have been born if gender, race, postcolonial, and Marxist theory had never been brought together. Digital art history or computational methodologies are not a threat to art history any more than visual culture studies were. Nor are they useless and unfit for the study of art.

Returning to the example of internet art, it is clear that there are questions that might be answered by doing a computational analysis of the <nettime> mailing list. In the early to mid-'90s, there was a network of artists (many of whom were from former Soviet countries) who created art that was meant to be viewed in an internet browser window. Theorists and artists from both the Netherlands and Germany interacted with these artists on <nettime> and in other venues, namely in-person conferences and events. One could perhaps devise a network analysis or even a geographical map of <nettime> participants based on which figures responded to and interacted with each other. Key phrases and topics could be pinpointed or often-cited artists or theorists might be uncovered. This is, of course, just one set of data that does not encapsulate the entirety of the discourse, but that does not mean it cannot be usefully analyzed with computational methodologies. Before demonstrating what can be done with a given set of data, however, one has to ask why it should be done in the first place. Every piece of research starts from a clue, whether that is an errant observation, a hunch, or the existence of a set of data. It should be quite obvious that the demonstration of a technique—digital or otherwise—falls under the category of pedagogy rather than scholarly argument. That is not to say that pedagogy is a less valid pursuit for scholars, but pedagogical demonstration and argument are so often confused with each other in debates on digital art history that it seems necessary to make that distinction explicit.

Despite the *availability* of this already-digitized database, which is often not the case for older periods of artistic production, it may not hold the answers a researcher finds interesting or important to highlight. As it turns out, none of the questions that can be drawn from the <nettime> mailing list were ones that were relevant to the research I conducted, but that does not mean that they might not be in future research on the same topic. Critics and scholars have approached internet art with a variety of other methodologies that I have found more or less useful. For example, some of the earliest attempts to bring internet art into the fold of art historical scholarship applied Greenbergian formalism to the work, which is still a dominant thread in writing on this period.[43] This type of analysis was less interesting to me than, for example, the economic and political framework that scholars like Julian Stallabrass have explored.[44]

The difference between utilizing digital humanities methodologies on older artforms and researching internet art, as well as

[43] Clement Greenberg was an American art critic who devised a theory of art where work was judged based on its truth to its medium, or "medium specificity." Internet art critics have picked up on the idea that there are certain qualities of the "medium" of internet art that are unique and that the best work highlights these attributes. Some of Rosalind Krauss's ideas on art follow Greenberg and are also highlighted by writing on internet art. See Tilman Baumgärtel, *Net.art: Materialien zur Netzkunst* (Nuremberg: Verlag für Moderne Kunst, 1999); Tilman Baumgärtel, *Net.art 2.0: New Materials Towards Net Art* (Nuremberg: Verlag für Moderne Kunst, 2001); Josephine Bosma, *Nettitudes: Let's Talk Net Art* (Amsterdam: Institute of Network Cultures, 2011).

[44] Julian Stallabrass, *Internet Art: The Online Clash of Culture and Commerce* (London: Tate Publishing, 2003).

other born-digital works, is that digital works were created in the same systemic paradigm that computational methodologies utilize. I would not wish to overstate the importance of this, but it undoubtedly creates opportunities that pairing analog works with digital methods does not.

Just Another Archive?

The extent to which the minutia of our everyday life exists as data on computers, phones, or in server farms (the "cloud") is unprecedented. Blockbuster data security revelations like the Facebook/Cambridge Analytica scandal of 2018, as well as the Edward Snowden disclosures of 2012, have brought some of the issues of data privacy and personal data ownership to the surface in political discourse. While there are certainly sinister and undesirable implications of this kind of mass data collection and surveillance, there is also a potential wealth of information that could be available to historians (after accounting for any ethical issues of using it, of course). If an artist creates their work on a computer, the process through which they made the work is often documented in more detail than ever before.

An interesting example of how this has been used in scholarship is Doug Reside's research on the musical *RENT*. Reside was appointed as the first digital curator at the New York Public Library's Performing Arts division, where he has written about and experimented with digital humanities methodologies. After learning that the Library of Congress had 180 floppy disks from *RENT* creator Jonathan Larson, who died from a heart condition at the age of thirty-five in 1996, Reside set about analyzing the music and text files contained on the disks.[45] Not only was Reside able to pinpoint sections of the *RENT* libretto that had been edited but he was also able to see previous and even deleted versions of the text. Additionally, he was able to investigate the notes that Larson took on various books he read as well as the exact moment he created or edited certain files. This type of archive is

[45] Doug Reside, "Last Modified January 1996: The Digital History of RENT," *Theatre Survey* 52, no. 2 (November 2011): 335–340.

really no different from any paper archive except that it contains such down-to-the-second details. It is an archive on steroids.

Artists themselves have been quick to explore the poetic potential of digital traces. Bishop cites Walid Raad's use of data and visualization in his work *Scratching on Things I Could Disavow* (ongoing).[46] The work reflects on the nature of globalized networks and the concept of researching the relationship between art institutions and economic and political power in the Arab world. Bishop states that the work is "at first glance very DH [digital humanities]," a statement that still somehow treats "the digital" as a separate field, something apart, rather than something inextricably entwined in every part of social, economic, and political life today.

Another example of this sort of exploration of data in artistic practice is Forensic Architecture's digital reconstructions and models of instances of political violence. Without readily accessible trails of data, the projects of Forensic Architecture would not be possible. Their most well-known work, *Triple-Chaser*, documents the way in which tear gas cannisters made by a company called Safariland were used on the US–Mexico border against migrants. The work was inspired by the controversy surrounding the owner of Safariland, Warren B. Kanders, who was a board member of the Whitney Museum of American Art. The connection between the Whitney and Kanders demonstrated the extent to which the moral and ethical position of the artists whose work is on display in the museum is disregarded in order to flatter wealthy donors such as Kanders. After Forensic Architecture were invited to participate in the 2019 Whitney Biennial, they decided to investigate and recreate this story, training "'computer vision' classifiers to detect Safariland tear gas canisters among the millions of images shared online."[47] Kanders eventually resigned under the pressure of protests against his continued position at the museum.

These projects demonstrate that data and the use of data need not be dull. Is it so unfathomable that art historical research could also explore some of the poetic potential of artworks' data? That

---

[46] Drucker and Bishop, "A Conversation on Digital Art History."

[47] Forensic Architecture, "Triple-Chaser: Forensic Architecture," https://forensic-architecture.org//investigation/triple-chaser.

insights collected from large repositories of data could feed into a compelling argument about works of art? As with any other methodology, the interesting part of a piece of research in digital art history is inevitably what it has to *say* about a work of art, an artist, or a moment of artistic creation.

In the cases above, both artistic and scholarly, the researchers in question were dealing with readily available large datasets. Doug Reside's research is perhaps a glimpse into the future, when archival work will have to contend with personal data. Perhaps when every artist any art historian wants to study has a hard drive full of working files, correspondence, and notes, computational methodologies to sort and unpack the massive amounts of data will start to seem more relevant. Visualizations like those created by artists like Walid Raad and Forensic Architecture could eventually be useful in telling the story of complex sets of images and complicating the basis upon which they were collected.

The Problem of Data

The main hindrance for art historical uptake of computational methodologies is not primarily that they are boring or neoliberal; it is that preexisting datasets for art history are often flawed, biased, or incomplete. For noncanonical or niche projects, datasets rarely exist and almost always need to be created first before they are analyzed. More generalized analyses using existing datasets, on the other hand, are often working with flawed or biased data. The generalized art datasets that already exist tend to be composed of primarily Western, canonical painting. Part of the rationale for this is, one would assume, that existing image analysis tools work best on two-dimensional works. There are also copyright issues to contend with for modern and contemporary works. While attempts have been made to unite the various museum collections of images and their attendant metadata, no such comprehensive dataset yet exists.[48] Even it if did, it would likely

[48] Gjorgji Strezoski and Marcel Worring, "OmniArt: A Large-Scale Artistic Benchmark," *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, no. 4 (October 23, 2018): 88:1–88:21.

replicate the biases of Western museums, an aspect that would need to be dealt with in any "general" analysis of art.

Indeed, existing art datasets, such as the popular WikiArt dataset, do not meet academic standards for art historical research in that the sources of the information provided is not properly documented. WikiArt includes metadata that is at least partially crowdsourced from the general public, which even researchers in computer science have found problematic as they worked with the dataset.[49] Most contemporary digital art history projects that might benefit from using computational methodologies need to actually *create* the data first before studying it. This of course poses a whole host of problems, not least that creating, collecting, managing, and storing data is a very expensive endeavor. Paul Jaskot writes that art history has long been in the business of collecting encyclopedic accumulations of artworks for iconographic analysis and, so, the era of big data collection is nothing new for art history.[50] These typological and formalist aims, however, are no longer the primary motivation for art historical scholarship.[51] For both 19th-century art collections and contemporary datasets, collecting is not a neutral activity. The accumulation of non-Western artworks and artifacts during the modern era was part of the Western colonialist project, which museums continue to perpetuate. Big data collection today, on the other hand, assumes a comprehensiveness that elides systemic biases.

Looking at the projects that Johanna Drucker cites in her published debate with Claire Bishop as examples of successful digital art history projects, every single one of them required the creation of the data they intended to study. Furthermore, many of them have goals more in line with archaeological research rather than art historical research. She cites projects, such as the Dunhuang Cave Project funded by Getty and UNESCO, which required enormous

49 Ahmed Elgammal et al., "The Shape of Art History in the Eyes of the Machine," January 23, 2018, http://arxiv.org/abs/1801.07729v2, 5–6.

50 Paul B. Jaskot, "Digital Methods and the Historiography of Art," in *The Routledge Companion to Digital Humanities and Art History*, ed. Kathryn Brown (New York: Routledge, 2020), 9.

51 Anna Dahlgren and Amanda Wasielewski, "The Digital U-Turn in Art History." Submitted.

amounts of image and other data collection, and work on burial mounds that required the development of techniques to begin collecting data on the sites.[52] Certainly now that, for example, much of the Dunhuang caves are digitized and available to explore in a virtual environment, art historical study of them is much easier. However, this is no different from an art historian searching online archives of medieval manuscripts or other digitized art objects. Drucker also cites the compilation of the Getty Provenance Index, a project that has the primary purpose of creating a complex dataset (not actually developing an argument based on it).[53] She goes on to cite a few projects that utilize preexisting textual datasets, which are certainly easier to deal with than image-based data but still require a concerted effort to compile in many cases.

As mentioned, born-digital data may eventually solve some of the issues for studying art via computational methods, but, for now, the major obstacle for meaningfully implementing some of these methodologies in art history is that datasets simply do not exist or are too partial, distorted, or full of noise to really be useful. The solution to this, however, is not necessarily more mass digitization projects. We need to think deeply about the implications of such projects, as there are also myriad political, social, and ethical quandaries involved.[54]

As much as art historians might resist the turn toward digital methods, they will only grow more useful for future generations of scholarship (if both technological progress and the discipline

52 Neville Agnew, ed., *Conservation of Ancient Sites on the Silk Road: Proceedings of the Second International Conference on the Conservation of Grotto Sites, Mogao Grottoes, Dunhuang, People's Republic of China, June 28–July 3, 2004* (Los Angeles, CA: Getty Publications, 2010); Melanie A. Riley, *Date Automated Detection of Prehistoric Conical Burial Mounds from LIDAR Bare-Earth Digital Elevation Models* (Master's thesis, Northwest Missouri State University, 2009); Gunnar Liestøl and Terje Rasmussen, "In the Presence of the Past: A Field Trial Evaluation of a Situated Simulation Design Reconstructing a Viking Burial Scene," in *Media Inspirations for Learning: Proceedings of EDEN 2010*, eds. A. Szucs and A.W. Tait (Budapest: Budapest University of Technology, 2010).

53 Getty Research Institute, "Provenance Index Databases," https://www.getty.edu/research/tools/provenance/search.html.

54 See Nanna Bonde Thylstrup, *The Politics of Mass Digitization* (Cambridge: MIT Press, 2019).

itself continue on in the same vein). The biggest issue right now for digital art history is available data to study, but, before we begin the arduous task of compiling such data, it is worthwhile to think what we might hope to achieve with it. As for the datasets that already exist, it is up to art historians to unlock insights they might hold while maintaining a critical eye toward biases. As Emma Stanford writes, "In attempting to maximize the potential impact of a proposed digitization project, an institution may focus on the 'important' parts of its collection, but importance is an extremely subjective measure dependent on many variables."[55] Computational methodologies can be relevant, be useful, *and* lead researchers to interesting avenues of inquiry, but that may only happen if the digital paradigm is no longer ghettoized as a separate field of inquiry. We live and operate as scholars almost completely within digital culture, and, so, we need not subscribe to either methodological fundamentalism or digital ghettoization.

Acknowledgments

The research for this chapter was conducted within the project *Sharing the Visual Heritage* (metadataculture.se) at Stockholm University. It was made possible through the generous support of the Swedish Research Council.

References

Agnew, Neville, ed. *Conservation of Ancient Sites on the Silk Road: Proceedings of the Second International Conference on the Conservation of Grotto Sites, Mogao Grottoes, Dunhuang, People's Republic of China, June 28–July 3, 2004*. Los Angeles, CA: Getty Publications, 2010.

Alberro, Alexander. "A Messier Coherence: Inventing Abstraction at the Museum of Modern Art." *Modernism/Modernity* 20, no. 2 (June 27, 2013): 371–381.

[55] Emma Stanford, "A Field Guide to Digital Surrogates: Evaluating and Contextualizing a Rapidly Changing Resource," in *The Routledge Companion to Digital Humanities and Art History*, ed. Kathryn Brown (New York: Routledge, 2020), 205.

Baeza-Yates, Ricardo. "Bias on the Web." *Communications of the ACM* 61, no. 6 (June 2018): 54.

Baker, George, ed. "Artist Questionnaire: 21 Responses." *October* 100 (2002): 6–97.

Baumgärtel, Tilman. *Net.art 2.0: New Materials Towards Net Art*. Nuremberg: Verlag für Moderne Kunst, 2001.

Baumgärtel, Tilman. *Net.art: Materialien zur Netzkunst*. Nuremberg: Verlag für Moderne Kunst, 1999.

Bogost, Ian. "The AI-Art Gold Rush Is Here." *The Atlantic*, March 6, 2019. https://www.theatlantic.com/technology/archive/2019/03/ai-created-art-invades-chelsea-gallery-scene/584134.

Bishop, Claire. "Against Digital Art History." *International Journal for Digital Art History*, no. 3 (July 27, 2018): 122–131.

Bosma, Josephine. *Nettitudes: Let's Talk Net Art*. Amsterdam: Institute of Network Cultures, 2011.

Brown, Kathryn, ed. *The Routledge Companion to Digital Humanities and Art History*. New York: Routledge, 2020.

Buchloh, Benjamin H. D., and Rachel Churner, eds. "Questionnaire: In What Ways Have Artists, Academics, and Cultural Institutions Responded to the U.S.-Led Invasion and Occupation of Iraq?" *October* 123 (2008): 3–184.

Buck-Morss, Susan. "Visual Culture Questionnaire." *October* 77 (1996): 29–31.

Buck-Morss, Susan. "Visual Studies and Global Imagination." *Papers of Surrealism*, no. 2 (2004).

Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp. *Digital_Humanities*. Cambridge: MIT Press, 2012.

Copeland, Huey, et al. "A Questionnaire on Decolonization." *October* 174 (December 1, 2020): 3–125.

Dahlgren, Anna and Amanda Wasielewski. "The Digital U-Turn in Art History." Submitted.

Debatty, Régine. "Interview with Marisa Olson." *We Make Money Not Art* (blog), March 28, 2008. http://we-make-money-not-art.com/how_does_one_become_marisa.

Dickerman, Leah, Hal Foster, David Joselit, and Carrie Lambert-Beatty, eds. "A Questionnaire on Monuments." *October* 165 (August 1, 2018): 3–177.

"Digital Humanities Laboratory at Rutgers." https://sites.google.com/site/digihumanlab/home.

Drucker, Johanna, and Claire Bishop. "A Conversation on Digital Art History." In *Debates in the Digital Humanities 2019*. Minneapolis, MN: University of Minnesota Press, 2019. https://dhdebates.gc.cuny.edu/read/untitled-f2acf72c-a469-49d8-be35-67f9ac1e3a60/section/3aedfd2c-280f-4029-b3f1-3e9a11794c01#ch27.

Drucker, Johanna, Anne Helmreich, Matthew Lincoln, and Francesca Rose. "Digital Art History: The American Scene." *Perspective: Actualité en histoire de l'art*, no. 2 (December 5, 2015).

Elgammal, Ahmed, Marian Mazzone, Bingchen Liu, Diana Kim, and Mohamed Elhoseiny. "The Shape of Art History in the Eyes of the Machine," January 23, 2018. http://arxiv.org/abs/1801.07729v2.

Ferguson, Andrew G. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. New York: New York University Press, 2017.

Forensic Architecture. "Triple-Chaser: Forensic Architecture." https://forensic-architecture.org//investigation/triple-chaser.

Foster, Hal, ed. "Questionnaire on 'The Contemporary.'" *October* 130 (2009): 3–124.

Garcia, Megan. "Racist in the Machine: The Disturbing Implications of Algorithmic Bias." *World Policy Journal* 33, no. 4 (2016): 111–117.

Getty Research Institute. "Provenance Index Databases." https://www.getty.edu/research/tools/provenance/search.html.

Gold, Matthew K., ed. *Debates in the Digital Humanities*. Minneapolis, MN: University of Minnesota Press, 2012.

"Information, n." In *OED Online*. Oxford University Press. http://www.oed.com/view/Entry/95568.

Joselit, David, and Carrie Lambert-Beatty, eds. "[15 Responses to a Questionnaire on Occupy Wall Street]." *October* 142 (2012): 26–73.

Joselit, David, Carrie Lambert-Beatty, and Hal Foster, eds. "A Questionnaire on Materialisms." *October* 155 (Winter 2016): 3–110.

Kant, Immanuel. *Critique of Judgement*, edited by Nicholas Walker, translated by James Creed Meredith. Oxford: Oxford University Press, 2007.

Kester, Grant H. "Aesthetics after the End of Art: An Interview with Susan Buck-Morss." *Art Journal* 56, no. 1 (1997): 38–45.

Klinke, Harald. "Big Image Data within the Big Picture of Art History," *International Journal for Digital Art History*, no. 2 (October 18, 2016): 14–37.

Kovacs, Claire L. "Mapping Paris: Social and Artistic Networks, 1855–1889." *Leonardo* 49, no. 5 (March 31, 2016): 446–446.

Liestøl, Gunnar, and Terje Rasmussen. "In the Presence of the Past: A Field Trial Evaluation of a Situated Simulation Design Reconstructing a Viking Burial Scene." In *Media Inspirations for Learning: Proceedings of EDEN 2010*, edited by A. Szucs and A. W. Tait. Budapest: Budapest University of Technology, 2010.

Manovich, Lev. "Data Science and Digital Art History." *International Journal for Digital Art History*, no. 1 (June 26, 2015): 12–35.

Moretti, Franco. *Distant Reading*. London: Verso, 2013.

"<nettime> Mailing List Archive." https://nettime.org.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press, 2018.

"Questions of Feminism: 25 Responses." *October* 71 (1995): 5–48.

Reside, Doug. "Last Modified January 1996: The Digital History of RENT." *Theatre Survey* 52, no. 2 (November 2011): 335–340.

Riley, Melanie A. *Date Automated Detection of Prehistoric Conical Burial Mounds from LIDAR Bare-Earth Digital Elevation Models*. Master's thesis, Northwest Missouri State University, 2009.

Ross, Nancy. "Teaching Twentieth Century Art History with Gender and Data Visualizations." *The Journal of Interactive Technology and Pedagogy*, no. 4 (December 2, 2013).

Schreibman, Susan, Raymond George Siemens, and John Unsworth, eds. *A Companion to Digital Humanities*. Malden, MA: Blackwell Publishing, 2004.

Stallabrass, Julian. *Internet Art: The Online Clash of Culture and Commerce*. London: Tate Publishing, 2003.

Strezoski, Gjorgji, and Marcel Worring. "OmniArt: A Large-Scale Artistic Benchmark." *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, no. 4 (October 23, 2018): 88:1–88:21.

Terras, Melissa M., Julianne Nyhan, and Edward Vanhoutte, eds. *Defining Digital Humanities: A Reader*. London: Routledge, 2016.

Thylstrup, Nanna Bonde. *The Politics of Mass Digitization*. Cambridge: MIT Press, 2019.

Vierkant, Artie. "The Image Object Post-Internet." *Jstchillin* (blog). 2010. http://jstchillin.org/artie/pdf/The_Image_Object_Post-Internet_a4.pdf.

"Visual Culture Questionnaire." *October* 77 (1996): 25–70.

Whitelaw, Mitchell. "Representing Digital Collections." In *Performing Digital: Multiple Perspectives on a Living Archive*, edited by David Carlin and Laurene Vaughan, 77–96. Farnham: Ashgate Publishing, 2015.

Zhang, Shaozeng, Bo Zhao, and Jennifer Ventrella. "Towards an Archaeological-Ethnographic Approach to Big Data: Rethinking Data Veracity." *Ethnographic Praxis in Industry Conference Proceedings* 2018, 1 (2018): 62–85.

Zweig, Ben. "Defining Digital Art History: What It Is, Is Not, and Should Be." Paper presented at the Digital Humanities Seminar, Uppsala University, September 25, 2019.

PART TWO:
LEGAL AND ETHICAL DILEMMAS

Elevating Legal Informatics in the Digital Age

Stanley Greenstein

Introduction

We are living in an increasingly digitalized society where we are connected to the internet constantly and where a large portion of our social and commercial interaction takes place. It is in this context that almost all human behavior is or can be reduced to data, where human-to-human interaction is being replaced by human-to-machine interaction, and where machines incorporating artificial intelligence (AI) are making the decisions that affect humans. The popular euphoria associated with AI has resulted in a mass hysteria concerning the advantages to society. However, there are also risks with this technology and it is imperative to take cognizance of the fact that technology is a double-edged sword.

The law has many functions within society. It provides a mechanism for resolving disputes once they occur, for ensuring predictability within society, for facilitating commercial activity, for reflecting ethical norms, and for protecting society from risks and vulnerabilities by means of both reactive and proactive regulation. In other words, the overarching function of the law is that of a mechanism for handling problems and disputes.[1] The increasing use of information technology incorporating elements of AI

[1] Peter Wahlgren, "Automatiserade juridiska beslut," in *Juridisk metodlära*, eds. Maria Nääv and Mauro Zamboni (Lund: Studentlitteratur, 2018), 407.

How to cite this book chapter:
Greenstein, Stanley. "Elevating Legal Informatics in the Digital Age." In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 155–180. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.g.

is resulting in societal challenges and potential risks and harms that have a multidisciplinary imprint. A logical consequence is to deliberate the extent to which the law is equipped to face these new problems.

Considering the function of the law as a problem-solver, this chapter investigates the suitability of the law for solving the complex problems arising from the use of AI. It begins by examining the technology. This focus on technology, more specifically AI, is done not with the aim of investigating AI as a phenomenon in itself but is rather an illustration of the power of technology. Two legal methods are then introduced. The first is the "traditional legal science method," incorporating the legal dogmatic method, and the second that of legal informatics.[2] These two methods are contrasted in relation to the complex nature of problems with AI.

The main aim of this chapter is to illustrate that, considering the interdisciplinary nature of the problems arising in the digital society, a more interdisciplinary legal approach is required to tackle problems and protect society from the ensuing risks. Complex problems require complex solutions. This necessitates an interdisciplinary approach, including research within and the application of the law. The legal informatics approach facilitates this by being more receptive to influences from other disciplines. Therefore, legal informatics not only provides an enhanced response by the law to problems arising from digitalization. It also maneuvers the law to better be a part of interdisciplinary research and thereby be a part of societal solutions.

The Power of Data

A conception shared by many is that technology is advancing at a rapid pace. New technologies appear as fast as old ones disappear and the digital environment seems to be in a state of continual

[2] At this point, it is important to place this chapter in perspective. First, it is written through the lens of the Swedish legal context; however, it is likely to be relevant in relation to other legal contexts. In addition, while multiple legal methods do exist, it is submitted that the legal dogmatic method is by far that legal method that is most widely applied by scholars, taught to students and applied by practitioners and judges.

flux. It is within this context that certain buzzwords seem to pop up from nowhere, for example "data mining," "big data," "predictive analytics," "profiling," and, more recently, "deep learning," "neural networks," and "artificial intelligence" (AI). What constitutes AI as a concept is subjective and is best described as moving target. What AI is for one person may not necessarily be AI for another, what was considered AI, say, 15 years ago is nowadays considered commonplace, and even the question "what is intelligence?" can be contested and debated. It is most probably for the above reasons (and many more) that there is as yet no legal definition of AI.

In attempting to define AI, it is important to acknowledge its existence as an academic discipline:

> "Artificial Intelligence," or AI, is a cross-disciplinary approach to understanding, modelling, and creating intelligence of various forms. It is a critical branch of cognitive science, and its influence is increasingly being felt in other areas, including the humanities. AI applications are transforming the way we interact with each other and with our environment, and work in artificially modelling intelligence is offering new insights into the human mind and revealing new forms mentality can take.[3]

Within the realm of AI, models are used that explain various dimensions of human and animal cognition, where the focus can be on the engineering of smart machines and applications.[4] AI research areas include knowledge representation, heuristic search, planning, expert systems, machine vision, machine learning, natural language processing, software agents, intelligent tutoring systems, and robotics.[5] In further seeking a definition of AI, it can be described as "a cross-disciplinary approach to understand, modelling, and replicating intelligence and cognitive processes by invoking various computational, mathematical, logical,

---

[3] In *The Cambridge Handbook of Artificial Intelligence*, eds. Keith Frankish and William M. Ramsay (Cambridge: Cambridge University Press, 2014), front matter.

[4] Keith Frankish and William M. Ramsay, eds., "Introduction," in *The Cambridge Handbook of Artificial Intelligence* (Cambridge: Cambridge University Press, 2014), 1.

[5] Frankish and Ramsay, "Introduction," 24–27.

mechanical and even biological principles and devices."[6] AI can therefore be described as the pursuit of providing machines with intelligent capabilities modeled on those of human beings.

Machine learning (ML) involves the use of algorithms that allow AI systems to learn and which has become more important as these systems have started to operate more autonomously and in increasingly complex and dynamic areas of application.[7] An algorithm is "[a] process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer."[8] Algorithms are used to identify patterns in data, after which these patterns are transposed into rules that are then utilized in decision-making systems. Algorithms learn from historical data, after which this knowledge can then be applied to novel data or situations. The extent to which we, with the help of machine learning, are able to extract knowledge from data has also altered the manner in which we use data to base our decisions on. Here reference is made to the notion of data-driven practices and evidence-based practices, where the former is problematic not least from the ethical perspective.[9]

AI technologies are characterized by two attributes, namely, "autonomy," being "the ability to perform tasks in in complex environments without guidance by a user" and "adaptivity," being "[t]he ability to improve performance by learning from experience."[10] Machine learning, therefore, is an area of AI that has progressed rapidly of late, one concrete form of this application of technology being the development and use of predictive models.

Models allow us to generalize about our surroundings in order to gain valuable insights that can be used to predict events.[11]

---

6 Frankish and Ramsay, "Introduction," 1.
7 Frankish and Ramsay, "Introduction," 26.
8 English Oxford Living Dictionaries, entry "Algorithm," https://en.oxforddictionaries.com/definition/algorithm.
9 For a more in-depth discussion of data-driven practices and evidence-driven practices see Teresa Cerratto Pargman and Cormac McGrath in this volume.
10 "Elements of AI," online course on artificial intelligence, https://www.elementsofai.com.
11 Simon Winter and Per Johansson, "Digitalis filosofi: Människor, modeller och maskiner," *SE:s Internetguide*, no. 13, version 1.0 (2009): 21–34.

While models generally have been used to gain insight into data within the natural sciences, they are increasingly being used in social contexts in order to assist with the making of decisions. For example, commercial actors steer their interaction with clients via models built into digital decision-making systems that potentially incorporate AI elements and that interact with clients/potential clients. This increases commercial effectivity and cuts costs but it also allows commercial actors to gain insight from the data about their clients/potential clients. In other words, by identifying "risky" human behavior, companies can make better commercial decisions by separating "desirables" from "undesirables."[12] Models can also have a predictive aim:

> A predictive model captures the relationships between predictor data and behaviour, and is the output from the predictive analytics process. Once a model has been created, it can be used to make new predictions about people (or other entities) whose behaviour is unknown.[13]

Consequently, data left behind by humans in the digital environment reveals lots about them, especially regarding their behavior and personalities. Knowledge is power, yet it is only a few who have the means to afford this power-enhancing technology.

Enter the cognitive sciences

The disciplines that study human cognitive behavior are continually gaining new insights into how human beings think and make decisions. It has been stated that "[w]e live in an age of psychology and behavioural economics—the behavioural sciences."[14] Knowledge concerning how people think, psychologically and

---

[12] In search of a definition of "risk" many alternatives exist. For example, the Society for Risk Analysis (SRA) defines it as "the potential for realization of unwanted, adverse consequences to human life, health, property, or the environment," in Peter Wahlgren, *Legal Risk Analysis: A Proactive Legal Method* (Stockholm: Jure, 2013), 20.

[13] Steven Finlay, *Predictive Analytics, Data Mining, and Big Data: Myths, Misconceptions and Methods* (Basingstoke: Palgrave Macmillan, 2014), 215.

[14] Cass R. Sunstein, *The Ethics of Influence: Government in the Age of Behavioural Science* (Cambridge: Cambridge University Press, 2016), 1.

physiologically, can be incorporated into decision-making systems and be used also to manipulate behavior. This is not particularly difficult considering that human beings are creatures of habit, habit accounting for 45% of the choices humans make every day.[15] From this vantage point, historical data becomes important as it provides a window into the future.

For example, behavioral economics and behavioral finance are areas where the cognitive ability of humans is studied. The former is an area of study that is attributed to Daniel Kahneman and Amos Tversky, who authored an article that referred to psychological cognitive techniques in order to explain deviations regarding the making of decisions in relation to classical economic theory, thereby explaining why investors behave the way they do.[16] Behavioral finance is described as "a sub-field of behavioral economics and proposes psychology-based theories to explain stock market anomalies, such as severe rises or falls in stock price."[17] Kahneman, in his book *Thinking Fast and Slow*, also describes humans' thought processes by distinguishing between two systems that drive human beings' cognitive processes and, depending on how one wants to manipulate a person, one of these systems for thinking could be addressed.[18] On the topic

---

[15] David T. Neal, Wendy Wood, and Jeffry M. Quinn, "Habits – A Repeat Performance," *Current Directions in Psychological Science* 15, no. 4 (August, 2006): 198, https://journals.sagepub.com/doi/full/10.1111/j.1467-8721.2006.00435.x. See also Kelly Rae Chi, "Why Are Habits So Hard to Break?" Duke Today website, January 21, 2016, https://today.duke.edu/2016/01/habits.

[16] Daniel Kahneman and Amos Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* 47 (1979): 263–291.

[17] "Behavioural Finance," Investopedia website, last modified November 8, 2019, https://www.investopedia.com/terms/b/behavioralfinance.asp.

[18] Daniel Kahneman, *Thinking Fast and Slow* (New York: Farrar Straus Giroux, 2011). For a description on duel-process theories, see Shelly Chaiken and Yaakov Trope, eds., "Preface," in *Dual-Process Theories in Social Psychology* (New York: The Guilford Press, 1999), ix. Here, it is stated that "[d]ual-process models […] all share the basic assumption that two qualitatively different modes of information processing operate in making judgements and decisions and in solving problems. In essence, the common distinction in dual-process models is between a fast, associative information-processing mode based on low-effort heuristics, and a slow,

of manipulation, Richard Thaler and Cass R. Sunstein also refer to the term "nudge," which illustrates humans' susceptibility to manipulation: "[a] nudge, as we will use the term, is any aspect of the choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives."[19] Therefore, by understanding the cognitive methods humans use to make decisions, attempts can be made to alter these decision-making processes.

In summary, then, the data people leave behind in the digital environment, when mined by machine learning mechanisms, can reveal much about their emotional, physical, and physiological circumstances. This information, combined with knowledge from the cognitive sciences, in turn reveals two important notions: the first is that people are not especially rational and the second is that people are exposed to being manipulated. This, in turn, leads to a number of potential harms.

Harms Associated with Digital Technologies

The reliance merely on data to predetermine the interaction between humans may have negative repercussions. Data may be defective or tainted with bias, which in turn will result in the technology reflecting this same bias. In addition, any technology may reflect the bias of its developers, be this intentional or unintentional. Also, long-term harms from the use of technology may not be apparent in the short term. Another problem with decision-making systems based on machine learning is called "overfitting," which occurs when a predictive model is faced with

---

rule-based information-processing mode based on high-effort systematic reasoning. Related dual-processing perspectives distinguish between controlled versus uncontrolled, conscious versus unconscious, and affective versus cognitive models of processing." The first mode of information processing is the intuitive and automatic way of thinking, which is fast and instinctive. It occurs almost as a reflex and is associated with a gut reaction. The second is the reflective and rational way of thinking, which is deliberate, self-conscious and associated with conscious thought.

19 Richard Thaler and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth and Happiness* (New York: Penguin Books, 2008), 6.

a situation not found in the training data and must fit this new set of circumstances to a rule that does not quite fit the new set of circumstances.[20] Another harm referred to is that of "self-fulfilling prophecies."[21] The rationale is that, if a predictive model designates a person to be a credit risk due to a poor economic situation, no credit institution will dare give that person credit, an effect being that the person's economic situation will more than likely deteriorate (most people requiring credit to make something of their lives). Those who operate the predictive model will argue that the predictive model was correct in that it predicted the deterioration of the person's economic situation. But a question arises: was the person's deteriorating economic situation predicted by the model or caused by it? In other words, predictive models can "[create] the situation they claim merely to predict."[22] Finally, a harm that arises with the use of predictive models is "what to do when the computer says 'no!'." In other words, having been on the receiving end of a negative and potentially incorrect decision taken by a predictive model, what can a person do to rectify the situation if they even are aware of this situation?

The main overarching risk with the use of the above represented technologies is the potential harm to human autonomy. In their contact with diverse decision-making systems daily, humans are constantly being served information, data, pictures, music, and other content that commercial entities think they will like, the aim being customer satisfaction. This results in humans seeing the world though the lenses provided by the operators of the technology, which ultimately affects human autonomy.[23]

---

[20] Foster Provost and Tom Fawcett, *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking* (Sebastopol, CA: O'Reilly Media, 2013), 111.

[21] Danielle Keats Citron and Frank A. Pasquale, "The Scored Society: Due Process for Automated Predictions," *Washington Law Review* 89 (2014): 18, https://ssrn.com/abstract=2376209.

[22] Citron and Pasquale, "The Scored Society," 18.

[23] For a more in-depth analysis of the potential harms associated with predictive modeling, see Stanley Greenstein, *Our Humanity Exposed: Predictive Modelling in a Legal Context*. Dissertation (Stockholm: Stockholm University, 2017), http://su.diva-portal.org/smash/record.jsf?pid=diva2%3A1088890&dswid=2661.

The Traditional Legal Science Approach

The law is a mechanism for solving problems. In applying the law to societal problems, lawyers, judges, and legislators generally apply the "traditional legal science method" in order to determine the existing law.[24] This method incorporates a "legal dogmatic" approach. It is also described as an approach to legal problem-solving taught to law students, providing them with a number of tools for problem-solving, such as interpretive arguments, modalities of decision, and conflict-solving maxims.[25] It is important to concede that there is no single legal method that is universal, transcending all legal systems, and that, in fact, the contrary is true, namely that there are many diverse legal methods, each having a different point of emphasis.[26] However, here three aspects are noted: first, the use of a legal method other than the traditional legal science method is usually done as an "add-on" (i.e., an extension of the traditional legal science method); second, the study of alternative legal methods remains restricted to the academic realm; and, third, legal practitioners are more adept at applying the traditional legal science method than they are at describing it, and there is little unanimity surrounding a single definition. Therefore, the contention here is that the traditional legal science method dominates the Swedish legal realm.[27]

Kleineman provides a comprehensive analysis of the traditional legal science method.[28] Its application can be explained as a

[24] This chapter relies on the broad generalization that most legal practitioners (scholars, practicing lawyers, and judges) in the Swedish context, confronted with a problem, will automatically apply the traditional legal science method in order to establish the law as it is.

[25] Torben Spaak, *Guidance and Constraint: The Action-Guiding Capacity of Theories of Legal Reasoning* (Uppsala: Iustus Förlag, 2007), 12, footnote 5.

[26] Maria Nääv and Mauro Zamboni, eds., "Sammanfattning," in *Juridisk metodlära* (Lund: Studentlitteratur, 2018), 17–20.

[27] Minna Gräns, "Om hjälpvetenskapernas betydelse för rättstillämpning och rättsvetenskapen," *Juridisk tidskrift*, no. 3 (2006–07): 791: "[i]n Sweden, most legal scientists formulate their problems from a traditional legal dogmatic perspective by using conventional legal dogmatic methods." (Paraphrased loosely from the Swedish by the author.)

[28] Jan Kleineman, "Rättsdogmatisk metod," in *Juridisk metodlära*, eds. Maria Nääv and Mauro Zamboni (Lund: Studentlitteratur, 2018), 21.

solution to a legal problem by means of the application of a legal rule, the point of departure being the principles for the use of the commonly accepted sources of law.[29] The goal is the establishment of the law independently of any other scientific discipline and without taking policy considerations or ethical considerations into account.[30] Therefore, in seeking the solution to a problem, the main sources of law are laws, preparatory works, legal precedent (case law), and legal dogmatic literature.[31] When faced with a problem, the main goal is the application of the abovementioned sources of law in order to determine "the law as it is."[32] Of these sources of law, it is laws and case law from the formal judicial instances that hold the greatest formal authority, while legal dogmatic doctrine (the writings of specialists) carries weight to the extent that the argumentation it puts forward is persuasive.[33] This method aims at acquiring a "coherent picture of the law [...] presenting the law as a network of principles, rules, meta-rules, and exceptions, at different levels of abstraction, connected by support relations," this achieved not only by means of description and logic but also by including evaluative or normative steps.[34] Consequently, it is necessary to have a solid knowledge of the law as well as of the general principles that are accessed by means of studying preparatory works, precedent, and literature.[35] The traditional legal science method is to a certain degree rigid, constraining a judge, the aim being predictability and equality before the law.[36]

This article is in Swedish and its content has been roughly translated by the author.

29 Kleineman, "Rättsdogmatisk metod," 21.

30 Minna Gräns, "Allmänt om användning av andra vetenskaper inom juridiken," in *Juridisk metodlära*, eds. Maria Nääv and Mauro Zamboni (Lund: Studentlitteratur, 2018), 429.

31 Alexander Peczenik, "A Theory of Legal Doctrine," *Ratio Juris* 14, no. 1 (March 2001): 78, https://onlinelibrary.wiley.com/doi/epdf/10.1111/1467-9337.00173.

32 Kleineman, "Rättsdogmatisk metod," 26.

33 Kleineman, "Rättsdogmatisk metod," 28.

34 Peczenik, "A Theory of Legal Doctrine," 79.

35 Kleineman, "Rättsdogmatisk metod," 21.

36 Spaak, *Guidance and Constraint*, 43.

An aspect of the legal dogmatic method concerns questions of interpretation, which involves the abstract task of identifying the legal rule that is relevant to a legal problem, to describe it as well as its relevance to that situation but also to explain how that legal rule should be applied to that legal problem.[37] Considering the formal authoritative weight assigned to law and case law, and the fact that any legal solution to a problem requires a basis from within the above central sources of law, Kleineman asserts that alternative academic methodologies, for example from within the social sciences, are regarded as alien.[38]

Two concepts central to the notion of argumentation are "the law as it is" (*de lege lata*) and "the law as it should be" or "justified recommendations for the lawgiver" (*de lege ferenda*). In researching the law, both these concepts are used to produce coherent theories, from both the enacted law (statutes and judicial decisions) and evaluations thereof. However, it should be noted that research *de lege lata* has a greater importance and weight attached to it.[39] Also, these concepts determine the extent to which an argument is permitted.[40] The formalism surrounding laws and court decisions prohibits a critical analysis of the law, while the accepted writings of legal scholars do provide a mechanism for a criticism of the law. However, it can be problematic putting forward arguments from a legal policy perspective without taking a position in relation to the law as it is.[41] In other words, even the most innovative and creative ideas on the law must receive legitimacy by applying them to the law as is. This in turn places restrictions on the capabilities of the law to provide creative solutions to problems.

Kleineman indicates that it is not uncommon that the law is combined with other subjects from within the social sciences, for example law and philosophy; however, it is primarily the traditional legal science method that is still applied and, even though alternative disciplines can identify inadequacies with the law, the

---

37 Kleineman, "Rättsdogmatisk metod," 30.
38 Kleineman, "Rättsdogmatisk metod," 29.
39 Peczenik, "A Theory of Legal Doctrine," 79–80.
40 Kleineman, "Rättsdogmatisk metod," 36.
41 Kleineman, "Rättsdogmatisk metod," 44.

primary objective is to establish the law as it is. The consequence is that it becomes imperative to distinguish the legal dogmatic and sociological argumentation. For example, a study to quantitively determine how many times a court has determined a case in a certain manner does not necessarily identify the law as it is. It may be that, of the judgments examined, none has the status of precedent, the consequence being that what was assumed to be the law as it is is overturned by a single judgment of the High Court, which has the authority of a precedent-giving institution.[42] It can be argued, however, that there are insights to be gained from quantitative analyses, although the traditional legal science method may look somewhat unapprovingly to incorporating such analyses in legal argument.

A central question put by Kleineman is whether it is possible to incorporate achievements from the social sciences into the application of the traditional legal science method. The answer provided is that this is possible when examining the law from a critical perspective but not when determining the law as it is, a compounding factor being that there is an unwillingness of legal practitioners to include achievements from the social sciences in the application of the traditional legal method, resulting in a mistrust of the legal fraternity.[43] This highlights two important issues. The first relates to what has been reiterated above, namely that the traditional legal science approach is reluctant to consider influences from other disciplines. The second is the mistrust that this creates in relation to the legal fraternity. Both these issues in turn make it difficult for legal researchers and practitioners potentially to work with representatives from other disciplines in finding creative solutions to the complex problems that technology creates but also in embarking on interdisciplinary initiatives.

Illuminating Legal Informatics

Legal informatics is a branch of traditional legal science, which means that "problems are defined and dealt with according to

---

[42] Kleineman, "Rättsdogmatisk metod," 38–39.
[43] Kleineman, "Rättsdogmatisk metod," 41.

criteria, which the legal community consider relevant and comprehensible."[44] Seipel states that it "strives to go beyond traditional, text oriented analysis of valid law, (normative or 'dogmatic' legal science)."[45] This in turn is done by providing perspectives from the field of informatics, which can be described as an academic discipline that "encompasses many fields where information plays a central role; system theory, computer science, communication theory, information security theory, cognitive science, and library science, to mention a few" and "is intrinsically associated with higher ambitions and a strive to develop a theoretical platform that extends beyond traditional (dogmatic) legal science."[46] Legal informatics, therefore, can be described as a legal method that incorporates yet advances the traditional legal science methodology.

A central pillar of legal informatics is the study of the relationship between two areas, namely law on the one hand and technology (represented by ICT [information and communication technology]) on the other.[47] Central to legal informatics is the subfield called "law and ICT," which is characterized by the study of this bidirectional relationship between these two areas, law and ICT. This bidirectional relationship can be formulated in two ways: first in terms of the legal regulation of ICT, that is, the use of substantive law to regulate technology, and, second, "the use of ICT for legal purposes," that is, the use of technology for regulatory purposes.[48] Put another way, ICT law is a combination of the field of "the legal regulation of computers," or "rules" and the study of the field of "the legal use of computers," namely "tools."[49] At the core of legal informatics, therefore, is the study of the intersection

---

[44] Peter Seipel, "IT Law in the Framework of Legal Informatics," in *IT Law*, ed. Peter Wahlgren, Scandinavian Studies in Law 47 (Stockholm: Scandinavian Institute for Scandinavian Law, 2004), 32.

[45] Seipel, "IT Law in the Framework of Legal Informatics," 32.

[46] Seipel, "IT Law in the Framework of Legal Informatics," 33.

[47] Peter Seipel, "ICT Law – A Kaleidoscope View," in *Information & Communication Technology: Legal Issues*, ed. Peter Wahlgren, Scandinavian Studies in Law 56 (Stockholm: Scandinavian Institute for Scandinavian Law, 2010), 37.

[48] Peter Seipel, ed., "Law and ICT: A Whole and Its Parts," in *Law and Information Technology: Swedish Views*, Swedish Government Official Reports 2002:12 (Stockholm: Fritzes offentliga publikationer, 2002), 23.

[49] Seipel, "IT Law in the Framework of Legal Informatics," 33.

between these two branches.[50] Legal informatics therefore by definition encapsulates a two-way perspective, the notion referring not only to the manner in which ICT affects society ("tools") but also the manner in which society impacts on technology ("rules"). An example is the study of the extent to which social choices, preferences, and tradition affect technology (and the interaction of this bidirectional relationship).[51] In other words, legal informatics examines "how the relationships between legal regulation and technical tools ought to be dealt with."[52]

In theorizing on how to examine ICT law, a question put by Seipel regards whether there is a link between examining ICT law from a "use" point of view as well as from a "regulation" point of view.[53] The common denominator, according to Seipel, is that both require a solid understanding of ICT, more specifically ICT in the legal perspective.[54] A superficial understanding of ICT and how it interacts with the law does not suffice—rather, a deeper understanding of this interaction is required.[55] Here reference is made to the use of ICT by creating new infrastructures that subsequently become a legal concern. It is this relationship that allows for the considering of issues that are not strictly speaking "legal issues" yet are important from the regulatory perspective.[56] It is best put by Seipel himself:

> In this way it is signalled that rules and tools constitute a dynamic whole, that rule elements and tool elements are interconnected, and that a deep understanding of law and IT is related to both

[50] Seipel, "ICT Law – A Kaleidoscope View," 41.

[51] Peter Seipel, "Legal Informatics Broad and Narrow," in *Legal Management of Information Systems: Incorporating Law in e-solutions*, ed. Cecilia Magnusson Sjöberg (Lund: Studentlitteratur, 2005), 25. In addition, for an in-depth and more practical example of the application of a legal informatics approach, see Cecilia Magnusson Sjöberg in this volume. Here Magnusson Sjöberg applies the bidirectional approach in the examination of the interaction of technology, in the form of algorithms and machine learning and the law, represented by the notions of data protection and privacy.

[52] Seipel, "IT Law in the Framework of Legal Informatics," 32.

[53] Seipel, "Law and ICT: A Whole and Its Parts," 26–27.

[54] Seipel, "Law and ICT: A Whole and Its Parts," 26–27.

[55] Seipel, "Law and ICT: A Whole and Its Parts," 25.

[56] Seipel, "Law and ICT: A Whole and Its Parts," 25.

> rule and tool elements. In other words, in order to understand the interplay of law and IT, it is necessary to consider both rule and tool aspects and only a combination of the two can lead to a full understanding.[57]

The above is illuminated by Seipel when he states that a new technology may create possibilities to improve the application of a legal right, such as the legal right to access information, while an existing regulation may prohibit this same right, considering it harmful or risky. Also of importance is the actual degree of interplay between law and ICT. Seipel, in referring to the "and" in "law and ICT" states that it signifies "interplay," "interaction," and "mutual dependencies."[58] A consequence of this necessary way of thinking is that both the theoretical and practical facets of this relationship must be studied, one example of a theoretical facet being how the automation of information processing affects legal thinking.[59] In addition, it is not random phenomena that require scrutiny but, rather, there are a number of general categories that comprise the field of information technology: automation, information, communication, integration, penetration, and sensation.[60] Sensation concerns how information processing tools interact with human senses, experiencing, and thinking. Two further attributes constitute legal informatics as a result of this back-and-forth examination of "rules" and "tools." First, concepts and definitions play an important role and are in focus and must continually be examined in the light of the "rules" versus "tools" distinction. Second, it fosters an interest in the information processing structures of society, where topics such as power relations based on information processing and legal steering of the information society are addressed.[61]

Central to legal informatics is that, rather than solving problems reactively and after the fact (*ex post*), there is an element of

[57] Seipel, "IT Law in the Framework of Legal Informatics," 35.

[58] Seipel, "Legal Informatics Broad and Narrow," 26. For a more in-depth discussion of the "and" in "law and ICT," see Seipel, "IT Law in the Framework of Legal Informatics," 35.

[59] Seipel, "IT Law in the Framework of Legal Informatics," 35–36.

[60] Seipel, "IT Law in the Framework of Legal Informatics," 36. The description of what these categories entail is provided by Seipel.

[61] Seipel, "IT Law in the Framework of Legal Informatics," 38.

problem-identification and problem-solving in advance (*ex ante*). A commonly used concept in this regard is "proactive law."[62] There are a number of ways of describing proactive law: as a language for problem formulation, analyses, and theory-building in a fragmented environment and where scholars from different disciplines are able to communicate; another way is as a "perspective" or "world view." An advantage with the latter is that it negates the view that proactive law is something totally novel and a threat to "traditional law."[63] The notion of perspective is also important from the point of view that "[e]ven small changes of perspective can make us see things differently and in a new way."[64] An important basis accepting proactive law is derived from the manner in which one views the function of law. In this regard, a useful starting point is to identify legal systems as dynamic and to view "law as a conceptual system subject to change and law as a system intended to produce as well as to accommodate changes in social structures."[65]

Legal informatics is also concerned with looking into the future and Seipel refers to the field of "legal futurology," where traditional legal research could be complemented with a "prognostication of developments in the legal system and with future-oriented policy issues."[66] Seipel therefore suggests, in addition to the established categories of law *de lege lata* and *de lege ferenda*, a third category of compartmentalization of the law, namely *lex ponderanda*, meaning "probing law" or "speculating law," it also being described as "a speculative, critical analysis of the law," reflecting the proactive way of working.[67]

Legal informatics is also influenced by a number of elements that are central to the law and ICT bidirectional relationship,

[62] Seipel, "IT Law in the Framework of Legal Informatics," 40. See also, Peter Seipel, "Nordic School of Proactive Law Conference, June 2005: Closing Comments," in *A Proactive Approach: Law Libraries*, ed. Peter Wahlgren, Scandinavian Studies in Law 49, (Stockholm: Stockholm Institute for Scandinavian Law, 2006), 360.

[63] Seipel, "Nordic School of Proactive Law Conference," 359–360.

[64] Seipel, "Nordic School of Proactive Law Conference," 359.

[65] Seipel, "Nordic School of Proactive Law Conference," 362.

[66] Seipel, "Nordic School of Proactive Law Conference," 362.

[67] Seipel, "Law and ICT: A Whole and Its Parts," inside back cover. See also Seipel, "Nordic School of Proactive Law Conference," 362.

namely: automation, information, communication, integration, and sensation.[68] The element of sensation is relevant to the extent that human beings sense with their brain and body but also with their tools and that there is no dividing line between the "inside" and "outside" of a person's mind.[69] Seipel states,"[a]s for ICT, we are only beginning to understand the consequences. And a legal understanding hardly exists."[70]

The historical perspective of legal informatics also gives it its distinct nature. It is the acceptance of subjects such as legal thinking based on quantitative and formal reasoning, by Loevinger in 1949, and the use of cybernetics in law, by Wiener in 1954, which can be seen as the forerunner to legal informatics.[71] It is therefore argued that the greater acceptance of such subjects by legal informatics gives it access to a larger body of knowledge for solving problems but also endows the legal profession with a greater receptiveness toward solutions that may not necessarily fall within the strict confines of the traditional legal science method of addressing problems.

A final insight that is illuminated by Magnusson Sjöberg relates to the fact that, in studying the bidirectional relationship between technology and the law, it is not only technology that bears inherent complexities. A study of the law too may reveal that there are multiple relevant regulatory frameworks that must be considered, and that these regulatory frameworks may not necessarily be in alignment with each other, in turn compounding the complexities.[72]

Enriching the Tools for a Legal Analysis

From the above it is clear that there are some major differences between the traditional legal science method and legal informatics.

[68] Seipel, "Law and ICT: A Whole and Its Parts," 22–23.
[69] Seipel, "Law and ICT: A Whole and Its Parts," 22.
[70] Seipel, "Law and ICT: A Whole and Its Parts," 22.
[71] Seipel, "IT Law in the Framework of Legal Informatics," 43. Here reference is made to Lee Loevinger, "Jurimetrics the Next Step Forward," *Jurimetics Journal* 12, no. 1 (September 1971) and Norbert Weiner, *The Human Use of Human Beings. Cybernetics and Society* (London: Eyre and Spottiswoode, 1954).
[72] Cecilia Magnusson Sjöberg in this volume.

The former remains a static, rigid, and relatively closed system for ascertaining the law as it is, while legal informatics is a more dynamic system, open to external influences in the form of knowledge from other sciences, and potentially having a more liberal attitude toward the function of the law.

Having said this, it may very well be that there are valid reasons for any legal system to be rigid in nature and that the traditional legal science method promotes these qualities with a purpose in mind. As part of its problem-solving function, legal systems are required to promote stability. Legal systems usually incorporate the values, morals, and norms held dear by a society. These in turn find their expression in the form of legal rights and duties and entrench values such as dignity, autonomy, and privacy, to mention but a few examples. A well-functioning society also requires predictability, for example in economic matters but also criminal matters. Consequently, a person must be aware that, if he or she performs a criminal act, a punishment will follow but also that entering into a contract will have predictable consequences that can be relied on. The main point argued here is that legal rigidity does have an important function so that the common societal values that have been worked out over hundreds of years are not discarded overnight. It would also be detrimental to society if these values fluctuated from one day to the next. For example, the law should have an internal rigidity that does not sway under popular incentives in, say, times of economic hardship. Simply put, there is a very legitimate reason underlying the composition of the traditional legal science method as well as the complexity and lengthy process of traditional legal regulation. However, as argued here, there is the risk that many potential creative solutions addressing the vulnerabilities associated with technology may go wanting to the extent that they are automatically discarded by too restrictive a legal approach.

Probably the most telling distinction between the traditional legal science method and legal informatics is the latter's obligation of focusing on the bidirectional relationship between law and ICT that constitutes legal informatics. This obligation, it is argued, is not inherent in the traditional legal science method and a legal analysis of just "rules" or just "tools" would suffice. Seipel states:

> rules and tools constitute a dynamic whole, that rule elements and tool elements are interconnected, and that a deep understanding of law and IT is related to both rule and tool elements. In other words, in order to understand the interplay of law and IT, it is necessary to consider both rule and tool aspects and only a combination of the two can lead to a full understanding.[73]

In addressing a problem that arises as a result of technology, a legal practitioner applying the traditional legal science method could choose to focus on only the regulation of ICT or only the effect of the increased use of ICT on the law or society, and this would suffice as an acceptable application of the traditional legal science method. However, according to legal informatics, this would not suffice. Legal informatics demands an in-depth investigation of both the technology at the heart of a problem and the legal aspects triggered by the technology.

Another major distinction between the traditional legal science method and legal informatics relates to the degree of receptiveness to research from other sciences. It is argued that there are limitations to treating the law as a system segregated from society. The law does not operate in a vacuum but influences and is influenced by society at large.[74] This in turn leads to the contention that law, having the problem-solving function that it has and being a tool for dealing with reality, should be built upon an as accurate a view of reality as possible and therefore be based on scientific study.[75] Therefore, the law should be dependent on material from other sciences besides law.[76] The extent to which traditional legal science is and should be receptive to other sciences is, however,

---

73 Seipel, "IT Law in the Framework of Legal Informatics," 33.

74 Fredric Korling, *Rådgivningsansvar – särskilt avseende finansiell rådgivning och investeringsrådgivning*. Dissertation (Stockholm: Jure, 2010), 49.

75 See Niklas Luhmann, *Law as a Social System*, trans. Klaus A. Ziegert, eds. Fatima Kastner and Richard Nobles (Oxford: Oxford University Press), 136. "The resulting dissolution of the sharp demarcation between jurisprudence and sociology has given rise, since the beginning of this century, to the hope that sociology will be able to make a contribution to the administration of justice. From the perspective of the law, however, sociology's function remains more that of an auxiliary science." The above author and citation referred to in Korling, *Rådgivningsansvar*, 49.

76 Korling, *Rådgivningsansvar*, 49.

a subject of much debate.[77] While some argue that the influence from other sciences should be limited, there are those that argue that all sciences are potentially relevant for solving legal problems and that there should be no hierarchy as to the importance of these other sciences.[78] In this regard, the law cannot be viewed as an independent phenomenon that is separate from society, a potential result of this being that seemingly rational solutions to problems ultimately lead to irrational and unacceptable consequences.[79] Researchers who rely on knowledge from the other sciences create new knowledge regarding the rationality of the law as opposed to traditional legal dogmatic doctrine, which stays within the frame of the central sources of law and therefore the constructed and closed legal system.[80] The aim of research is to improve not only the law but also the rationality of making decisions in relation

77 Reference is made to the debate within the Swedish context concerning the acceptability of referring to other sciences in order to interpret a legal rule. In this regard, see Lars Heuman, "Hjälpvetenskapernas betydelse för rättstillämpning och rättsvetenskapen," *Juridisk tidskrift*, no. 4 (2005–06): 768. Heuman states that other sciences may be used as a source of law only as the exception and where there exists a legal rule that lacks linguistic precision or where it has certain objectives or requires a balancing of interests (789, paraphrased by the author). Heuman states that if another science, besides legal science, can prove that a legal rule or precedent is built upon incorrect prerequisites, the courts cannot ignore or retranslate the legal rule due to the fact that law and precedent shall be respected even where criticism can be levied at the solutions chosen by the legislator (772, paraphrased by the author). Finally, Heuman distinguishes between different sciences in the event that they are of relevance for legal science, some more suitable than others (789, paraphrased by the author). See also Gräns, "Om hjälpvetenskapernas betydelse för rättstillämpning och rättsvetenskapen," 782, where Gräns states that it is not the linguistic precision of a law that should determine the relevance of other sciences, but rather the extent to which knowledge can be gained from the other sciences in order to interpret that legal rule (786, paraphrased by the author). Finally, for a description of the extent to which traditional legal science already incorporates research from other sciences, see Gräns, "Allmänt om användning av andra vetenskaper inom juridiken."

78 Gräns, "Om hjälpvetenskapernas betydelse för rättstillämpning och rättsvetenskapen," 782.

79 Gräns, "Allmänt om användning av andra vetenskaper inom juridiken," 437.

80 Gräns, "Allmänt om användning av andra vetenskaper inom juridiken," 438.

to reality where the consequences will be realized.[81] The above debate highlights the varying opinions regarding the application of the traditional legal science method. This, it is argued, can result in uncertainty and potentially lead to a more conservative approach to applying the law. Finally, the debate concerning whether law should take other sciences into account might be merely theoretical, some arguing that a more pragmatic approach has long been the norm.[82] What is clear is that it would be foolish to attempt to argue that all practitioners of the traditional legal science method apply it strictly as laid out above. Even among proponents of this method, there are divergent opinions as to whether the traditional legal science method is receptive to other disciplines or not, as well as the extent to which this is so. However, its theoretical basis does lead one to believe that it is more conservative in its acceptance of knowledge from other disciplines and is lacking in its ability to address effectively the challenges associated with complex technologies.

Another advantage of legal informatics, especially in relation to AI, is that legal informatics by its very nature forces an examination of the incorporation of the cognitive sciences in technology. This is done via the concept of "informatics," which, as was mentioned above, incorporates the study of the cognitive sciences.

---

[81] Gräns, "Allmänt om användning av andra vetenskaper inom juridiken," 437–438.

[82] Luhmann, *Law as a Social System*, 36: "In the classical division of labor between jurisprudence and sociology, jurisprudence is concerned with norms, and sociology, in contrast, with facts. The jurist's task is to interpret norms and apply them. The sociologist may concern himself only with the existing context of the law, with its social conditions and consequences. But this classical view was already out of date, if not anachronistic, even at the time when Hans Kelsen gave it its most precise formulation. 'Social-engineering' approaches and the jurisprudence of interests had tied the application of law to facts that had not been taken into account in formulating norms but instead had to be ascertained subsequent to the formulation of the legal text. Pragmatism had postulated that all practical application of the law should consider how different constructions of the law would affect legal outcomes; it was concerned not only with the impact on future decisions within the legal system but also with controlling actual consequences within social reality." The above author and citation referred to in Korling, *Rådgivningsansvar*, 49.

Another argument favoring legal informatics is that the traditional legal science method should not take precedence automatically, which sometimes seems to be the case. In other words, the knee-jerk reaction of automatically merely applying the traditional legal science method should be reflected upon. In fact, it is stated that the traditional legal science method should never prevail at the expense of other legal processes or methods that solve a problem in a better fashion, while at the same time upholding basic quality demands that legal doctrine requires.[83] This means that the legal method that best solves a problem while at the same time upholding the required principles of law, for example the rule of law, must be applied.

In addition, applying the traditional legal science method potentially limits the notion of what law is. This method only acknowledges a limited number of central legal sources, which can be characterized by their form, namely black ink on white paper. In other words, the traditional attitude is that laws are only that which is written in natural language. However, it is argued that legal informatics allows for and even promotes a broader notion of what the law is. This notion incorporates the view that anything that regulates human behavior should be recognized when addressing regulatory aspects of the law. Therefore, programming code, technical solutions, standards, and ethical codes should also be incorporated in the notion of law due to regulatory effect. This can in turn be connected to the above statement regarding the choice of the best solution to a problem principle, where alternative forms of law, such as soft law, should be the first-choice solution.

A final argument concerns the need for speed. In other words, considering the pace at which technologies such as AI develop, to what extent is the traditional legal science method able to keep abreast of these developments? The response once again is offered by Seipel, who states that it is:

> not enough to rely on slow and minimal adjustments of legal terminology *et cetera*. Whole legal norm structures may need to

---

[83] Wahlgren, "Automatiserade juridiska beslut," 404–405.

be reconsidered, including such elements as basic aims and ideas, stake holders, means of legal steering, and ways of implementation (how to reach goals).[84]

Conclusions

This chapter began with an illustration of how complex the digital environment has become by describing the use of AI in conjunction with the cognitive sciences. By doing so it highlighted the complexity of the risks and problems associated with these technologies and in doing so the necessity for complex solutions. This chapter then set out to describe the traditional legal science method in the context of law as a mechanism for problem-solving. It then proceeded to elevate legal informatics as a legal approach, extending the traditional legal science method and promoting its suitability for solving complex societal problems in the context of complex technologies. In doing so, legal informatics was compared to the traditional legal science method, a legal method that legal informatics is built upon but also extends. What this chapter has hopefully illuminated is that, unlike technology, the subjective nature, composition, and application of the law as a phenomenon is as far from the digital "on or off," nature of technology as could possibly be. There is no "right or wrong" or "good or bad" alternative and the choice of legal method and approach depends on personal preferences and on subjective (and potentially even biased) considerations. The aim of the chapter was to illuminate the strengths of legal informatics due to its acceptance of other sciences and due to its interdisciplinary nature. In other words, there is a point to accepting knowledge from other sciences into the legal realm in order to make sound legal judgments that mirror reality. The extent to which this ought to occur may for all intents and purposes remain the topic of many a debate. What is certain, however, is that legal informatics is an approach that by default advances the ideals of the digital human sciences. It advances the applicability of law as a solution to the challenges associated with the complexity of digitalization with not only a focus on technology itself but also its two-way relationship with society.

---

[84] Seipel, "ICT Law – A Kaleidoscope View," 37.

Furthermore, it is more receptive to interdisciplinary cooperation, thereby facilitating research initiatives that transcend the rigid borders of various disciplines and advocate solutions based on reality.

References

Chaiken, Shelly, and Yaacov Trope, eds. “Preface.” In *Dual-Process Theories in Social Psychology*, ix. New York: The Guilford Press, 1999.

Chi, Kelly Rae. “Why Are Habits So Hard to Break?” DukeToday website, January 21, 2016. https://today.duke.edu/2016/01/habits.

Citron, Danielle Keats, and Frank A. Pasquale. “The Scored Society: Due Process for Automated Predictions.” *Washington Law Review* 89 (2014): 1–33. https://ssrn.com/abstract=2376209.

“Elements of AI.” Online Course on Artificial Intelligence. https://www.elementsofai.com.

English Oxford Living Dictionaries. Entry “Algorithm.” https://en.oxforddictionaries.com/definition/algorithm.

Finlay, Steven. *Predictive Analytics, Data Mining, and Big Data: Myths, Misconceptions and Methods*. Basingstoke: Palgrave Macmillan, 2014.

Frankish, Keith, and William M. Ramsay, eds. *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press, 2014.

Greenstein, Stanley. *Our Humanity Exposed: Predictive Modelling in a Legal Context*. Dissertation, Stockholm: Stockholm University, 2017. http://su.diva-portal.org/smash/record.jsf?pid=diva2%3A1088890&dswid=2661.

Gräns, Minna. “Om hjälpvetenskapernas betydelse för rättstillämpningen och rättsvetenskapen.” *Juridisk tidskrift*, no. 3 (2006–07): 782–792.

Gräns, Minna. “Allmänt om användning av andra vetenskaper inom juridiken.” In *Juridisk metodlära*, edited by Maria Nääv and Mauro Zamboni, 429–441. Lund: Studentlitteratur, 2018.

Heuman, Lars. “Hjälpvetenskapernas betydelse för rättstillämpning och rättsvetenskapen.” *Juridisk tidskrift*, no. 4 (2005–06): 768–789.

Kahneman, Daniel. *Thinking Fast and Slow*. New York: Farrar Straus Giroux, 2011.

Kahneman, Daniel, and Amos Tversky. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (1979): 263–291.

Kleineman, Jan. "Rättsdogmatisk metod." In *Juridisk metodlära*, edited by Maria Nääv and Mauro Zamboni, 21–46. Lund: Studentlitteratur, 2018.

Korling, Fredric. *Rådgivningsansvar – särskilt avseende finansiell rådgivning och investeringsrådgivning*. Dissertation, Stockholm: Jure, 2010.

Loevinger, Lee. "Jurimetrics the Next Step Forward." *Jurimetics Journal* 12, no. 1 (September 1971): 3–41.

Luhmann, Niklas. *Law as a Social System*, translated by Klaus A. Ziegert, edited by Fatima Kastner and Richard Nobles. Oxford: Oxford University Press, 2004.

Neal, David T., Wendy Wood, and Jeffry M. Quinn. "Habits – A Repeat Performance." *Current Directions in Psychological Science* 15 no. 4 (August 2006): 198–202. https://journals.sagepub.com/doi/full/10.1111/j.1467-8721.2006.00435.x.

Nääv, Maria, and Mauro Zamboni eds. "Sammanfattning." In *Juridisk metodlära*, 17–20. Lund: Studentlitteratur, 2018.

Peczenik Alexander. "A Theory of Legal Doctrine." *Ratio Juris* 14, no. 1 (March 2001): 75–105. https://onlinelibrary.wiley.com/doi/epdf/10.1111/1467-9337.00173.

Provost, Foster, and Tom Fawcett. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. Sebastopol, CA: O'Reilly Media, 2013.

Seipel, Peter, ed. "Law and ICT: A Whole and Its Parts." In *Law and Information Technology: Swedish Views*, 21–32. Swedish Government Official Reports 2002:12. Stockholm: Fritzes offentliga publikationer, 2002.

Seipel, Peter. "IT Law in the Framework of Legal Informatics." In *IT Law*, edited by Peter Wahlgren, 31–47. Scandinavian Studies in Law 47. Stockholm: Scandinavian Institute for Scandinavian Law, 2004.

Seipel, Peter. "Legal Informatics Broad and Narrow." In *Legal Management of Information Systems: Incorporating Law in e-solutions*, edited by Cecilia Magnusson Sjöberg, 17–36. Lund: Studentlitteratur, 2005.

Seipel, Peter. "ICT Law – A Kaleidoscope View." In *Information & Communication Technology: Legal Issues*, edited by Peter Wahlgern, 34–56. Scandinavian Studies in Law 56. Stockholm: Scandinavian Institute for Scandinavian Law, 2010.

Seipel, Peter. "Nordic School of Proactive Law Conference, June 2005: Closing Comments." In *A Proactive Approach: Law Libraries*, edited by Peter Wahlgren, 359–363. Scandinavian Studies in Law 49. Stockholm: Stockholm Institute for Scandinavian Law, 2006.

Spaak, Torben. *Guidance and Constraint: The Action-Guiding Capacity of Theories of Legal Reasoning*. Uppsala: Iustus Förlag, 2007.

Sunstein, Cass R. *The Ethics of Influence: Government in the Age of Behavioural Science*. Cambridge: Cambridge University Press, 2016.

Thaler, Richard, and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth and Happiness*. New York: Penguin Books, 2008.

Wahlgren, Peter. "Automatiserade juridiska beslut." In *Juridisk metodlära*, edited by Maria Nääv and Mauro Zamboni, 401–427. Lund: Studentlitteratur, 2018.

Wahlgren, Peter. *Legal Risk Analysis: A Proactive Legal Method*. Stockholm: Jure, 2013.

Weiner, Norbert. *The Human Use of Human Beings: Cybernetics and Society*. London: Eyre and Spottiswoode, 1954.

Winter, Simon, and Per Johansson. "Digitalis filosofi: Människor, modeller och maskiner." *SE:s Internetguide*, no. 13, version 1.0 (2009): 21–34.

Legal AI from a Privacy Point of View: Data Protection and Transparency in Focus

Cecilia Magnusson Sjöberg

Topic Introduction and Outline

The title of this book chapter implies that the topic chosen is multifaceted. Here, this means that there is a diversity of components to be considered in order to investigate the problem area. Starting points for forthcoming studies are also consequently artificial intelligence (AI)[1] within the legal domain, from the perspective of

[1] A sample of explanatory references in the field comprises Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (London: Penguin Books, 2017); Olle Häggström, *Here Be Dragons: Science, Technology and the Future of Humanity* (Oxford: Oxford University Press, 2016). See further the Swedish Foundation for Strategic Research, *Livet med AI*, SSF report no. 29 (Stockholm: Swedish Foundation for Strategic Research, 2018); the Swedish Ministry of Enterprise and Innovation, *Regeringens nationella inriktning för artificiell intelligens*, N2018.14 (Stockholm: Swedish Ministry of Enterprise and Innovation, 2018); Vinnova, *Artificiell intelligens i svenskt näringsliv och samhälle: Analys av utveckling och potential*, VR 2018:08 (Stockholm: Vinnova, 2018). A policy initiative worth mentioning in this context is addAI, whose mission can be summarized as follows: "What will it mean to be a human in the future? The Swedish-based initiative addAI is collaboration between experts in academia, government and companies to discuss and explore the impact of smart algorithms and AI on society. Sociology: What are the best ways to interact with AI and how may it change the relations between humans? Law: How much responsibility should AI have? AI and the rule of law? Business: What does a AI strategy mean for an organization or a country?" The author of this chapter is a cofounder of addAI.

How to cite this book chapter:
Magnusson Sjöberg, Cecilia. "Legal AI from a Privacy Point of View: Data Protection and Transparency in Focus." In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 181–202. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.h.

privacy taking data protection and transparency into particular consideration when aiming also for openness.

In terms of an outline, the notion of legal AI[2] will be addressed first. Then attention will be paid to the surrounding privacy framework. Thereafter, challenges of transparency will be discussed. The quest for regulatory management is illustrated by the Swedish case of being a digitalized European Union (EU) Member State including automated procedures and decision-making. Finally, a selection of ways forward will serve as concluding remarks.

The general hypothesis is that legal AI presupposes privacy in the context of personal data processing. This comprises transparency, which is a kind of overall data protection principle associated with openness and access rights that, in turn, needs to be effectively implemented and managed in order to provide legal safety.[3]

The overall methodological approach in this study can roughly be described as both interdisciplinary and multidisciplinary. The interdisciplinary character is shown by the interplay of law (legal

[2] Wikipedia is a questionable fact-finding source. In a legally oriented text, a limited use can be justified when it comes to the technical domain. (Otherwise it can be questioned if a non-techie is competent to choose one particular definition within the field of computer science.) Here, this concerns a general description of what AI means: "In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and animals. Colloquially, the term 'artificial intelligence' is used to describe machines that mimic 'cognitive' functions that humans associate with other human minds, such as 'learning' and 'problem solving'." ("Artificial intelligence," *Wikipedia*, https://en.wikipedia.org/wiki/Artificial_intelligence). For a more official source, see Nationalencyklopedin ("Artificiell intelligens," *Nationalencyklopedin*, https://www.ne.se/uppslagsverk/encyklopedi/l%C3%A5ng/artificiell-intelligens), explaining AI as first intelligence ascribed to computer systems and second as a research field oriented toward computer systems exhibiting intelligent behavior.

[3] According to Article 5.1 a) Regulation (EU) 2016/679 of the European Parliament and of the Council of April 27, 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation): "Personal data shall be: (a) processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency')."

science) and informatics (computer and systems sciences) being integrated with each other.[4] More precisely, the core of this legal field concerns how law can be proactively embedded at early stages of system design, development, implementation, and management.[5] To exemplify, the quite common question concerning whether a formal legal requirement of a signature in a contract can be fulfilled electronically requires both an understanding of technical means and governing legislation. Adequate problem-solving in such a case would also require an understanding of civil law and/or public law in order to conclude whether an electronic signature at a certain security level would be sufficient to meet evidential requirements, *et cetera*.[6] This kind of legal system management must, in practice, be supplemented by substantive IT law regarding how to interpret and apply legal rules and regulations in digital environments such as the internet. It is here that the multidisciplinary character of this legal field becomes visible by way of requiring command of several legal subject areas, such as security legislation, data protection law, contract law and intellectual property rights.[7]

---

[4] This paradigm becomes quite clear in Peter Seipel, *Computing Law: Perspectives on a New Legal Discipline* (Stockholm: Liber Förlag, 1977); Cecilia Magnusson Sjöberg, ed., *Rättsinformatik: Juridiken i det digitala informationssamhället* (Lund: Studentlitteratur, 2021); Peter Wahlgren, ed., *50 Years of Law and IT – The Swedish Law and Informatics Research Institute 1968–2018* (Stockholm: Stockholm Institute for Scandinavian Law, 2018). The interdisciplinary approach has also characterized the Swedish Ministry of Finance inquiry, chaired by the author of this book chapter, about how law can support the digitalization of the public sector of society. Swedish Ministry of Finance Public Inquiry, *Juridik som stöd för förvaltningens digitalisering*, SOU 2018:25 (Stockholm: Swedish Ministry of Finance, 2018). On this theme see also Cecilia Magnusson Sjöberg, "Förvaltningslagen och digitaliseringen," *Förvaltningsrättslig tidskrift*, no. 3 (2018): 519–530; Markku Suksi, "Automatiserat beslutsfattande enligt den svenska förvaltningslagen," *JFT* 154, no. 6 (2018): 463–472.

[5] See attempts and achievements in this direction in Christopher Millard, ed., *Cloud Computing Law* (Oxford: Oxford University Press, 2013).

[6] Yet another perspective that has emerged relatively recently is referred to as legal tech, which can be described as modern business models offering legal services through digital means and methods.

[7] See for instance Daniel Westman, "AI, big data och upphovsrätten," in *Rättsinformatik: Juridiken i det digitala informationssamhället*, ed.

A study of this kind requires certain delimitations. The jurisdictional scope is primarily the EU and its Member State Sweden. However, this does not exclude reflections about the state of affairs in other jurisdictions as well as references to literature published internationally. Technologically speaking, the text has been authored by a "non-techie," which means that the author's expertise does not lie in, for instance, analyses of specific code (computer programs) but rather in the ability to reveal how the development and use of AI has legal consequences. Neither will there be specific attempts to clarify how to logically represent open textures and ambiguities in law. What will be addressed though, is a set of focal points regarding the kind of dynamic algorithms, emanating from machine learning taking advantage of huge datasets commonly referred to as big data. As already stated, the perspective of this development will be privacy and more specifically legal means for accomplishing transparency.

Legal AI

Applying compliant legal AI

With reference to the above-presented scientific approach within the research field of law and informatics, we can distinguish two major aspects of legal AI. First, there is the methodological theme about integrating law into different kinds of AI-based applications. Second, there is IT law oriented toward substantive issues of how to interpret and apply legislation (broadly speaking) as well as case law in digital environments.

Consequently, AI applied in the legal domain has potential to enhance e-government in the context of decision-making. Public information supply is another application area comprising, in particular, conventional legal sources such as legislation, decided court cases, documents reflecting the history of lawmaking, and doctrine authored by legal scholars. In this context, hopes are to improve recall and precision within information retrieval and to make relevance ranking mechanisms more intelligent, not least using probabilistic (statistical) and linguistic methods and

Cecilia Magnusson Sjöberg, 4th ed. (Lund: Studentlitteratur, 2021), 639–668.

advanced mathematics. In order to reuse public sector information (PSI) and generate open data, AI attracts increasing attention.

Of course, legal AI has much to contribute also applied in the private sector of society. With regard to the legal profession, conventional and often quite burdensome due diligence processes associated with mergers and acquisitions, which are nowadays commonly carried out by younger lawyers at business law firms, are in the long run expected to be replaced by AI solutions.[8] Similarly, so-called smart contracts based on blockchain technologies have become topical. Mention should also be made of customer profiling, which is a business activity that is already using AI-based methods for assessments of creditworthiness, *et cetera*.

So, while we have AI applied in the legal domain supporting major aspects of legal system management, there must in parallel be supplementary assessments of whether current AI applications are legally compliant. What about self-driving cars and liabilities?[9] Are pricing algorithms on the competitive market at all permissible?[10] To what extent, if at all, are automated assessments of creditworthiness adherent to data protection regulation? There are many questions open for discussion and a selection must be made that follows from the below.

Problem area

Legal AI is a problem area of great interest to both legal scholars and practitioners. Over the years different questions have attracted attention. This is true also when it comes to descriptions of what more specifically is meant by the notion of "legal AI." Though there are many more or less successful attempts to capture what AI stands for, today there is a variety of concepts that

---

8 See further Richard Susskind, *Tomorrow's Lawyers: An Introduction to Your Future* (Oxford: Oxford University Press, 2017).

9 Read more in the report of the Swedish Ministry of Enterprise and Innovation Public Inquiry, *Vägen till självkörande fordon – Introduktion*, SOU 2018:16 (Stockholm: Swedish Ministry of Enterprise and Innovation, 2018).

10 See further Stanley Greenstein, *Our Humanity Exposed: Predictive Modeling in a Legal Context*. Dissertation (Stockholm: Stockholm University, 2017).

appears to be adequate if not comprehensive. This means that the analysis needs to navigate in a landscape characterized by, in particular, digitalization, automation, robots, and what may be referred to as core AI. For instance, many people refer to AI without being precise about whether the term “automation” in a given context refers to conventional use of static algorithms or dynamic ones based on machine learning and trained data.

From a technical perspective, it is understandably important to be specific about what kind of technology is being referred to, but this does not necessarily apply to discussions and analyses in the legal domain. The point is that in this study it is mostly not necessary to uphold a strict distinction between, for instance, soft and hard AI. (The so-called singularity is an extreme situation when AI is tentatively in control of everything.) Currently, it can be argued that there is a scale of AI that gradually challenges legal infrastructures of different kinds. To summarize, the scope here is broad, allowing an open-minded approach to the topic.

The standpoint above means that certain relevance should be attributed to what may be referred to as the “old school” of AI, thriving some 30 years ago.[11] At that time AI developers were struggling with fundamental tasks of rule-based versus case-based reasoning, discussing forward chaining and/or backward chaining, when using so-called inference engines that could only be applied on relatively modest datasets. At that time, major questions concerned the distinction—if there were one—between a decision support system and a decision-making system, how to understand the expert system label,[12] and what the role of a so-called

[11] See, e.g., Anne von der Lieth Gardner, *An Artificial Intelligence Approach to Legal Reasoning* (Cambridge, MA: MIT Press, 1987); Patrick Henry Winston, *Artificial Intelligence* (Reading, MA: Addison-Wesley Publishing Company, 1984). More modern references can be found in Marcelo Corrales, Mark Fenwick, and Helena Haapio, eds., *Perspectives in Law, Business and Innovation: Legal Tech, Smart Contracts and Blockchain* (Singapore: Springer & Kyushu University, 2019); Marcelo Corrales, Mark Fenwick, and Nikolaus Forgó, eds., *New Technology, Big Data and the Law* (Singapore: Springer & Kyushu University, 2017); Marcelo Corrales, Mark Fenwick, and Nikolaus Forgó, eds., *Robotics, AI and the Future of Law* (Singapore: Springer & Kyushu University, 2018).

[12] See, e.g., Richard Susskind, *Expert Systems in Law: A Jurisprudential Inquiry* (Oxford: Oxford University Press, 1988). See also Cecilia

knowledge engineer would be in practice. Critical factors were no doubt the transformation of law into algorithms that would be coded so that computers could execute the programs on certain data input.

Once again it should be remembered that the overall position in this study is that transparency of transformation procedures of this kind is a condition for rightful privacy in the context of personal data processing taking place in an AI setting. However, as will be further deliberated, transparency conceived as a kind of openness is dependent on the existence of access rights and their implementation in various contexts.

Privacy Framework

Governing normative structure

Without here going into an in-depth analysis, there is no doubt that the current privacy framework[13] is important in a discussion about legal AI. Today's governing normative structure can be summarized as follows. Of major relevance is of course conventional law in terms of primarily legislation, case law, government bills, and other documents reflecting the history of lawmaking, doctrine authored by legal scholars, and contract law. One example is the General Data Protection Regulation (GDPR).

Adding to the picture is nowadays also what commonly is referred to as "soft law." Generally speaking, this expression refers

---

Magnusson Sjöberg, *Rättsautomation: Särskilt om statsförvaltningens datorisering*. Dissertation (Stockholm: Norstedts Juridik, 1992); Peter Wahlgren, *Automation of Legal Reasoning: A Study on Artificial Intelligence and Law*. Dissertation (Deventer-Boston, MA: Kluwer Law and Taxation Publisher, 1992).

[13] Here is not the place to seek understanding and definitions of the notion of privacy beyond a right to be let alone and to have a private sphere. However, a few topical references will be made. One is the Swedish Ministry of Justice public inquiry report about the state of art concerning personal integrity: Swedish Ministry of Justice Public Inquiry, *Hur står det till med den personliga integriteten? – En kartläggning av Integritetskommittén*, SOU 2016:41 (Stockholm: Swedish Ministry of Justice, 2016). Another is an anthology mirroring the modern digital information society: Russel Weaver, Jane Reichel, and Steven Friedland, eds., *Comparative Perspectives on Privacy in an Internet Era* (Durham, NC: Carolina Academic Press, 2019).

to legal steering documents that mostly are not formally binding. In this context, mention could also be made of the independent European Data Protection Board with the overall task of contributing to the consistent application of data protection rules throughout the EU. In addition to some formal decision-making, this is to be accomplished through general guidance and advice and also promoting cooperation between national supervisory authorities, *et cetera*.[14] It is somewhat challenging to add a third perspective of the privacy framework that could be referred to as AI law or digital law. Nevertheless, the major point is to acknowledge the normative steering mechanisms associated with, in particular, dynamic algorithms applied in the legal domain. Furthermore, AI indicates emerging new legal infrastructures past imagination merely a few years ago. Personal data processing comprising very large datasets, which was previously impossible to carry out, is now on the "to-do list" of both private enterprises and public authorities.

Organizational framework

A study of legal AI should not disregard the surrounding kind of organization as a subset of the privacy framework. This relates to the fact that the legal conditions for applying AI vary considerably between the public and the private sectors of society. To briefly illustrate, a public agency must adhere to all public law governing its activities. In addition to constitutional law, this comprises general as well as special administrative procedures legislation, principles of openness and secrecy, and of course data protection rules directed toward authorities specifically. A private party, on the other hand, is not burdened by the same rules and regulations. However, the market needs to be aware of, for instance, consumers' rights, potential liability, competition law, and also data protection regulation applicable to commercially active data controllers and processors.[15] Getting back to the public

[14] See further "About EDPB," European Data Protection Board, https://edpb.europa.eu/about-edpb/about-edpb_en.

[15] See Article 6 of GDPR, establishing legal grounds making processing lawful to certain extent depending on whether the controller is a public agency or not.

sector, it is worthwhile to note legal conditions and constraints also with regard to, for instance, litigation support in courts in comparison to public administration. The latter may in its turn often need to be divided into state applications on the one hand, and municipal ones on the other. This is all due to the fact that the governing legal framework for introducing AI varies considerably between organizations.

Information security

Every comprehensive digitalization effort nowadays gives rise to debates among specialists as well as the general public about how the latest kind of information and communication technology (ICT) will have an impact on society as a whole. Commonly, reflections concern both pitfalls and potentials with regard to anything from freedom of information and expression to job opportunities on the labor market. In this context AI may be referred to as a milestone rarely seen before. To some extent this reflects the overall privacy framework and its aforementioned normative structures. What appears missing, though, or at least too little discussed, concerns the impact of AI on information security and vice versa.[16] Consequently there is a demand for more studies of this perspective. Risk analyses of personal data processing that are directly oriented toward AI applications and managed accordingly therefore appear to be crucial for safe use of this kind of technology.

At a general level, the interplay between information security and data protection applied—also when using AI—could be read as follows: Personal data protection is one way of accomplishing information security. Information security is multifaceted. It covers at least requirements of confidentiality (contractual and/or legislative), integrity (correct/fair), and availability (agreed upon or laid down in law). At the same time, information security is a way of achieving privacy protection. Privacy protection comprises, in its turn, personal data processing (informative privacy) and/

[16] See for example Cyril Holm, ed., *Secure Digitalisation: Nordic Yearbook of Law and Informatics 2016–2018. The Swedish Law and Informatics Research Institute*. Skrifter utgivna av Juridiska fakulteten vid Stockholms universitet nr 86 (Stockholm: Poseidon Förlag, 2019).

or bodily protection. From the above follows that this study concerns privacy in the context of personal data processing.

Transparency Management

Critical factors

The heading of this section indicates that AI proves to be a challenge to transparency.

The underlying argument and associated circumstances can be summarized in the following points:

(a) AI applied in the legal domain requires transformation of law (broadly understood) into algorithms that can be coded and executed by computers.[17]
(b) AI does not merely take advantage of traditional static algorithms but also dynamic self-learning and possibly self-improving ones.
(c) AI is not only about algorithms but also has to do with, for instance, big data management.

Of major concern is how the abovementioned core features challenge transparency, which is a fundamental building block with regard not only to data protection but also to the rule of law as a whole.

Artificial intelligence—at least when it concerns more advanced applications—inherently includes a "black box" of complicated procedures of a kind that not even (the original) programmer is

---

[17] See, e.g., Magnusson Sjöberg, *Rättsautomation*. The legal implications of code as a kind of law are also discussed by Marek Sergot et al., "The British Nationality Act as a Logic Program," *Communications of the ACM* 29, no. 5 (May 1986): 370–386; See further Joe Collenette, Katie Atkinson, and Trevor Bench-Capon, *An Explainable Approach to Deducing Outcomes in European Court of Human Rights Cases using ADFs*, Department of Computer Science, University of Liverpool, April 2020. Read more about ADF (Abstract Dialectical Frameworks): https://cgi.csc.liv.ac.uk/~katie/comma20.pdf. Lawrence Lessig, *Code and Other Laws of Cyberspace* (New York: Basic Books, 1999) and more recently Boris Melvås, "En formaliserad rättsgrammatik," *Förvaltningsrättslig tidskrift*, no. 5 (2018): 937–972.

able to fully grasp. This implies that principles of transparency, legality, and equality are all at danger if clarity about what goes on internally in an AI application is not achieved. Considering the rapid technological development, it does not seem to be any ready-made solution to the problem of hidden data processing. Right now, the least we can do to master this kind of machine learning, based on trained data and encapsulated in a diversity of other data processing functions, is to aim for legally valid transparency management. Put simply: why not let a legal shield enclose the "black box," ensuring at least awareness of legal rules and principles and thus enhancing trustworthy AI?[18] Below follow a few reflections in this direction heading toward transparency management.

Access rights

Given the above starting point that transparency is a condition for privacy in the context of personal data processing based on AI methods, it is relevant to further examine the legal implications. A major keyword in this context is, as already pointed out, openness, which, however, is not equivalent to transparency. This is explained by the fact that an organization may very well be governed by principles of openness but still not provide transparency due to insufficient access rights taking into consideration also their implementation.

Aiming for a holistic approach—here referred to as transparency management—a need for a legal shield emerges (as acknowledged above) in order to partially cope with the AI "black box" problem. Such a legal shield could very well address and hopefully proactively cover a whole set of legal issues depending on the kind of AI application that is at hand. Is it, for instance, dedicated to the

[18] The idea of a legal shield is not new as such but has been introduced within the framework of GDPR and the regulation of transborder flows of personal data (see Articles 44–49). An EU policy initiative worth mentioning in this context is the establishment of a High-Level Expert Group on Artificial Intelligence (AI HLEG): https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence. Among many different deliverables, it is here interesting to note ethical as well as legal guidelines aiming for trustworthy AI.

public and/or private sector of society, does it give rise to questions concerning intellectual property rights such as copyright, is the dimension of international private law central, *et cetera*?

Within the framework of the analysis carried out here, a major legal shield component concerns access rights. In the Swedish legal system, which will serve as an example for embedded law by way of a legal shield adhering to AI applications, there are three major categories of access rights and another one of a somewhat different kind.

Without any type of internal ranking, the first one concerns the Swedish principle of openness laid down in Chapter Two of the Freedom of the Press Act, dating back to 1766. In brief, this right gives anyone, whether a natural person or a legal person, a Swede or a foreigner, for any kind of purpose (nonprofit or commercial interest), a right of access to official documents that are deemed public. The second access right concerns case-based material. With reference to provision 10 of the Swedish Administrative Procedure Act (SFS 2017:900), any party has a right to be made aware of all material that has been added to the case from external sources. The third access right relates to data controllers' information duties and in particular data subjects' right of access according to Article 15 of the GDPR.

As indicated above, there is yet a legal framework that calls for attention in spite of not quite qualifying as an access right *per se*. It concerns the European Directive on the reuse of PSI.[19] This EU PSI regulation provides a legal platform enhancing open data, but also taking into consideration constraints associated with primarily third parties' rights. Important to note here is that the EU PSI legislation does not provide a strict right of access on behalf of the general public. It is rather a kind of law embracing public authorities to engage in open data solutions that will facilitate the digital society in general.

Implementation

The benefits of access rights are no more than what their implementation shows. To summarize, true openness presupposes

[19] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the reuse of public sector information.

transparency that, in its turn, depends on the presence of access rights that are rightfully implemented. Stakeholders are to be found both among the general public and in the private sphere. In practice, it necessitates both adequate rule interpretation and rule application of the legal framework of openness.

From the above follows that there is a whole chain of components that are crucial in order to accomplish openness based on transparency management. This reasoning gets even more complicated when AI is added to the picture. With this in mind, the three (almost four) aforementioned access rights will be commented upon in a more AI-oriented perspective. To begin with, the (Swedish) principle of openness with regard to the right of access to public official documents should be mentioned. Then there is the regulation authorizing parties in administrative procedures to receive all kinds of relevant case material. A fundamental aspect of privacy connected to personal data processing is transparency based on a right of access that once again calls for attention.

(a) One of the conditions in the Swedish Freedom of the Press Act for accessing what is referred to as compilations of data from already existing official documents requires that this be possible to achieve by so-called routine measures. The fundamental law itself does not explain what is to be understood by this concept, but some guidance can be found in the documents reflecting the history of lawmaking, in particular the underlying government bills, and also some case law. More precisely, it is made clear that so-called routine measures should be understood in terms of a simple effort, without any significant costs or other complications on behalf of the public authority keeping the digital compilation in question.

The key AI issue here is the dynamic character of the notion of routine measure. The major reason is that what qualifies as a simple effort in a more traditional digital environment will most probably not give the full picture in a future characterized by AI-based ICT. When (and if) AI is used, much more information will naturally fall within the scope of the principle of openness. Whether this is good or bad is a question with potential political consequences that needs a separate analysis regarding not only privacy but also freedom of expression and information. This also applies to options for a legal shield.

(b) In the Swedish Administrative Procedure Act, the major access right is to be found in 10 §. The context is that of case handling by public agencies and the right of external parties to take part of what has been added to their matter. The overall regulatory approach is characterized by technical neutrality, that is, without specifying any particular digital solution to (electronic) document management in the Act. There is, however, one exception to this and that is the fact that 28 § of the Act explicitly states that decisions can be made automatically. This leads us into yet another AI reflection.

The provision as such does not provide a right of access to a public agency's internal material such as coded algorithms and selected datasets used for machine learning. So what are the legal consequences of public agencies engaging in cross-organizational digital platforms using AI methods? Today there are already indications of governmental agencies being actively involved in sustainable and innovative AI solutions of this kind. This indicates that AI solutions will rapidly fall within the scope of 28 § of the Administrative Procedure Act, regulating automated decision-making. From the perspective of legal system management, this development is also relevant in a discussion about a legal shield embracing the "black box" commonly referred to in discussions about AI.

(c) Shifting the focus to rights of access by a data subject, current information duties on behalf of a controller are primarily to be found in Articles 13–15 and 22(1) of GDPR. Article 13 is applicable where personal data is collected from the data subject and Article 14 where personal data has not been obtained from the data subject but from another source. Article 15(1) h) is triggered as a right of access by a (potential) data subject, that is, a natural person as the GDPR is not applicable to legal persons (Article 1). Article 22(1) h) governs automated individual decision-making, including profiling. All these provisions except Article 15 are for the controller to be aware of and fulfilled on its own initiative. Article 15 is instead triggered upon a request by a data subject.

What strikes as particularly noteworthy from an AI perspective is the requirement, under certain circumstances, to provide meaningful information about the logic involved in the context of automated decision-making in particular (Article 22(1) and (4), Recital 63). This kind of informed logic can be anticipated to cause problems considering that one central feature of AI is nonexplanatory machine learning capacities. The AI approach therefore becomes even more interesting as a tool for safeguarding privacy.

As a consequence, the following question arises: to what extent—if at all—is there reason do make a distinction between the logic behind a certain legal decision and the logic involved in the complete system?

(d) Mention should also be made of the legal framework for open data and reuse of PSI. As already pointed out, this is not an access right for individuals *per se*. Rather, it is a conditional right of access second hand, that is, in situations when public agencies have already assessed that certain information should be accessible for reuse outside the public sector.

One point of bringing AI into the discussion is to shed light on the fact that information resources management, including by way of dissemination of information, will probably be enhanced and more powerful in future digital environments.

Being a Digitalized EU Member State

This section provides a practical example of the challenges of being a digitalized EU Member State. For obvious reasons, the author being a Swede makes Sweden a good object for this minor excursion into the societal development of the digital information society in which AI already plays a central role. First, a few national characteristics will be noted. Throughout the text below AI will be referred to as both a facilitator and a risk.

Sweden has a long history of using personal identification numbers in digital environments. Generally speaking, this kind of data processing takes place without the general public being particularly upset, with some exceptions. One obvious explanation to this is historical reasons associated with the government and other public agencies being trusted. This has enabled early and smooth computerization of different kinds—including record linkages. This applies in particular to national transaction systems related to social insurance, social security,[20] taxation, study administration, *et cetera*. Furthermore, the moderate number of Swedish citizens[21] has favored complete and early use of information technologies. A disadvantage of this early adaptation to ICT is rather unexpectedly that Sweden has had to cope with a digital legacy[22] when entering the AI society.

From a legislative perspective, Sweden has been an early adopter of digitalization. A sign of this is the fact that Sweden was the first country to implement a national Data Protection Act (SFS 1973:289). Proactively, there have also been constitutional amendments in order to keep pace with digitalization. This is, however, not similar to saying that there is no need for a legal shield around Swedish AI applications. On the contrary, it all boils down to Sweden, in spite of being relatively well prepared for AI in society, having a large amount of work to carry out so as to avoid a surveillance society in contrast to a democracy. To a large extent, this is related to automatic decision-making based on personal data processing.

In spite of Sweden's long history of data protection legislation, being an EU Member State has involved quite a few difficult questions. A major concern has, as already pointed to, been and still is how to combine personal data processing with transparency. The scope of the long-established openness principle is quite far away from today's discussions about needs for data ownership and control. Therefore, it cannot be taken for granted that a Swedish legal approach to the mandatory GDPR will hold in front of the Court of Justice of the European Union.

Articles 85 (Processing and freedom of expression and information) and 86 (Processing and public access to official documents) of GDPR are of particular interest here. At first glance, the EU regulation seems to be compatible with Swedish law but a more detailed analysis reveals that it is not obvious how to combine the

---

20 The so-called Robot Ernst is an early illustration of AI in a social security environment taking place in a municipal community. Fredrik Adolfsson, "Robot styr försörjningsstöd i Trelleborg," *Voister*, July 12, 2017, https://www.voister.se/artikel/2017/07/robot-styr-forsorjningsstod-i-trelleborg.

21 Today's migration can no doubt be seen as a critical success factor when it comes to how AI might promote a multicultural digitalized Swedish.

22 This could also be expressed in terms of a technological heritage (history).

right of access to compilations according to the Swedish Freedom of the Press Act with the scope of Article 86 of GDPR. Another example concerns the fact that it is unclear whether the regulation in 28 § of the Swedish Administrative Procedures Act permitting automatic decision-making meets the requirements of national legislation according to Article 22 of GDPR, regulating automated individual decision-making, including profiling. The doubts can partly be explained by the fact that the Swedish Administrative Procedure Act is subsidiary to other deviating national laws, rules, and regulations.

Ways Forward and Final Remarks

With reference to the above, the anticipation is that AI will have an immense impact on privacy-related personal data processing. This is, however, not similar to saying that development will be all good or bad. Instead, potentials and pitfalls depend highly upon how responsibly people and bodies will implement emerging AI. This may in turn be referred to as a kind of digital climate change risking privacy infringements—which is of course different from the current environmental crises affecting mother earth as a natural resource but still very severe. Legal AI calls not merely for sustainable and innovative technical infrastructures but also for legal infrastructures that are fit to master conditions for privacy in an open society not only today but also in the long run.

Through this lens, transparency is a condition for privacy in the context of personal data processing based on AI methods. In practice, this requires openness, which is not necessarily the same as transparency. This has to do with the fact that principles of openness and associated transparency might not reach out sufficiently due to lack of access rights and how those are implemented. Based on this reasoning, a few ways forward will be presented as concluding remarks in recognition of algorithms, machine learning, and big data.

By way of letting law play a proactive role instead of merely a traditional reactive one when things have already gone wrong, transparency issues can be captured at early stages of system design, development, implementation, and further on to the management of applications.

In terms of innovative law, the notion of the "digital person" as a new legal entity could be discussed. The overall idea is that such an approach to AI would supplement long-established categories of "natural person" and "legal person," not least in the context of responsibilities and liabilities of, for example, robots in connection with AI. Being able to explain the logic involved is no doubt crucial for transparency.[23]

Remedies are also important. One approach could be a kind of algorithmic scrutiny oriented toward embedded (substantive) law concerning social insurance, social security, taxation, study administration, *et cetera*. Such an analysis would be based on a legal informatics approach bringing AI to the fore.[24]

What happens in the future is difficult—if not impossible—to foresee. However, one observation is that the impact of ethics seems to increase in the context of AI. This also has consequences for the legal domain. Expressed in another way, traditional law, be it civil or common, as well as the roles of legal professionals acting as judges, attorneys, *et cetera*, might have to step back in favor of ethical advice, vetting, and guidelines.

At the same time, it is of course important to protect and adjust legal safeguards toward what may be referred to as a rule of law 2.0, offering predefined datasets (reducing biased data), capacity restrictions scoring, and exploring different levels of automation, *et cetera*. In this context, the interplay of law and information security is a critical success factor. This all boils down to a quest for a transparent legal shield around the black boxes of AI algorithms.

In addition to the suggested ways forward above, the important understanding of the interplay between privacy, personal data processing, and modern technologies should finally be emphasized. Not least, means and methods for transparency management

---

[23] See further Cecilia Magnusson Sjöberg, "Digitala personer – en ny rättsfigur," in *Människor och AI: En bok om artificiell intelligens och oss själva*, eds. Daniel Akenine and Jonas Stier (Stockholm: Books on demand, 2019), 65–79; Morgan M. Broman and Pamela Finckenberg-Broman, "AI & Lagen – RAiLE© Projektet," *Arkiv Information Teknik*, no. 1 (2019): 18–20, including references representing a critical approach and for further reading in general. For an in-depth analysis see Visa A. J. Kurki, *A Theory of Legal Personhood* (Oxford: Oxford University Press, 2019).

[24] See further Stanley Greenstein in this volume.

appear to be an important task for further studies. In this context, autonomy of technology calls for particular attention and needs to be challenged from multiple perspectives. A legal approach to digital human sciences[25] appears to be a comprehensive resource for research when data subjects are exposed to AI for better or for worse.

References

"About EDPB." European Data Protection Board. https://edpb.europa.eu/about-edpb/about-edpb_en.

Adolfsson, Fredrik. "Robot styr försörjningsstöd i Trelleborg." *Voister*. July 12, 2017. https://www.voister.se/artikel/2017/07/robot-styr-forsorjningsstod-i-trelleborg.

"Artificial intelligence." Wikipedia. https://en.wikipedia.org/wiki/Artificial_intelligence.

"Artificiell intelligens." *Nationalencyklopedin*. https://www.ne.se/uppslagsverk/encyklopedi/l%C3%A5ng/artificiell-intelligens.

Broman, Morgan M., and Pamela Finckenberg-Broman. "AI & Lagen – RAiLE© Projektet." *Arkiv Information Teknik* 2019, no. 1 (2019): 18–20.

Collenette, Joe, Katie Atkinson, and Trevor Bench-Capon. *An Explainable Approach to Deducing Outcomes in European Court of Human Rights Cases using ADFs*. Department of Computer Science, University of Liverpool, April 2020. https://cgi.csc.liv.ac.uk/~katie/comma20.pdf

[25] One approach to digital humanities is found in Per-Olof Erixon and Julia Pennlert, eds., *Digital humaniora – Humaniora i en digital tid* (Gothenburg: Daidalos, 2017). For a somewhat broader approach, see the digital human sciences committee at Stockholm University. The committee's definition, originally expressed in Swedish, may be translated into English the following way: Digital human sciences means interdisciplinary studies of digital artifacts and environments and their meaning for human beings and society. This includes (1) studies concerning actors and their role in the digital society; (2) social and legal aspects of responsibility and ethics; and (3) interaction between human beings in digital systems and between human beings and digital entities.

Corrales, Marcelo, Mark Fenwick, and Helena Haapio, eds. *Perspectives in Law, Business and Innovation: Legal Tech, Smart Contracts and Blockchain*. Singapore: Springer & Kyushu University, 2019.

Corrales, Marcelo, Mark Fenwick, and Nikolaus Forgó, eds. *New Technology, Big Data and the Law*. Singapore: Springer & Kyushu University, 2017.

Corrales, Marcelo, Mark Fenwick, and Nikolaus Forgó, eds. *Robotics, AI and the Future of Law*. Singapore: Springer & Kyushu University, 2018.

Erixon, Per-Olof, and Julia Pennlert, eds. *Digital humaniora – Humaniora i en digital tid*. Gothenburg: Daidalos, 2017.

Gardner, Anne von der Lieth. *An Artificial Intelligence Approach to Legal Reasoning*. Cambridge, MA: MIT Press, 1987.

Greenstein, Stanley. *Our Humanity Exposed: Predictive Modelling in a Legal Context*. Dissertation, Stockholm: Stockholm University, 2017.

Holm, Cyril, ed. *Secure Digitalisation: Nordic Yearbook of Law and Informatics 2016–2018*. The Swedish Law and Informatics Research Institute. Skrifter utgivna av Juridiska fakulteten vid Stockholms universitet nr 86. Stockholm: Poseidon Förlag, 2019.

Häggström, Olle. *Here Be Dragons: Science, Technology and the Future of Humanity*. Oxford: Oxford University Press, 2016.

Kurki, Visa A. J. *A Theory of Legal Personhood*. Oxford: Oxford University Press, 2019.

Lessig, Lawrence. *Code and Other Laws of Cyberspace*. New York: Basic Books, 1999.

Magnusson Sjöberg, Cecilia. *Rättsautomation: Särskilt om statsförvaltningens datorisering*. Dissertation, Stockholm: Norstedts Juridik, 1992.

Magnusson Sjöberg, Cecilia. “Förvaltningslagen och digitaliseringen.” *Förvaltningsrättslig tidskrift*, no. 3 (2018): 519–530.

Magnusson Sjöberg, Cecilia, ed. *Rättsinformatik: Juridiken i det digitala informationssamhället*. 4th ed. Lund: Studentlitteratur, 2021.

Magnusson Sjöberg, Cecilia. "Digitala personer – En ny rättsfigur." In *Människor och AI: En bok om artificiell intelligens och oss själva*, edited by Daniel Akenine and Jonas Stier, 65–79. Stockholm: Books on demand, 2019.

Melvås, Boris. "En formaliserad rättsgrammatik." *Förvaltningsrättslig tidskrift*, no. 5 (2018): 937–972.

Millard, Christopher, ed. *Cloud Computing Law*. Oxford: Oxford University Press, 2013.

Seipel, Peter. *Computing Law: Perspectives on a New Legal Discipline*. Stockholm: Liber Förlag, 1977.

Sergot, Marek, Fariba Sadri, Robert Kowalski, Frank Kriwaczek, Peter Hammond, and Terese Cory. "The British Nationality Act as a Logic Program." *Communications of the ACM* 29, no. 5 (May 1986): 370–386.

Suksi, Markku. "Automatiserat beslutsfattande enligt den svenska förvaltningslagen." *JFT* 165, no. 6 (2018): 463–472.

Susskind, Richard. *Expert Systems in Law: A Jurisprudential Inquiry*. Oxford: Oxford University Press, 1988.

Susskind, Richard. *Tomorrow's Lawyers: An Introduction to Your Future*. Oxford: Oxford University Press, 2017.

Swedish Foundation for Strategic Research. *Livet med AI*. SSF report no. 29. Stockholm: Swedish Foundation for Strategic Research, 2018.

Swedish Ministry of Enterprise and Innovation. *Regeringens nationella inriktning för artificiell intelligens*. N2018.14. Stockholm: Swedish Ministry of Enterprise and Innovation, 2018.

Swedish Ministry of Enterprise and Innovation Public Inquiry. *Vägen till självkörande fordon – Introduktion*. SOU 2018:16. Stockholm: Swedish Ministry of Enterprise and Innovation, 2018.

Swedish Ministry of Finance Public Inquiry. *Juridik som stöd för förvaltningens digitalisering*. SOU 2018:25. Stockholm: Swedish Ministry of Finance, 2018.

Swedish Ministry of Justice Public Inquiry. *Hur står det till med den personliga integriteten? – En kartläggning av Integritetskommittén*. SOU 2016:41. Stockholm: Swedish Ministry of Justice, 2016.

Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence*. London: Penguin Books, 2017.

Vinnova. *Artificiell intelligens i svenskt näringsliv och samhälle: Analys av utveckling och potential*. VR 2018:08. Stockholm: Vinnova, 2018.

Wahlgren, Peter. *Automation of Legal Reasoning: A Study on Artificial Intelligence and Law*. Dissertation, Deventer-Boston, MA: Kluwer Law and Taxation Publisher, 1992.

Wahlgren, Peter, ed. *50 Years of Law and IT – The Swedish Law and Informatics Research Institute 1968–2018*. Stockholm: Stockholm Institute for Scandinavian Law, 2018.

Weaver, Russel, Jane Reichel, and Steven I. Friedland, eds. *Comparative Perspectives on Privacy in an Internet Era*. Durham, NC: Carolina Academic Press, 2019.

Westman, Daniel. "AI, big data och upphovsrätten." In *Rättsinformatik: Juridiken i det digitala informationssamhället*, edited by Cecilia Magnusson Sjöberg, 4th ed., 639–668. Lund: Studentlitteratur, 2021.

Winston, Patrick Henry. *Artificial Intelligence*. Reading, MA: Addison-Wesley Publishing Company, 1984.

Be Careful What You Wish For! Learning Analytics and the Emergence of Data-Driven Practices in Higher Education

Teresa Cerratto Pargman and Cormac McGrath

Introduction

In an age of data measurement,[1] we are witnessing the developments of tools and techniques to capture, transmit, store, and analyze data that enable businesses, governments, healthcare, and welfare institutions to identify "knowledge" about human behavior in unprecedented ways. Big data and analytics offer the promise and potential of providing a better foundation for financial and organizational decisions and streamline, as do practices associated with data more effectively.[2] The higher education sector is no exception to the emerging data-driven practices in society. With the pervasive use of learning management systems (LMSs) facilitating access to and storage of large-scale datasets, higher education institutions (HEIs) have started to pay attention to the promises entrenched in big data and data mining techniques to support learning, teaching, and administrative activities in more efficient ways.[3] In this context, HEIs, particularly in the USA, the

[1] Ben Williamson, *Big Data in Education: The Digital Future of Learning, Policy and Practice* (Los Angeles, CA: Sage, 2017).

[2] Ben Daniel, "Big Data and Analytics in Higher Education: Opportunities and Challenges," *British Journal of Educational Technology* 46, no. 5 (2015): 904–920.

[3] Yi Shan Tsai et al., "Complexity Leadership in Learning Analytics: Drivers, Challenges and Opportunities," *British Journal of Educational Technology* 50, no. 6 (November 1, 2019): 2839–2854.

How to cite this book chapter:
Cerratto Pargman, Teresa, and Cormac McGrath. "Be Careful What You Wish For! Learning Analytics and the Emergence of Data-Driven Practices in Higher Education." In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 203–226. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.i.

UK, and Australia, are implementing learning analytics to understand better and support student learning.[4] Even elsewhere, we see vast investments in learning management systems that enable the analysis of learning behavior by capturing student data on their academic interactions.

The purpose of this chapter is to introduce learning analytics (LA), exemplify how LA has currently been implemented in higher education, and discuss critically the ethical issues and concerns that arise when LA is introduced into HE.

What Is Learning Analytics?

Learning analytics is a fast-developing research area within the field of technology-enhanced learning (TEL) that has emerged during the last decade.[5] In particular, LA has its roots in various fields such as data science, artificial intelligence, practices of recommender systems, online marketing, and business intelligence.[6] The Society for Learning Analytics Research (SOLAR)[7] situates it at the intersection of learning (e.g., educational research, educational technology), analytics (e.g., statistics, visualization, computer/data sciences, artificial intelligence), and human–computer interaction (e.g., usability, participatory design, sociotechnical systems thinking). As a consequence of such a cross-disciplinary characterization, the field has been defined in varied ways. Still, a common understanding is: "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs."[8]

---

[4] Clara Schumacher and Dirk Ifenthaler, "Features Students Really Expect from Learning Analytics," *Computers in Human Behavior* 78 (2018): 397–407.

[5] Rebecca Ferguson, "Learning Analytics: Drivers, Developments and Challenges," *International Journal of Technology Enhanced Learning* 4, no. 5/6 (2012): 304–317.

[6] Yishay Mor et al., "Learning Design, Teacher Inquiry into Student Learning and Learning Analytics: A Call for Action," *British Journal of Educational Technology* 46, no. 2 (2015): 221–229.

[7] Society for Learning Analytics website, https://www.solaresearch.org/about.

[8] George Siemens, and Phil Long, "Penetrating the Fog: Analytics in Learning and Education," *Educause Review* 46, no. 5 (2011): 33.

What is new with capturing a vast amount of student data is the teachers' (and the higher education institutions') chance to design pedagogical interventions based on analyzing the massive amount of data and the links between them. Among the types of analyses allowing for changes in educational practices, SOLAR includes: (1) predicting student success, (2) identifying students at risk of failing or dropping out of their studies, (3) supporting student development of lifelong learning skills and strategies, (4) providing personalized and timely feedback to students regarding their learning, (5) supporting the development of essential skills such as collaboration, critical thinking, communication, and creativity, (6) developing student awareness by supporting self-reflection, and (7) supporting quality learning and teaching by providing empirical evidence on the success of pedagogical innovations.

The technologies of learning analytics

Combining analytics technologies, their mighty analytical power, and cutting-edge data mining techniques with large-scale datasets, learning analytics in higher education practices promises to identify patterns to enhance student learning and make teaching more effective and efficient. In the specific context of higher education, data is generated basically from the use of learning management systems (e.g., ItsLearning), open online learning environments (e.g., Moodle), open social platforms (e.g., LinkedIn), e-learning, intelligent tutoring systems, and forums, among others. Students' data can most often be accessed, combined, and analyzed using complex algorithms to inform decisions.[9] The large-scale datasets, or "big data," are often described according to the following specific attributes: volume (massive size), velocity (of updating data), validity (accuracy), variety (various formats), venue (location), vocabulary (context), and value (usefulness).[10] Anthony G. Picciano explains that, as university student record-keeping systems maintain information on students' grades in each course, institutions

[9] Mervat Adib Bamiah et al., "Big Data Technology in Education: Advantages, Implementations, and Challenges," *Journal of Engineering Science and Technology* 13, special issue on ICCSIT (2018): 229–241.

[10] Kristine Brands and Mark Holtzblatt, "Business Analytics: Transforming the Role of Management Accountants," *Management Accounting Quarterly* 16, no. 3 (2015): 1–12.

could hypothetically use it to trace student performance patterns over time.[11] Similar processes can be illustrated in the analytics applications used in e-commerce.[12] For instance, companies like Netflix, using recommender models and algorithms, examine website traffic, customers' purchases, or navigation patterns to determine which customers are more or less likely to buy a particular product (e.g., series). Based on these patterns, companies like Netflix send notifications to customers of new series or movies as they become available.[13]

Following Picciano, analytics of this sort are beginning to be used in higher education for predicting student performance, outcomes, and persistence. Consequently, capturing the data generated by groups of students attending, for instance, a 15-week online course could generate thousands of transactions per student and can be used to feed an LA application to optimize students' future choices. Such transactions can also be integrated with other data sources (e.g., personal data, health data, police records, bank data, etc.) coming from university information systems or others. This type of real-time monitoring of student transactions can, in turn, generate alerts that give course instructors a chance to intervene to assist the student in time.[14]

Learning analytics: some examples

Among the most common data analytics, monitoring individual student performance is most commonly applied in today's LA. This type of analytics can be better grasped by the example of

[11] Picciano argues that: "In a big data scenario, data would be collected for each student transaction in a course, especially if the course was delivered electronically online. Every student entry on a course assessment, discussion board entry, blog entry, or wiki activity could be recorded, generating thousands of transactions per student per course. Furthermore, this data would be collected in real or near real time as it is transacted and then analyzed to suggest courses of action. Analytics software is evolving to assist in this analysis." Anthony G. Picciano, "The Evolution of Big Data and Learning Analytics in American Higher Education," *Journal of Asynchronous Learning Networks* 16, no. 3 (2012): 9–20, at 12.

[12] Picciano, "The Evolution of Big Data."

[13] Picciano, "The Evolution of Big Data."

[14] Picciano, "The Evolution of Big Data."

diagnostic testing via diagnostic assessment systems. As a teacher, at the end of a learning sequence, let's suppose that you assign your students a diagnostic test. You design your test, choosing 20 questions from the reading material and the material discussed and lectured about in class. You have subdivided the questions into five categories, A, B, C, D, and E. The students do the diagnostic test, and, when you analyze the results, you notice that a majority of students lack knowledge on category D. What you do next is design a follow-up class to address the lack of understanding in category D, assign additional reading to the students, or create a digital learning resource with a video or something similar where you clarify the category in question. As a teacher, you could also use the data to see how your teaching on different themes develops over a more extended time by collecting the same or similar data points. That is a form of LA in its most basic form: it requires data, data analysis, and ideally some change in behavior on behalf of the teacher and/or the student to close potential gaps that arise from the analysis.

In a similar vein, students may be encouraged to self-identify the weaknesses they have in their understanding. Following the same steps as described above, students could do a test and be presented with their score, usually via a dashboard. From the information offered via the dashboard, the student could then identify gaps in their knowledge and then study to reduce those gaps, and, in the best of worlds, the students would be better prepared for exams, score better tests, and so on.

There are also more advanced ways to work with LA. For example, through machine learning algorithms, one could pull together multiple sources of data from students' attendance and interaction on the universities' LMSs, to make predictions about the risk for student dropout or the success or failure of courses and programs. Some of the most common data points include login information, tracking not only when students log in but also how long they interact with teaching material, student performance and activity data on downloads, quizzes, and video views online via the LMS. All of this could also include third-party tools and data from social media.[15] Some examples of

[15] Picciano, "The Evolution of Big Data."

use are highlighted by Picciano, who points to the LA application developed at Rio Salado Community College in Arizona in the USA, which enrolls more than 41,000 students in online courses. Rio Salado has implemented the PACE (Progress and Course Engagement) analytics application, which automatically tracks student progress and prompts teachers to introduce interventions if needed. The PACE application emphasizes personalizing the learning experience, meaning that it helps nontraditional students reach their educational goals through programs and services tailored to individual needs. Another example is Pattern, developed, and implemented at Purdue University. Pattern is a service offered to students to measure study habits and provide analytics and insights to their learning.

Furthermore, the use of predictive analytics, presenting the potential to improve the feasibility of effective early intervention strategies aimed to support at-risk students before they fail, have earlier been identified.[16] Such interventions may include specific recommendations for improvement facilitated by the mapping of student activity and student profiles.[17]

Learning analytics: promises and expectations

LA applications are often presented as offering unbounded possibilities and a grand narrative of modernity that resonate with techno-romanticism's rhetoric.[18] Such narratives have recently gained prominence and can be associated with the idea that access to and analyzing students' data will improve the quality and value of the learning experience within schools and universities.[19] More specifically, the use of LA in universities is expected to

[16] Abelardo Pardo and George Siemens, "Ethical and Privacy Principles for Learning Analytics," *British Journal of Educational Technology* 45, no. 3 (2014): 438–450.

[17] Lynne D. Roberts et al., "Student Attitudes toward Learning Analytics in Higher Education: 'The Fitbit Version of the Learning World,'" *Frontiers in Psychology* 7 (December 2016): 1–11.

[18] Neil Selwyn, *Distrusting Educational Technology: Critical Questions for Changing Times* (New York: Routledge, 2013).

[19] Cormac McGrath and Anna Åkerfeldt, "Educational Technology (Ed Tech): Unbounded Opportunities or Just Another Brick in the Wall," in *Digital Transformation and Public Services: Societal Impacts in Sweden and*

contribute "to quality assurance and quality improvement, boosting retention rates, assessing and acting on differential outcomes for students and as an enabler for the introduction of adaptive learning,"[20] as well as to enhance students' success by identifying students at risk,[21] and consequentially increasing organizational productivity[22] and competitiveness.[23] The use of LA is also expected to facilitate students' informed decision-making, helping them change their learning strategies accordingly, and support self-regulated learning.[24] On this note, Simon Knight et al.,[25] as well as Anna Kruse and Rob Pongsajapan,[26] identify LA as offering much potential to understand how services to students and student learning can be improved.

The promises and expectations associated with LA are ingrained in the novel technical developments in artificial intelligence and the higher education sector's socioeconomic landscape. As such, stories about the progressive label of LA technologies and the emergence of data-driven practices these technologies stimulate may be due to the following reasons: (1) Universities worldwide are operating in an increasingly complex and competitive environment. They are expected to adjust to national and

---

Beyond, eds. Anthony Larsson and Robin Teigland (London: Routledge, 2019), 143.

20 Deborah West et al., "Putting an Ethical Lens on Learning Analytics," *Educational Technology Research and Development* 64, no. 5 (2016): 903–922, at 124.

21 John T. Avella et al., "Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review," *Journal of Asynchronous Learning Network* 20, no. 2 (2016): 13–29.

22 James Manyika et al., "Big Data: The Next Frontier for Innovation, Competition, and Productivity," report from McKinsey Global Institute, May 2011, https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation.

23 David Kiron et al., "Raising the Bar with Analytics," *MIT Sloan Management Review* 55, no. 2 (2014): 29–33.

24 Sharon Slade and Paul Prinsloo, "Learning Analytics: Ethical Issues and Dilemmas," *American Behavioral Scientist* 57, no. 10 (2013): 1510–1529.

25 Simon Knight et al., "Epistemology, Assessment, Pedagogy: Where Learning Meets Analytics in the Middle Space," *Journal of Learning Analytics* 1, no. 2 (2014): 23–47.

26 Anna Kruse and Rob Pongsajapan, "Student-Centered Learning Analytics," *CNDLS Thought Papers* (2012): 1–9.

global economic changes while ensuring that their quality is relevant.[27] (2) Many universities worldwide have adopted learning management systems and are offering online education. As a consequence of the increasing use of these digital learning platforms, it is today possible to access online data repositories and make sense of the data to intervene in current teaching practices. (3) Learning management systems are evolving into more powerful data collection devices with the power to generate clickstream data on all student interactions online, presenting universities with means to better monitor and predict students' learning trajectories and predict teaching and assessment strategies.[28] (4) There are today concrete opportunities for higher education institutions to "go beyond the individual student level and share their disparate datasets or even link data at a federal level which presents further opportunities for analytical insights at an even larger scale."[29] (5) With the increasing trust in data, there is a tendency to "value what we measure rather than to measure what we value." As such, it has become a common practice to "focus discussions about education almost exclusively on the measurement and comparison of educational outcomes."[30] Furthermore, access and analysis of large-scale datasets have been suggested to be the opportunity for higher education to reinvent its business model and making decision processes about educational outcomes.[31]

In this context, it is not surprising that LA is presented as something inherently positive and approached as an excellent means to understand better the complexity inherent to student learning, current teaching practices, and how to contribute to

[27] Daniel, "Big Data and Analytics."
[28] Bamiah et al., "Big Data Technology in Education."
[29] Kyle M. L. Jones, "Learning Analytics and Higher Education: A Proposed Model for Establishing Informed Consent Mechanisms to Promote Student Privacy and Autonomy,"*International Journal of Educational Technology in Higher Education* 16, no. 24 (2019): 1–22, at 2.
[30] Gert Biesta, "Why 'What Works' Won't Work: Evidence-Based Practice and the Democratic Deficit in Educational Research," *Educational Theory* 57, no. 1 (2007): 1–22, at 1.
[31] Daniel, "Big Data and Analytics."

the institution's success rate.[32] However, as pointed out by Paul Prinsloo and Sharon Slade,[33] real-time automatic gathering, processing, storage, and students' data analysis are not neutral acts. Accessing, storing, and analyzing such data entail making a series of several decisions that involve ethical, legal,[34] and moral considerations. In this context, the use of LA in education is contested on several points, and we will return to those shortly.

Learning Analytics and Data-Driven Practices: Understanding the Complexity of a Sociotechnical Phenomenon in Higher Education

Looking at the LA research field's development over time, we observe a gradual shift away from the technical-driven promises associated with LA toward the real and tangible complexities of introducing LA into educational institutions. In this regard, Olga Viberg et al. underscore that, while the literature in the field highlights that LA has the potential to improve learning and teaching, overall there is little evidence that shows such improvements take place.[35] The emerging data-driven practices associated with the use of LA by universities have yet to demonstrate the claimed effectiveness of these technologies for improving student learning, teaching, and, consequently, higher education quality. To date, the use of LA seems to have been realized by the extraction and reporting of student data, instead of "prescribing personalized and timely support strategies that can aid teaching quality and improve student learning experiences."[36] In this regard, Mervat Adib Bamiah et al. add that the use of big data and analytics promised in the adoption of LA by HEIs is today not fully implemented for a diversity of reasons, including "the complexity of

32 Daniel, "Big Data and Analytics."

33 Paul Prinsloo and Sharon Slade, "Big Data, Higher Education and Learning Analytics: Beyond Justice, Towards an Ethics of Care," in *Big Data and Learning Analytics in Higher Education: Current Theory and Practice* (Cham: Springer International Publishing, 2017), 109–124.

34 See Greenstein in this volume.

35 Olga Viberg et al., "The Current Landscape of Learning Analytics in Higher Education," *Computers in Human Behavior* 89 (December 2018): 98–110.

36 Tsai et al., "Complexity Leadership in Learning Analytics," 2840.

identifying the relevant data which requires knowledge, resources and time."[37] Here there is also the need to develop specific competencies such as data literacy and ethical literacy and transparent and human-centered interfaces to LA applications.[38]

Stakeholders in learning analytics

Despite the availability of large-scale datasets, powerful analytical tools, and cutting-edge data mining techniques, deploying LA in the higher education sector is a complex endeavor. Part of the complexity is given by the different types of stakeholders and practices involved in developing, designing, deploying, and using learning analytics technologies.[39] Questions such as: Who gets access to aggregate and communicate the data? What does all this data mean? For what ends is data collected and analyzed? Who benefits from data-driven practices? are all critical questions[40] for teachers, students, administrators, academics, EdTech developers, and university technical staff to discuss within educational communities. Such questions help clarify the various stakeholders' roles and responsibilities in the emergence of data-driven practices in higher education and map the vast array of technologies and practices involved. More precisely, following Olugbenga Adejo and Thomas Connolly, administrators, teachers, and students are acknowledged as the primary stakeholders, followed by course developers/researchers, computer/network administrators, technicians, and data analysts. They are all involved in strategizing about LA implementation and deciding about the importance, acceptance, and suitability of such analytics technologies for the students, their learning, and the learning environment.[41] In turn,

---

37 Bamiah et al., "Big Data Technology in Education," 231.

38 Simon Buckingham Shum et al., "Human-Centred Learning Analytics," *Journal of Learning Analytics* 6, no. 2 (2019): 1–9.

39 Teresa Cerratto Pargman and Cormac McGrath, "Mapping the Ethics of Learning Analytics in Higher Education: A Systematic Literature Review of Empirical Research," *Journal of Learning Analytics* (2021): 1–17.

40 danah boyd and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication & Society* 15, issue 5 (2012): 662–679.

41 Olugbenga Adejo and Thomas Connolly, "Learning Analytics in a Shared-Network Educational Environment: Ethical Issues and Countermeasures,"

the social and epistemic relationships maintained by such diverse stakeholders are configured and shaped by data, files, programs, and other resources used and shared across different systems bound to different educational practices.

Within this complex sociotechnical configuration, two points of view can be distinguished.[42] On the one hand, the point of view focused on the HEI, most often called academic analytics, to support the institutional, operational, and financial decision-making process[43] is a point of view that is concerned with data governance, and, on the other hand, the point of view focused on the students and their rights vis-à-vis their data, which Slade and Prinsloo exemplify in their socio-critical framework for learning analytics.[44]

These—at times conflicting—points of view involve a series of challenges that lie at the core of the deployment of LA applications.[45]

The conflation of diverse types of techniques, applications, multiple stakeholders' roles, and mandates constitutes dynamic, complex, and sociotechnical arrangements that make the appraisal of ethical and moral issues and their implications challenging for education. This is perhaps one reason why so little is still understood about ethical and moral concerns linked to the use of LA and the emergence of data-driven practices in universities.

International Journal of Advanced Computer Science and Applications 8, no. 4 (April 2017): 22–29.

42 Cerratto Pargman and McGrath, "Mapping the Ethics."

43 Celeste Lawson et al., "Identification of 'at Risk' Students Using Learning Analytics: The Ethical Dilemmas of Intervention Strategies in a Higher Education Institution," *Educational Technology Research and Development* 64, no. 5 (2016): 957–968.

44 Slade and Prinsloo, "Learning Analytics."

45 Kaiwen Sun et al. point to the multiple tensions among and within the stakeholder groups (e.g., developers, academic advisors, students) in relation to awareness, understanding, access, and use of the system. These findings resonate with prior observations made with regard to "the ethical and transparent use of the data, student data privacy concerns and rights, considerations of informed consent, data access control, and responsible interpretation of student data." Kaiwen Sun et al., "It's My Data! Tensions Among Stakeholders of a Learning Analytics Dashboard," in *CHI '19: Proceedings of the 2019 Conference on Human Factors in Computing Systems* (New York: Association for Computing Machinery, 2019), 1–14.

Ethical and moral challenges concerning the use of learning analytics in higher education

Issues about ethical and moral considerations that are inherent to the access, storage, and analysis of student data were introduced by Sharon Slade and Paul Prinsloo in 2013, where the authors argued that, although using student data for gaining knowledge about learners' behavior may be advantageous for students, instructors, and institutions, and of value for the understanding of students' learning and construction of didactical interventions, it faces significant ethical and moral considerations. Such considerations are connected to the location and interpretation of data, informed consent, privacy, deidentification of data, and classification and data management.[46] Within this framework, these authors propose (1) understanding LA as "moral practice," (2) addressing students as agents; acknowledging that student identity and performance are temporal dynamic constructs, (3) asserting that student success is a complex and multidimensional phenomenon, (4) putting the focus on transparency, and (5) acknowledging the fact that higher education cannot afford not to use data.[47]

Comparatively, with studies on the use of LA technologies promising enhancement of students' performance, a small number of empirical research studies have engaged with the ethics of LA systems. More particularly, a review of such literature[48] reflects a tension between, on the one hand, a need to find solutions via concrete guidelines to help institutions to be able to deal with new ethical and moral challenges brought by LA and, on the other hand, a need to problematize student data and data mining techniques for higher education learning and teaching purposes.

Two examples are of interest here. One sets out to focus on an institutional perspective[49] or academic analytics[50] that engages

46 Slade and Prinsloo, "Learning Analytics."

47 Slade and Prinsloo, "Learning Analytics."

48 Cerratto Pargman and McGrath "Mapping the Ethics."

49 Niall Sclater, *Code of Practice for Learning Analytics: A Literature Review of the Ethical and Legal Issues* (Jisc, 2014), retrieved at http://repository.jisc.ac.uk/5661/1/Learning_Analytics_A-Literature_Review.pdf.

50 We here refer to the Open University's Policy on Ethical Use of Student Data (Open University, 2014), influenced by Slade and Prinsloo, "Learning Analytics," and the work put forward by Niall Sclater that contributed to

with ethics as barriers that halt the LA field's development and the institutional benefits associated with LA.[51] The other is driven by the interest to raise awareness and engender conversations about opportunities and concerns that come with developing data-driven practices from a student perspective.[52] Interestingly, the latter argues that producing ethical guidelines or policies on the issue only is *insufficient* and points instead to cultivating ethical practices to deal with the use of LA and their effects, including risks for students and teachers in higher education. This argument makes explicit that the implementation of policies concerning

---

the development of a code of practice for Jisc (Joint Information Systems Committee) in the UK. Sclater's code of practice identifies no fewer than 86 distinct issues comprised in a taxonomy of ethical, legal, and logistical issues for learning analytics aimed "to set out the responsibilities of educational institutions to ensure that learning analytics is carried out responsibly, appropriately and effectively, addressing the key legal, ethical and logistical issues which are likely to arise." Niall Sclater, "Developing a Code of Practice for Learning Analytics," *Journal of Learning Analytics* 3, no. 1 (2016): 16–42, at 31. The Open University's Policy on Ethical Use of Student Data influenced by Slade and Prinsloo offers a set of guiding principles aimed to provide a framework for the ethical application of learning analytics. These principles are: (1) "Learning analytics is an ethical practice that should align with core organizational principles, such as open entry to undergraduate level study. (2) The Open University has a responsibility to all stakeholders to use and extract meaning from student data for the benefit of students where feasible. (3) Students should not be wholly defined by their visible data or our interpretation of that data. (4) The purpose and the boundaries regarding the use of learning analytics should be well defined and visible. (5) The University is transparent regarding data collection, and will provide students with the opportunity to update their own data and consent agreements at regular intervals. (6) Students should be engaged as active agents in the implementation of learning analytics (e.g. informed consent, personalized learning paths, interventions). (7) Modelling and interventions based on analysis of data should be sound and free from bias. (8) Adoption of learning analytics within the Open University requires broad acceptance of the values and benefits (organizational culture) and the development of appropriate skills across the organization." "Policy on Ethical Use of Student Data for Learning Analytics," Open University, retrieved https://help.open.ac.uk/documents/policies/ethical-use-of-student-data/files/22/ethical-use-of-student-data-policy.pdf.

51 Sclater, "Developing a Code of Practice."

52 Slade and Prinsloo, "Learning Analytics."

the ethical use of student data is not enough. These instruments need to be carefully introduced via a communication strategy to explain "how individual subjects are affected by specific applications of learning analytics, to off-set this complexity" (cf. Open University Policy).[53]

Recent work that critically problematizes the adoption of LA in higher education points to the transient, context-sensitive and temporal character of the identity of the student,[54] the meaning of student success in university,[55] the risk of viewing students as sources of data and passive recipients of services and tuition,[56] the risks of putting into practice decision-making structures increasingly beholden to the algorithms without necessarily understanding how they work,[57] the educational triage in open distance education,[58] and the obligation to act.[59] All these themes seem to engage with the issue of information justice,[60] which not only concerns the ideological nature of the data captured and in use but also calls for a critical stance that underscores that "just

53 Open University website, "Policy on Ethical Use of Student Data for Learning Analytics," https://help.open.ac.uk/documents/policies/ethical-use-of-student-data/files/22/ethical-use-of-student-data-policy.pdf.

54 Slade and Prinsloo, "Learning Analytics."

55 Angelo Fynn, "Ethical Considerations in the Practical Application of the Unisa Socio-Critical Model of Student Success," *International Review of Research in Open and Distance Learning* 17, no. 6 (2016): 206–220.

56 Paul Prinsloo and Sharon Slade, "Student Privacy Self-Management: Implications for Learning Analytics," in *LAK '15: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (New York: Association for Computing Machinery, 2015), 83–92.

57 Fynn, "Ethical Considerations."

58 Paul Prinsloo and Sharon Slade, "Educational Triage in Open Distance Learning: Walking a Moral Tightrope," *International Review of Research in Open and Distance Learning* 15, no. 4 (2014): 306–331.

59 Paul Prinsloo and Sharon Slade, "An Elephant in the Learning Analytics Room – The Obligation to Act," in *LAK '17: Proceedings of the Seventh International Learning Analytics and Knowledge Conference* (New York: Association for Computing Machinery, 2017) 46–55.

60 Jeffery Alan Johnson, "Structural Information Justice," *Toward Information Justice: Technology, Politics, and Policy for Data in Higher Education Administration* (Cham: Springer International Publishing, 2018), 133–159.

because you can access, store and analyze student data doesn't mean you should."[61]

Looking Ahead

With learning analytics available in current institutional learning management systems, the era of data-driven educational practices may have just begun. As we have seen in this chapter, diverse stakeholders need to be involved in discussions about the potentials, benefits, and risks of capturing, analyzing, and making decisions based on the digital traces that the students leave behind, and without necessarily knowing what they leave behind. In this context, important questions arise:

- Will the integration of LA help students develop critical and creative thinking and collaborative and communication skills? Or will LA help students to pass university assignments and exams?
- Will emerging data-driven practices in higher education enable us to understand student learning and success better? Or will they encourage the emergence of surveillance practices and invasion of student privacy at universities?
- Will the adoption of LA in educational practices scaffold academics and institutional leadership to embrace education as a public good? Or will it become instrumental in supporting new public management ideals in higher education?

Considering the development of LA, we see it as imperative to build on a socio-critical approach[62] and engage with the multiple questions and multifaceted aspects of using such technologies in everyday higher educational practices. It is crucial to bear in mind that what is at play here is no less than the development of new data-driven, also known as evidence-based practices. They are underpinned and configured by a set of assumptions and potential

[61] Kyle M. L. Jones, "'Just Because You Can Doesn't Mean You Should': Practitioner Perceptions of Learning Analytics Ethics," *Portal* 19, no. 3 (July 1, 2019): 407–428.

[62] Slade and Prinsloo, "Learning Analytics."

biases that need to be detected, unpacked, and seriously discussed by the entire group of stakeholders (e.g., academics, administrative staff, institutional leaders, students, information technology support, developers, companies, and lawyers). In this respect, we call for discussion regarding both the potentials and the challenges entrenched in the use of LA in current educational practices. Moreover, we suggest the following insights to organize such a necessary discussion.

Educational data-driven practices are highly context sensitive

Notably, the approaches to data reflected in the LA literature seem to overlook the central role that the context of the captured data plays in the analysis and interpretation of such data. Techniques such as prediction, clustering, relationship mining, the distillation of data, and discovery with models,[63] just like applications for modeling user knowledge, behavior, experience, and knowledge domains, as well as applications for creating profiles of users, trend analysis, personalization, and adaptation,[64] all imply human judgment and understanding of the particular microcontexts in which the data is imbued.[65] As friction between microcontexts tied to different norms, values, goals are most likely to surface, engaging with them in student data access and analysis is imperative to make explicit and negotiate.[66]

Although there is a potential benefit for institutions, teachers, and students in analyzing data in educational settings, analyses

---

63 Ryan S. J. D. Baker and Kalina Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining* 1, no. 1 (2009): 3–17.

64 Marie Bienkowski et al., "Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief," *US Department of Education, Office of Educational Technology* 1 (2012): 1–57.

65 This is underscored by boyd and Crawford, "Critical Questions for Big Data," 670: "Because large data sets can be modelled, data are often reduced to what can fit into a mathematical model. Yet, taken out of context, data lose meaning and value."

66 See the concept of contextual suppression in Kyle M. L. Jones, "Advising the Whole Student: E-Advising Analytics and the Contextual Suppression of Advisor Values," *Education and Information Technologies* 24, no. 1 (January 2019): 437–458, https://doi.org/10.1007/s10639-018-9781-8.

that do not consider the context of such data remain incomplete, unreliable, and potentially nonsensical.[67] Considering context in large-scale datasets analysis is a challenging task requiring technical competence, judgment, and care if we understand the student identity and performance as temporal dynamic constructs associated with myriads of microcontexts.[68]

Educational data-driven practices are not synonymous with evidence-based practice

In Sweden, as we have seen in the UK and the USA, there is a growing trend toward "evidence-based educational practice."[69] In this sense, it is vital to distinguish between educational data and evidence. There seems to be confusion in the LA literature, which often draws parallels with the medical domain.[70] There is a body of knowledge and professional practice in medicine that develops based on establishing effective interventions. In the field of education, to speak about effective teaching becomes meaningless as "the question that always needs to be asked is, effective for what?"[71] Following Gert Biesta, what counts as effective in the educational sector is tightly related to what is "educationally desirable," which, in turn, makes educational data the object of subjective interpretation and human judgment.

---

67 Cf. boyd and Crawford, "Critical Questions for Big Data."

68 Slade and Prinsloo, "Learning Analytics."

69 This trend is explained by Gert Biesta, who points out that proponents of evidence-based education stress "that education is too important to allow it to be determined by unfounded opinion, whether of politicians, teachers, researchers or anyone else. They call for a culture in which evidence is valued over opinion and argue that any approach to decision making that is not evidence-based is simply pre-scientific." Biesta, "Why 'What Works' Won't Work," 4.

70 Biesta explains this distinction by putting attention to the view of professional practice that is embedded in evidence-based practice: "Central to evidence-based practice is the idea of effective intervention. Evidence-based practice conceives of professional action as intervention, and looks to research for evidence about the effectiveness of interventions. Research needs to find out, in other words, 'what works,' and the main if not the only way of doing this, so it is often argued, is through experimental research, most notably in the form of randomized controlled trials." Biesta, "Why 'What Works' Won't Work," 7.

71 Biesta, "Why 'What Works' Won't Work," 7.

Innovative educational data-driven practices are not sustainable *per se*

From previous experiences, we know that innovations, to be sustainable, need always to be woven into the fabrics of everyday practices of all stakeholders involved.[72] In that respect, we pay special attention to the university teachers whose data literacy, ethical standpoints (shaped by institutional practices), and legal considerations[73] are key for making sense of the student data captured and LA's technical potential to design sound pedagogical interventions. In the literature visited, it is notable that questions about the technological (i.e., data literacy) and ethical competence that, for instance, teachers need to have to harness the potential of LA are rarely even evoked. It seems that it is implicitly assumed that university teachers (and the university staff concerned by the treatment of student data) will become data-literate during the LA deployment process and will engage with it without resisting or discussing the pedagogical value of it.[74] In this context, it is essential to ask how the deployment of LA will contribute to the professional competence development of the university teachers and even whether the emergence of data-driven practices in higher education will liberate teachers or suppress them. And where will the data capturing and the data analysis stop?[75]

---

72 Teresa Cerratto Pargman and Marcelo Milrad, "Beyond Innovation in Mobile Learning: Towards Sustainability in Schools," in *Mobile Learning: The Next Generation*, eds. John Traxler and Agnes Kukulska-Hulme (Abingdon: Routledge, 2016), 154–178.

73 See Magnusson Sjöberg in this volume.

74 These matters are further underscored by boyd and Crawford: "In addition to questions of access, there are questions of skills. Wrangling APIs, scraping, and analyzing big swathes of data is a skill set generally restricted to those with a computational background. When computational skills are positioned as the most valuable, questions emerge over who is advantaged and who is disadvantaged in such a context. This, in its own way, sets up new hierarchies around 'who can read the numbers,' rather than recognizing that computer scientists and social scientists both have valuable perspectives to offer." boyd and Crawford, "Critical Questions for Big Data," 674.

75 Carrie Klein et al., "Technological Barriers and Incentives to Learning Analytics Adoption in Higher Education: Insights from Users," *Journal of Computing in Higher Education* 31, no. 3 (2019): 604–625.

Conclusion

In this chapter, we introduced learning analytics (LA). We discussed some examples of LA applications, and we touched upon the emergence of data-driven practices in higher education. We mapped the complexity of LA seen as a sociotechnical phenomenon in higher education. Lastly, we discussed the following three insights intending to provoke discussion about ethical issues inherent to the emergence of data-driven practices in higher education: (1) educational data-driven practices are highly context sensitive, (2) educational data-driven practices are not synonymous with evidence-based practices, and (3) innovative educational-data-driven practices are not sustainable *per se*.

As a concluding remark, we identify the almost boundless opportunities provided by LA. The capacity to capture, store, analyze, and predict based on large-scale datasets represents a profound change at the level of current educational ethos in current Nordic university practices. But, as practitioners in higher education, we need to be careful what we wish for. We are convinced that the emergence of LA creates a fundamental shift in how we think about learning and teaching in higher education. So we simultaneously acknowledge that data-driven practices need to be carefully configured by the academic freedom and educational values embedded in critical pedagogy.[76]

References

Adejo, Olugbenga, and Thomas Connolly. "Learning Analytics in a Shared-Network Educational Environment: Ethical Issues and Countermeasures." *International Journal of Advanced Computer Science and Applications* 8, no. 4 (April 2017): 22–29.

Avella, John T., Mansureh Kebritchi, Sandra G. Nunn, and Therese Kanai. "Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review." *Journal of Asynchronous Learning Network* 20, no. 2 (2016): 13–29. https://

[76] Henry A. Giroux, *On Critical Pedagogy* (New York; Bloomsbury Publishing, 2011).

www.scopus.com/inward/record.uri?eid=2-s2.0-84975321434&partnerID=40&md5=85c3e4fbfb31f561497048bd7df36fa3.

Baker, Ryan S. J. D., and Kalina Yacef. "The State of Educational Data Mining in 2009: A Review and Future Visions." *Journal of Educational Data Mining* 1, no. 1 (2009): 3–17.

Bamiah, Mervat Adib, Sarfraz Nawaz Brohi, and Babak Bashari Rad. "Big Data Technology in Education: Advantages, Implementations, and Challenges." *Journal of Engineering Science and Technology* 13, special issue on ICCSIT (2018): 229–241. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057115307&partnerID=40&md5=05c80df1bedcbbff3a1ec0ee987496f7.

Bienkowski, Marie, Mingyu Feng, and Barbara Means. "Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief." *US Department of Education, Office of Educational Technology* 1 (2012): 1–57.

Biesta, Gert. "Why 'What Works' Won't Work: Evidence-Based Practice and the Democratic Deficit in Educational Research." *Educational Theory* 57, no. 1 (2007): 1–22.

boyd, danah, and Kate Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15, issue 5 (2012): 662–679.

Brands, Kristine, and Mark Holtzblatt. "Business Analytics: Transforming the Role of Management Accountants." *Management Accounting Quarterly* 16, no. 3 (2015): 1–12

Buckingham Shum, Simon, Rebecca Ferguson, and Roberto Martinez-Maldonado. "Human-Centred Learning Analytics." *Journal of Learning Analytics* 6, no. 2 (2019): 1–9.

Cerratto Pargman, Teresa and Cormac McGrath. "Mapping the Ethics of Learning Analytics in Higher Education: A Systematic Literature Review of Empirical Research." *Journal of Learning Analytics* (2021).

Cerratto Pargman, Teresa, and Marcelo Milrad. "Beyond Innovation in Mobile Learning: Towards Sustainability in Schools." In *Mobile Learning: The Next Generation*, edited by John Traxler and Agnes Kukulska-Hulme, 154–178. Abingdon: Routledge, 2016.

Daniel, Ben. "Big Data and Analytics in Higher Education: Opportunities and Challenges." *British Journal of Educational Technology* 46, no. 5 (2015): 904–920.

Ferguson, Rebecca. "Learning Analytics: Drivers, Developments, and Challenges." *International Journal of Technology Enhanced Learning* 4, no. 5/6 (2012): 304–317.

Fischer, Gerhard. Unpublished presentation at the Symposium on Learning Analytics. Kyoto, 2018.

Fynn, Angelo. "Ethical Considerations in the Practical Application of the Unisa Socio-Critical Model of Student Success." *International Review of Research in Open and Distance Learning* 17, no. 6 (2016): 206–220. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85008204098&partnerID=40&md5=06d1ed057c1cf4766e5fc7247968e160.

Giroux, Henry A. *On Critical Pedagogy*. New York: Bloomsbury Publishing, 2011.

Johnson, Jeffery Alan. "Structural Information Justice." *Toward Information Justice: Technology, Politics, and Policy for Data in Higher Education Administration*, 133–159. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-70894-2_6.

Jones, Kyle M. L. "Advising the Whole Student: E-Advising Analytics and the Contextual Suppression of Advisor Values." *Education and Information Technologies* 24, no. 1 (January 2019): 437–458. https://doi.org/10.1007/s10639-018-9781-8.

Jones, Kyle M. L. "'Just Because You Can Doesn't Mean You Should': Practitioner Perceptions of Learning Analytics Ethics." *Portal* 19, no. 3 (July 1, 2019): 407–428. https://doi.org/10.1353/pla.2019.0025.

Jones, Kyle M. L. "Learning Analytics and Higher Education: A Proposed Model for Establishing Informed Consent Mechanisms to Promote Student Privacy and Autonomy." *International Journal of Educational Technology in Higher Education* 16, no. 24 (2019): 1–22.

Kiron, David, Pamela Kirk Prentice, and Renee Boucher Ferguson. "Raising the Bar with Analytics." *MIT Sloan Management Review* 55, no. 2 (2014): 29.

Klein, Carrie, Jaime Lester, Huzefa Rangwala, and Aditya Johri. "Technological Barriers and Incentives to Learning Analytics Adoption in Higher Education: Insights from Users." *Journal of Computing in Higher Education* 31, no 3, 604–625.

Knight, Simon, Simon Buckingham Shum, and Karen Littleton. "Epistemology, Assessment, Pedagogy: Where Learning Meets Analytics in the Middle Space." *Journal of Learning Analytics* 1, no. 2 (2014): 23–47.

Kruse, Anna, and Rob Pongsajapan. "Student-Centered Learning Analytics." *CNDLS Thought Papers* (2012): 1–9.

Lawson, Celeste, Colin Beer, Dolene Rossi, Teresa Moore, and Julie Fleming. "Identification of 'at Risk' Students Using Learning Analytics: The Ethical Dilemmas of Intervention Strategies in a Higher Education Institution." *Educational Technology Research and Development* 64, no. 5 (2016): 957–968. https://doi.org/10.1007/s11423-016-9459-0.

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburg, and Angela Hung Byers. "Big Data: The Next Frontier for Innovation, Competition, and Productivity." Report from McKinsey Global Institute, May 2011. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation.

McGrath, Cormac, and Anna Åkerfeldt. "Educational Technology (Ed Tech): Unbounded Opportunities or Just Another Brick in the Wall." In *Digital Transformation and Public Services: Societal Impacts in Sweden and Beyond*, edited by Anthony Larsson and Robin Teigland, 143–157. London: Routledge, 2019.

Mor, Yishay, Rebecca Ferguson, and Barbara Wasson. "Learning Design, Teacher Inquiry into Student Learning and Learning Analytics: A Call for Action." *British Journal of Educational Technology* 46, no. 2 (2015): 221–229.

Open University website. "Policy on Ethical Use of Student Data for Learning Analytics." https://help.open.ac.uk/documents/policies/ethical-use-of-student-data/files/22/ethical-use-of-student-data-policy.pdf.

Pardo, Abelardo, and George Siemens. "Ethical and Privacy Principles for Learning Analytics." *British Journal of Educational Technology* 45, no. 3 (2014): 438–450.

Picciano, Anthony G. "The Evolution of Big Data and Learning Analytics in American Higher Education." *Journal of Asynchronous Learning Networks* 16, no. 3 (2012): 9–20.

Prinsloo, Paul, and Sharon Slade. "An Elephant in the Learning Analytics Room – The Obligation to Act." In *LAK '17: Proceedings of the Seventh International Learning Analytics and Knowledge Conference*, 46–55. New York: Association for Computing Machinery, 2017. https://doi.org/10.1145/3027385.3027406.

Prinsloo, Paul, and Sharon Slade. "Big Data, Higher Education and Learning Analytics: Beyond Justice, Towards an Ethics of Care." In *Big Data and Learning Analytics in Higher Education: Current Theory and Practice*, 109–124. Cham: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-06520-5_8.

Prinsloo, Paul, and Sharon Slade. "Educational Triage in Open Distance Learning: Walking a Moral Tightrope." *International Review of Research in Open and Distance Learning* 15, no. 4 (2014): 306–331. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84907047331&partnerID=40&md5=7e0b21fba416f90c7a0eea2f8b0ef04c.

Prinsloo, Paul, and Sharon Slade. "Student Privacy Self-Management: Implications for Learning Analytics." In *LAK '15: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, 83–92. New York: Association for Computing Machinery, 2015. https://doi.org/10.1145/2723576.2723585.

Roberts, Lynne D., Joel A. Howell, Kristen Seaman, and David C. Gibson. "Student Attitudes toward Learning Analytics in Higher Education: 'The Fitbit Version of the Learning World.'" *Frontiers in Psychology* 7 (December 2016). https://doi.org/10.3389/fpsyg.2016.01959.

Schumacher, Clara, and Dirk Ifenthaler. "Features Students Really Expect from Learning Analytics." *Computers in Human Behavior* 78 (2018): 397–407.

Sclater, Niall. *Code of Practice for Learning Analytics: A Literature Review of the Ethical and Legal Issues*. Jisc, 2014. http://repository.jisc.ac.uk/5661/1/Learning_Analytics_A-Literature_Review.pdf.

Sclater, Niall. “Developing a Code of Practice for Learning Analytics.” *Journal of Learning Analytics* 3, no. 1 (2016): 16–42.

Selwyn, Neil. *Distrusting Educational Technology: Critical Questions for Changing Times*. New York: Routledge, 2013.

Siemens, George, and Phil Long. “Penetrating the Fog: Analytics in Learning and Education.” *Educause Review* 46, no. 5 (2011): 31–40.

Slade, Sharon, and Paul Prinsloo. “Learning Analytics: Ethical Issues and Dilemmas.” *American Behavioral Scientist* 57, no. 10 (2013): 1510–1529. https://doi.org/10.1177/0002764213479366.

Society for Learning Analytics website. https://www.solaresearch.org/about.

Sun, Kaiwen, Abraham H. Mhaidli, Sonakshi Watel, Christopher A. Brooks, and Florian Schaub. “It’s My Data! Tensions Among Stakeholders of a Learning Analytics Dashboard.” In *CHI ’19: Proceedings of the 2019 Conference on Human Factors in Computing Systems*, 1–14. New York: Association for Computing Machinery, 2019. https://doi.org/10.1145/3290605.3300824.

Tsai, Yi Shan, Oleksandra Poquet, Dragan Gašević, Shane Dawson, and Abelardo Pardo. “Complexity Leadership in Learning Analytics: Drivers, Challenges and Opportunities.” *British Journal of Educational Technology* 50, no. 6 (November 1, 2019): 2839–2854. https://doi.org/10.1111/bjet.12846.

Viberg, Olga, Mathias Hatakka, Olof Bälter, and Anna Mavroudi. “The Current Landscape of Learning Analytics in Higher Education.” *Computers in Human Behavior* 89 (December 2018): 98–110. https://doi.org/10.1016/j.chb.2018.07.027.

West, Deborah, Henk Huijser, and David Heath. “Putting an Ethical Lens on Learning Analytics.” *Educational Technology Research and Development* 64, no. 5 (2016): 903–922. https://doi.org/10.1007/s11423-016-9464-3.

Williamson, Ben. *Big Data in Education: The Digital Future of Learning, Policy and Practice*. Los Angeles, CA: Sage, 2017.

PART THREE:
NEW OBJECTS—NEW APPROACHES

How to Approach Hard Drives as Cultural Heritage

Johan Jarlbrink

Manuscripts, letters, and diaries are well established as documents in traditional archives. In collections of "papers" we can follow the lives and works of authors, scientists, politicians, and many lesser-known subjects. The documents have been donated by those who produced them, sometimes by relatives or the organizations they worked for. Many collections can be accessed by the public, or at least by researchers. Others are restricted in some way or another, often released after a preset date. Historians would have little to say about everyday life in the past if it were not for the files collected by archives.

We do not write very many letters on paper anymore. Susan Sontag's correspondence is a revealing example. She wrote letters and kept journals for most of her life, until she passed away in 2004. The materiality of her personal archive changed over time, however. In the mid-1990s she bought a PowerBook 5300, and later on a PowerMac G4 and an iBook. The library at UCLA did not know about her hard drives when they first acquired her personal papers in 2005, but in 2012 they got access to the three hard drives, with 18,000 emails, drafts, photographs, and other documents. The data is now part of Sontag's personal papers. Access is restricted, but not prohibited. Benjamin Moser, who published a biography about Susan Sontag in 2019, was able to go through some of the files as part of his research:

How to cite this book chapter:
Jarlbrink, Johan. "How to Approach Hard Drives as Cultural Heritage." In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 229–251. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.j.

> [R]eading papers and manuscripts is one thing. Looking through someone's e-mail is quite another, and the feeling of creepiness and voyeurism that overcame me as I sat with [the librarian] Gonzalez [and] struggled with the unstoppable curiosity that I feel about Sontag's life. To read someone's e-mail is to see her thinking and talking in real time. […] One sees Sontag, who had so many friends, elated to be in such easy touch with them ("I'm catching the e-mail fever!"); one sees the insatiably lonely writer reaching out to people she hardly knew and inviting them to pay a call.[1]

Personal papers are even more personal when they are digital, it seems. And Moser did not even look at all the other kinds of data stored on hard drives—the web browser history, the words she googled, stored geolocations, metadata. Such information would be a gold mine for intellectual historians, but valuable also for media scholars researching everyday media life in the digital age. Scholars in other disciplines might want to investigate the software someone has used, or the file formats, viruses, the devices once connected to the computer—the research potential is enormous. These kinds of data are highly sensitive, of course, but so are many of the paper documents kept by archives. Why, then, is it so easy to access and read a private diary from 1897 but so difficult to find a web browser history from 1997?

The use of computers has been widespread since the 1980s, but hard drives as archives are still new to most memory institutions. Tom Hyry, the former director of the special collections at UCLA, explains that the incorporation of Sontag's hard drives into the library "raised technical, ethical, philosophical, financial, and practical issues that still seem new to the archival endeavor."[2] The preservation, curation, and presentation of someone's hard drive require technical solutions, skills, guidelines, and routines. One reason why the process has been slow for many institutions is perhaps that few scholars within the humanities and social sciences

[1] Benjamin Moser, "In the Sontag Archive," *The New Yorker*, January 30, 2014, https://www.newyorker.com/books/page-turner/in-the-sontag-archives.

[2] Tom Hyry, "On Digital Archives: Lessons from the Susan Sontag Hard Drives." Paper presented at the Society of American Archivists meeting, Cleveland, OH, August 20, 2015, http://nrs.harvard.edu/urn-3:HUL.InstRepos:40918991.

have taken an interest in these kinds of born-digital archives. This is unfortunate since hard drives are important records of everyday life, with a great potential as empirical sources.

The aim of this chapter is to exemplify what an investigation of a hard drive implicates, the methods needed to conduct it, and what kind of results we can get out of it. To focus the investigation, I will approach hard drives as records of everyday media use. I am less interested in the content of private emails and documents, what the photographs show or what secrets a web browser might reveal. I am interested in more general patterns of media use and how they change over time. To develop a media archaeological approach to the digital traces of everyday life (see the introduction to this volume), I will suggest a computer forensic method used as a media ethnographic tool. Computer forensics is a method developed to examine digital traces in order to establish a user's activities. Media ethnography is a field of inquiry researching media production and audiences in natural settings based on interviews, observations, media diaries, and field notes. The two methodological traditions are different but overlap. Computer forensics and media ethnography both take an interest in people's routines, the way they organize things, what they do, and how they do it. The approach I will demonstrate is a forensic investigation guided by media ethnographic themes, findings, and questions. Drawing from ethnographical studies of computers in everyday life, I will suggest the broad categories of time and space as fruitful starting points in digital excavations.

Hard Drives as Cultural Heritage

Media scholar Pelle Snickars wrote in 2010 that the hard drive was "our most central tool" and "the material base of our digital memory culture."[3] Yet, it is almost invisible in our daily lives and often buried behind several layers of plastic and circuit boards. Few scholars within the humanities or social science have paid much attention to it. Matthew Kirschenbaum is one exception,

[3] Pelle Snickars, "Hårddisken och samtiden [The Hard Drive Today]," in *The Story of Storage I*, ed. Lars Björk, Jānis Krēsliņš, and Matts Lindström (Stockholm: Kungliga biblioteket, 2010), 44 (my translation).

excavating the different layers of data from the perspective of digital literature. More on his pioneering work in the next section.

Snickars included servers in his discussion on hard drives in 2010. Ten years later it is perhaps these storage units, rather than personal hard drives, that form the material base of our memory culture. Documents, photographs, and messages are uploaded, sent, and accessed on online platforms. Few people save their Facebook posts on their own computer or phone. Most of us consume music and movies through streaming, and many documents are written and stored directly on platforms such as Google Drive. The biographies of future Susan Sontags cannot rely on personal hard drives alone. These outsourced storage units, however, are often beyond the reach of archives and libraries.

Personal computers as the prime storage units for digital information might be disappearing, but they are still important to memory institutions collecting the digital traces of the 1990s and 2000s. These were the decades when computers were domesticated, when they (at least in the West) became part of many people's everyday lives. The British Library and the US Council on Library and Information Resources raised the questions on how to collect and preserve personal hard drives in the late 2000s. In their reports they presented guidelines to support institutions and archivist in their work. Digital forensics was suggested as the prime method to capture and transfer data.[4] The decade following the two reports has seen many initiatives in the field. Museums, archives, and libraries have slowly started to incorporate data from hard drives into their collections. Computer forensics is nowadays "a standard practice in memory institutions for the preservation of digital storage media and born-digital records."[5] Yet, very few scholars within the humanities and social science have actually

---

4 Jeremy Leighton John, Ian Rowlands, Peter Williams, and Katrina Dean, *Digital Lives: Personal Digital Archives for the 21st Century. An Initial Synthesis* (London: British Library 2010); Matthew G. Kirschenbaum, Richard Ovenden, and Gabriela Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* (Washington, DC: Council on Library and Information Resources, 2010).

5 Thorsten Ries and Gábor Palkó, "Born-Digital Archives," *International Journal of Digital Humanities* 1, no. 1 (April 2019): 4, https://doi.org/10.1007/s42803-019-00011-x.

used and analyzed data from hard drives as primary sources in their research. Ries and Palkó (2019) write that there is a gap between memory institutions and archival science researchers on the one hand and scholars from the humanities and the social sciences on the other. There is a need to:

> enable GLAM institutions, institutional networks and infrastructures to develop their born-digital collections in meaningful ways, improve preservation formats, curation workflows, repositories, services, and access for researchers. This can only be achieved by cross-sectoral and interdisciplinary collaboration to support active research on born-digital collections.[6]

Hard drives from public figures such as Susan Sontag and Salman Rushdie are often key examples when preservation and access is discussed (for another fascinating case, see Amanda Wasielewski in this volume, about the excavation of the musical *RENT* on floppy disks left by the creator Jonathan Larson). If the strategies for collecting and preserving are based on data from intellectuals and writers, there is actually a risk that a textual bias will be built into the infrastructure. The archival strategies implemented today will have an impact on future research possibilities.[7] What I want to highlight here is that there are other—and overlooked—kinds of data to collect and explore.

Computer Forensics as a Method

Kirschenbaum has argued for a media specific reading of electronic literature. The digital format is an important part of what constitutes digital text. It is produced differently compared to the typewritten or printed text, it behaves differently, and it can be interpreted in different ways. Digital text is more than text on a screen.[8] My concern here is not digital poetry, but a basic understanding of digital storage is necessary also for an investigation of the hard

6 Ries and Palkó, "Born-Digital Archives," 4.

7 Agiatis Benardou et al., eds., "Introduction: A Critique of Digital Practices and Research Infrastructures," in *Cultural Heritage Infrastructures in Digital Humanities* (Abingdon: Routledge, 2018), 4.

8 Matthew G. Kirschenbaum, *Mechanisms: New Media and Forensic Imagination* (Cambridge, London: MIT Press, 2008), xii–xiv.

drive as an archive. Data in digital form appears both fragile and stable. A sudden crash or a document not properly saved means that the data is lost for many users. A forensic investigation, however, will most likely be able to recreate such data.[9] To understand why, we need to know how data is stored on hard drives and how hard drives are managed within forensic investigations.

One way to describe the different appearances of digital objects is to differentiate between three basic forms or layers of the objects: the physical objects (inscriptions on a medium), logical objects (inscriptions read by software), and conceptual objects (as they appear on the screen).[10] Scholars within the humanities and social sciences have mostly dealt with digital objects in the third sense. Computer forensics, however, investigates all three forms. A basic definition states that it is concerned with "the examination of digital storage and digital environments in order to determine what has happened."[11] Tools and methods are developed in order to recreate (or monitor) events and actions based on digital traces. Some traces do not appear on the screen but can be identified as physical or logical objects. That is why all three layers or forms are important.

Most users know that the commands "delete file" or "empty trash" will remove a file's entry or address from the catalog but not delete the file itself from the hard drive. The space it takes up on the disk is flagged as available, but the data is only erased once it is overwritten with new data. With ever-increasing storage capabilities of disks, a space flagged as available might not be overwritten immediately. If the deleted file is overwritten, there are often copies in the form of temporary files or older versions saved elsewhere. Such files are created when files are modified, printed, copied, sent as attached files, *et cetera*.[12]

Since digital storage media are divided into clusters of a fixed length (often 4,096 bytes per allocated unit), individual files larger than a single cluster are stored in multiple places on the disk.

9 Kirschenbaum, *Mechanisms*, chap. 2.
10 Kirschenbaum, *Mechanisms*, 3.
11 Joakim Kävrestad, *Fundamentals of Digital Forensics: Theory, Methods and Real-Life Applications* (Cham: Springer, 2018), 3.
12 Kirschenbaum, *Mechanisms*, 52.

When a file is deleted, a part of it might be overwritten while other parts remain intact. These fragments can prove that a file existed even though most of the data is erased.[13]

In order to capture every fragment and every bit unaltered, the standard method within computer forensics is to copy the storage medium in a so-called bitstream. Similar to traditional archival practices, developed to keep documents and files in the same order as they were once arranged by those who produced them, a bitstream transfers every bit recorded in a linear sequence. To simply copy the files from one disk to another would miss the fragments and deleted files that do not show up among the indexed files in the graphical interface, and would add new metadata to the files. The computer or device should not even be turned on, as new data will be added as soon as the operating system starts running. Instead, the hard drive is taken out of the computer and is connected to a docking station. A bitstream captures every bit recorded on the disk and keeps the metadata intact. Since digital forensics was originally developed to support criminal investigations, it treats digital data like fingerprints and DNA on a crime scene, as evidence that should never be altered. The bitstream method makes the copy a stand-in for the original.[14]

Software for computer forensics usually lists the folders and files as they were arranged by the user, along with deleted files not yet overwritten, carved files (file fragments reassembled based on signatures in the code), and fragments in unallocated sectors of the disk. A USB stick might reveal a few hundred files, while a recent hard drive may contain tens of thousands of them. Most tools extract and highlight data of special interest to facilitate overview and orientation. Autopsy (4.11.0), the tool I am most familiar with, generates lists of image files, video and documents, EXIF metadata (the cameras used to take photos, among other things), encrypted files, accounts, emails and addresses, and very large files. A search function makes it possible to locate files and fragments containing particular keywords. The tool is obviously designed with criminal investigators in mind, for those looking

[13] Kirschenbaum, *Mechanisms*, 52.
[14] Kirschenbaum, *Mechanisms*, 53.

for specific files, contacts, or addresses. An example from *Digital Evidence and Computer Crime* (2011) is typical:

> <A HREF="http://www.google.com/search?hl=en&lr=&ie=ISO-8859-1&q=human+poison+herbs" ADD_DATE="1049641841" LAST_VISIT="1049642467" VISITATION_COUNT="3 "OBJECT_TYPE="LINK">[15]

Scholars interested in general patterns of user behavior and how they change over time have more limited options. The timeline view in Autopsy (Figure 1) gives a useful overview, though, showing events (files created, accessed, modified, deleted, web activity, installed programs, *et cetera*) in a sequence, with the possibility to zoom in (on events at a specific date, minute, second) or out (indicating changes from year to year).

An even better option might be to export the metadata to Excel, where it can be filtered and grouped depending on the research interest. The metadata for an individual file usually includes file name, extension (file format), when it was created, modified, accessed, and deleted (if it was); its size, address, hash value (an id derived from the data, making it possible to locate duplicates and similarities), and its full path. Based on this metadata, researchers can examine users' information management, how files are organized into folders, when different kinds of software were installed, how the generation of different file formats change over time, when files in specific folders were created and modified, and so on.

Files of special interest may also include log files, browser history, playlists, and traces of various devices connected to the computer, such as printers, scanners, USB sticks, and phones. Forensics is based on the idea that "[e]very contact leaves a trace," and this is no less true for computer forensics.[16] This does not mean that every contact is traceable, however. What data and metadata is available depends on the operating system and the file system. Older generations of Apple computers, for example, did not add extensions to the file name or timestamps for "last accessed."

[15] Eoghan Casey, *Digital Evidence and Computer Crime: Forensic Science, Computers and the Internet* (Burlington: Academic Press, 2011), 599.

[16] Kirschenbaum, *Mechanisms*, 49.

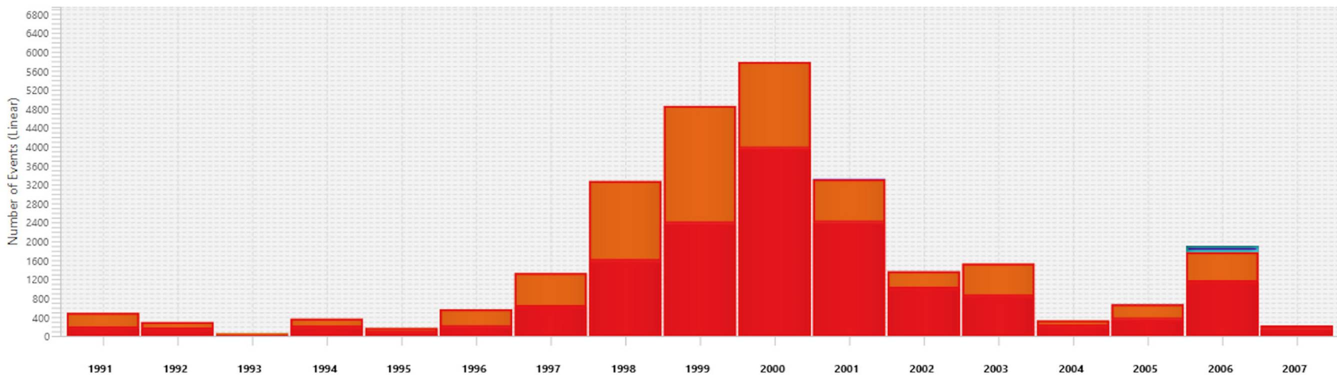


Figure 1: Timeline created with Autopsy 4.11.0, based on an Apple iMac G3 (id TEKS0047646) in the collections of the National Museum of Science and Technology, Stockholm, Sweden. Orange represents files created, red files modified, and blue miscellaneous files, in this case photographs. Copyright: Author. License: CC BY 4.0.

Deleted and carved files may lack most of the metadata.[17] Still, many aspects of a person's digital life can be recreated based on the files, logs, and metadata. Part of it is very personal; other parts may be less sensitive. The rise and fall of MP3 files on personal computers, or the history of preferred search engines, or the cycles of constantly new software updates, or the domestic ecosystems of connected devices—these issues might be examined without sensitive information being required or revealed.

Mapping Time and Space

In order to sketch potential research topics, where forensic methods and findings can complement and stimulate media ethnographical approaches, I will mainly draw on two previous studies of computers and digital communication in everyday life, Elaine Lally's *At Home with Computers* from 2002, and Maria Bakardjieva's *Internet Society: The Internet in Everyday Life* from 2005. They were not written as historical accounts of media use—but now that is what they are. Lally interviewed 95 individuals in 31 households in Australia. Most of the interviews covered experiences from the second half of the 1990s. Bakardjieva interviewed 23 respondents living in Canada about media use in the late 1990s and early 2000s. Lally took an interest in people's computer interaction in a broad sense; Bakardjieva was primarily focused on respondents' experiences of the internet, but she also described domestic computer usage in general.

Most of the respondents interviewed by Lally and Bakardjieva lived in families with only one computer. This computer was often their first. For many of them, the computer represented a "shared space," a technology available for every family member, but (mostly) for one person at a time. A shared computer called for more or less developed rules of "time-zoning."[18] The computer was used by certain family members for certain things at certain times, depending on internal hierarchies of power, values, skills, and needs. Thus, much of what Lally and Bakardjieva described

17 Casey, *Digital Evidence*, 588–591.

18 Maria Bakardjieva, *Internet Society: The Internet in Everyday Life* (London: Sage, 2005), 151.

was how the use of computers was organized and restricted in time and space.[19]

A hard drive itself is also a shared space, filled with data in temporal layers. A forensic analysis reveals how this space is divided and used, when, for what, and by whom. Lally and Bakardjieva were less interested in the content of people's emails, exactly what the respondents searched for online, or what kind of work or entertainment they were engaged in. What they examined was the way computer interaction was organized socially. My intention here is to build on their work in order to map potential overlaps between forensic inquiries and ethnographic approaches. What kind of questions derived from the ethnographic analysis can we investigate further by forensic means? What kind of forensic findings might be contextualized in ethnographical studies?

Time

Computers that end up in museums or archives have unique histories as objects in specific social settings. One way to approach a computer as a historical artifact is to examine the "cultural biography" of the object and its journey in time and space. When anthropologist Igor Kopytoff launched the idea of cultural biographies of things in 1986, he stated that we can ask similar questions about things as we do about people:

> Where does the thing come from and who made it? What has been its career so far, and what do people consider to be an ideal career for such things? What are the recognized "ages" or periods in the thing's "life," and what are the cultural markers for them? How does the thing's use change with its age, and what happens to it when it reaches the end of its usefulness?[20]

What is specific about a computer is that it keeps track of its own history. We can ask the former users about its history, and we

[19] Both of them have chapters on "Temporal Rhythms of the Computerized Home" (Lally) and "Making Room for the Internet" (Bakardjieva).

[20] Igor Kopytoff, "The Cultural Biography of Things: Commoditization as Process," in *The Social Life of Things: Commodities in Cultural Perspective*, ed. Arjun Appadurai (Cambridge: Cambridge University Press, 1986), 66–67.

can trace its history from within the object itself. In a biographical investigation of a computer it would be fruitful to distinguish different "ages," what characterized these, and how and why a computer transitioned from one age to the next. Jonathan Stern has pointed out that the aging of computers is in most cases a process driven by repeated introductions of new software. A computer bought a few years ago might be running perfectly fine with its original software, but it is not powerful enough to run new software. Thus, the computer *becomes* old when new software is introduced.[21] The history of installed and updated software can easily be tracked in log files and in the metadata of program files. Such metadata can also show the old within the new. The timeline in Figure 1 is based on a hard drive manufactured in 1998 and bought in 1999—and reveals that many of the files in the preinstalled software packages were created in the early 1990s. The users might think that what they buy is brand new, but much of it is old stuff packaged in a new box.

Several users interviewed by Lally and Bakardjieva were well aware that new media ages fast: "I bought a 486—at that time it was the best and today it is already old and out of date."[22] "[I]t seemed to be out-of-date as soon as we got it. It probably wasn't really quickly, but it just seemed really quick."[23] In another case, it was a new computer at work that made the home computer from the previous year look ancient.[24] Some of the old, slow, and discarded computers were given a second chance, however. Computers no longer used by adults were sometimes given to children, and some men handed over their old machines to their wives. One of the most common ways to acquire a home computer was to purchase (or just take over) an old computer from work, colleagues, and friends.[25] Thus, early adopters spread the technology to laggards in their social networks: "Non-professionals and

---

[21] Jonathan Sterne, "Out with the Trash: On the Future of New Media," in *Residual Media,* ed. Charles R. Acland (Minneapolis, MN: University of Minnesota Press, 2007), 22–23.

[22] Bakardjieva, *Internet Society*, 93.

[23] Elaine Lally, *At Home with Computers* (Oxford: Berg, 2002), 85.

[24] Lally, *At Home*, 117.

[25] Lally, *At Home*, 74–88.

'poor cousins' take up the computer waste and put it to uses of their own."[26]

A standard life journey of a computer takes the following route in Sterne's account: "it travels through categories from new, to useful, to obsolete, to unused, to trash."[27] As exemplified in the interviews, however, there might be several detours on the journey. A forensic analysis can trace the journey and reveal several temporal layers of data, from multiple users or owners. Such an analysis can show how usage changed over time. A new computer is most likely used differently from one that is four years old, and different owners use it for different things. The timeline in Figure 1 indicates some of the changes on a macro level, but a thorough analysis of the underlying data would provide us with a detailed biography as well as different user profiles. Much of the social context would be missing from the analysis, but it could be a useful addition to traditional ethnographical approaches.

New media studies are often occupied with the latest version, new gadgets, and new practice. A computer forensics approach may provide an alternative and reveal hardware and software in use long after new versions were launched. Historian of technology David Edgerton made the claim in *The Shock of the Old* (2006) that "many things we think of as old remained in practical use for longer than our future-oriented accounts of technological history allow."[28] What is true for spaceships and sewing machines is most likely true for computers: many of the technologies in use are surprisingly old. Some users transfer old software to the new computer when they upgrade. Others install emulators to be able to play old games on new machines. A forensic analysis of metadata makes it possible to distinguish the temporal layers.

We can also investigate a different kind of temporality, the daily rhythms of computer interaction. Such an interest was an important part of computer forensics right from the beginning. When astronomer Clifford Stoll in Berkeley tried to track a hacker in 1986, breaking into various military networks all over the USA

[26] Bakardjieva, *Internet Society*, 94.
[27] Sterne, "Out with the Trash," 23.
[28] David Edgerton, *The Shock of the Old: Technology and Global History Since 1900* (New York: Oxford University Press, 2006), 29.

via the computer in Stoll's own lab, he noticed that the hacker entered the systems at specific hours: "On the average, the hacker showed up at noon, Pacific time. Because of daylight savings time, I could stretch this to 12:30 or even 1 P.M., but there was no way that he was an evening person."[29] Most hackers worked evenings and late nights due to the lower costs of data traffic, but not this one. Could this indicate that the hacker came from overseas? Early afternoon in California was late night in Europe. Stoll broadened his search and managed to track him down—in Hannover, Germany. Stoll's pioneering work laid the ground for what is today computer forensics. His approach to time might be valuable also for scholars analyzing the mundane rhythms of digital life.

How the families studied by Lally and Bakardjieva used their computers varied with the seasons, with school semesters and breaks, weekdays, and weekends. How to divide, regulate, and spend computer time was one of the most frequently reoccurring issues. According to the domestic moral economy of time, some activities were encouraged, while others were restricted:

> this hierarchy may be institutionalized in explicit rules ("homework comes before games") and conventions ("the person whose homework deadline comes first has first turn"), but is also open to negotiation (adults can stay up later than children so the homework can sometimes come before adult income-generating work).[30]

Members of families with only one computer allocated time in the same way they allocated space. They had to define rules, and negotiate and divide time slots depending on authority and need. Many of the adults struggled to separate work and leisure time. Some spent evenings on work-related tasks in front of their computers, but computers from work were also used for activities not related to work.[31] How much computers were used, by whom and when, also changed over time, depending on the age of family members and the age of the computer itself. Initially, many families

[29] Clifford Stoll, *The Cuckoo's Egg: Tracking a Spy Through the Maze of Computer Espionage* (New York: Doubleday, 1989), 138.
[30] Lally, *At Home*, 131.
[31] Bakardjieva, *Internet Society*, 96.

used them in the same way as vacuum cleaners—turned on only when used. Over time they came to resemble fridges instead—never turned off.[32] Aged computers were sometimes used less and less, however, as they became slow and not compatible with new software.

How and when computers have been used has changed with the internet connections available. The tariffs were higher for daytime surfing well into the 2000s, explaining why some people chose to download heavy files, games, and large software during early mornings or late at night. When internet connections used the same landline as the regular phone, the time online had to be limited in order not to block phone calls.[33] New connections and flat rates have meant that the time spent online has steadily increased. Jonathan Crary has pointed out that modern media have rearranged our daily rhythm, how we spend our time, when we sleep, and how much (a related idea concerns the "acceleration of 'the pace of life'," see Jonas Stier in this volume). Computers in standby mode and always connected have turned weekends into workdays, while an endless supply of content online makes us stay up late at night.

> The notion of an apparatus in a state of low-power readiness remakes the larger sense of sleep into simply a deferred or diminished condition of operationality and access. It supersedes an off/on logic, so that nothing is ever fundamentally "off" and there is never an actual state of rest.[34]

A forensic analysis of hard drives could ground this analysis in actual user behavior, or reveal patterns that are more complex. What does a daily rhythm of computer use look like? How does it differ between individuals in different contexts? Does it change over time? How?

Figure 2 shows the number of files created at every hour in a personal folder. It is stripped of most of the details, but indicates a personal rhythm nevertheless. One possibility is to filter the data

32 Lally, *At Home*, 127.
33 Bakardjieva, *Internet Society*, 148, 151.
34 Jonathan Crary, *24/7: Late Capitalism and the End of Sleep* (London, New York: Verso, 2013), 13.

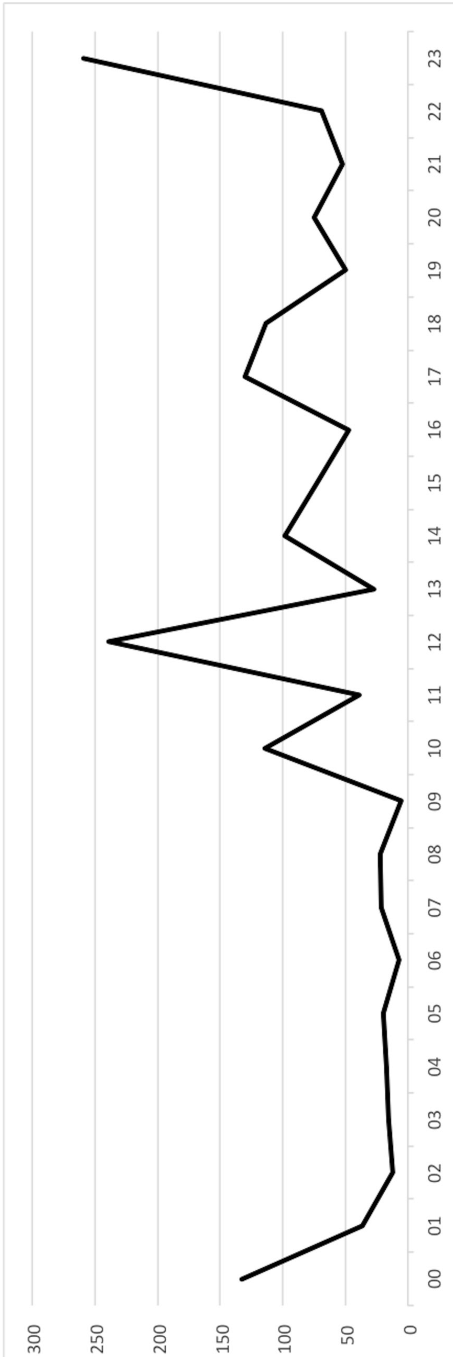


Figure 2: Timeline created in Excel, based on file metadata from a personal folder on an Apple iMac G3 (id TEKS0047646) in the collections of the National Museum of Science and Technology, Stockholm, Sweden. Copyright: Author. License: CC BY 4.0.

and examine timestamps for specific file formats. Do MP3 files follow a different rhythm than Word files and PowerPoints? We could also compare different personal folders within a family. If a folder is divided into "Private" and "Work," we can compare the timestamps in order to understand divisions and overlaps between work and leisure time. Data from the forensic examination can be explored further with ethnographical methods, and vice versa. What did you do around lunchtime? What kept you up late at night? What happened between 7 and 10 pm? Figure 2 gives a basic insight into what is possible, but there is a lot more that can be done.

Space

The life journey of a computer, as sketched by Jonathan Sterne (from new to trash), usually corresponds to movements in space: from work to home, from the living room or home office into the children's room, from bedroom to basement, and from visible areas into closets.[35] The geography of a family home is often based on notions of shared space and private, adult areas and underage, female and male.[36] Where a computer is placed has consequences for when it can be used, by whom, and for what. Parents might choose a specific location for the computer in order to control how children are using it. Some rooms are "neutral" and used by everyone, like living rooms and hallways, while others, such as bedrooms, are private. One of Bakardjieva's respondents explained that "the living room was where everybody could have access to it. Because we just have the one [computer], if we had more than one, it may have made sense to have it in a study, or a bedroom, but just the one, we put it, you know, central."[37]

In some families an individual member claims "ownership" of the computer and restricts other members' access by placing it in a private room. When such formal or informal ownership changes, the computer is often moved to another room.[38] How the computer is used may change with its physical location: private or noisy

[35] Sterne, "Out with the Trash," 25.
[36] Bakardjieva, *Internet Society*, 139.
[37] Bakardjieva, *Internet Society*, 151.
[38] Lally, *At Home*, 138.

stuff (games and music), or work that requires silence and concentration, in closed-off areas, and less private tasks in a shared space.[39]

The hard drive itself is a space that also needs to be shared and divided. In families with only one computer it is often divided into a common desktop and personal folders. To put personal files directly on the desktop is to colonize common space. One of Lally's respondents complained that her husband had "created a folder called 'Masters' for his university work which is outside all the individual family members' own folders." She saw this as "proprietorial": "You should have your 'Masters' inside 'Thomas'."[40] How the memory space available is used by the different family members can tell us something about formal or informal ownership. Take the personal folders of this 4 GB hard drive as an example (Table 1):

Who "owned" this computer? From file size alone, we can at least say that person 2 was the one who was taking up most storage space with his or her personal stuff. The files in the other persons' folders do not come close when it comes to size. If we add time to the equation, we would see that person 1 was actually creating most of the files in the personal folder in year one—but next to nothing the following years. Person 2 created most of the files in years two and three. Person 3 was creating files from year one until year seven—but small Word files that do not take up much space. "Ownership" and use usually change over time, and a forensic analysis makes it possible to follow changes in detail.

Many of the files on a hard drive are not personal at all. They come with the computer when it is bought, or are stored on the

Table 1: The size of personal folders, based on metadata from an Apple iMac G3 (id TEKS0047646) in the collections of the National Museum of Science and Technology, Stockholm, Sweden.

Person 1	169.2 MB
Person 2	1012.9 MB
Person 3	15.2 MB

39 Bakardjieva, *Internet Society*, chap. 6.
40 Lally, *At Home*, 138.

hard drive when new software is installed or old ones updated. The hard drive is a space that users share with software manufacturers. This part of the space is also worth an excavation.

The field of software studies (in Lev Manovich's version) is concerned with "the role of software in contemporary culture, and the cultural and social forces that are shaping the development of software itself."[41] Matthew Fuller writes about methods within the field that "software studies approaches might characteristically tend to identify specific algorithms, articulate their genealogy, recognize and work with their characteristics, and see them as part of a larger assemblage."[42] Forensic investigations of personal hard drives might provide software studies with individual stacks of software, what they were once used for, how they were used in combination, what was installed but immediately uninstalled, and so on. Another possibility is to analyze the way space taken up by software was once customized by individual users. To design and download "skins" for the MP3 player was very common around 2000, as were various icons used for ICQ accounts, *et cetera*.[43] Plug-ins and personal settings are other examples.

To map time and space in the ways I have suggested here is just meant to exemplify possible research. Other aspects to explore include what Lally describes as "the domestic ecology of objects," the way computers become connected to printers, floppy disks, CDs, USB sticks, modems, phones, and so on.[44] From here we can follow how domestic networks are connected to the world outside.

Concluding Remarks: Institutional and Methodological Challenges

Before we can follow any of the traces, they must be available for research. Unless researchers get their hands on hard drives

---

41 Lev Manovich, *Software Takes Command* (New York, London: Bloomsbury, 2013), 10.

42 Matthew Fuller, "Software Studies Methods," in *The Routledge Companion to Media Studies and Digital Humanities*, ed. Jentery Sayers (New York: Routledge, 2018), 251.

43 Jeremy Wade Morris, *Selling Digital Music, Formatting Culture* (Oakland, CA: University of California Press, 2015), 55.

44 Lally, *At Home*, chap. 9.

themselves, archives, libraries, and museums are logical repositories for this kind of material. These institutions have a long tradition of collecting personal papers and objects, and know the importance of protecting sensitive information. Analog and digital collections have much in common, but there are important differences. A home computer with an internet connection is likely to record data about everyday life in all its diversity, whether the user is aware of it or not. Data from and about other people will be recorded as well, without them knowing it. Personal papers are often more selective, at least those that reach archival institutions. Papers once shredded by a donor are gone forever, while many files deleted on computers can be recovered. Archives, libraries, and museums need to develop strategies, routines, and technical skills in order to handle born-digital collections in ways that are legal, are archivally sound, and protect private integrity. Donors must be informed about the possibility that hidden data might be recovered.[45]

Forensic procedures follow archival standards prescribing authenticity and provenance. Yet, the completeness that is a requirement in a criminal investigation can cause practical problems for archival institutions: a disk image contains more data than some of them can handle. Sensitive information, about the original user(s) as well as others, can be hidden among thousands of files. Some of it can be located automatically with forensic software, such as addresses, bank accounts, and social security numbers. To locate other kinds of data requires manual processing—and much time. In an ideal situation this is done in close contact with the donor, but this is not always possible. Some institutions choose to keep much of the unsorted and potentially sensitive data in a secure and restricted "dark archive," while providing access to a selection of nonproblematic files only—this is the way Salman Rushdie's hard drives are curated and made available.[46]

[45] Kirschenbaum, Ovenden, and Redwine, *Digital Forensics*, 29–39; Alyssa Hamer, "Ethics of Archival Practice: New Considerations in the Digital Age," *Archivaria* 85 (Spring 2018): 156–179.

[46] Ben Goldman and Timothy D. Pyatt, "Security Without Obscurity: Managing Personally Identifiable Information in Born-Digital Archives," *Library and Archival Security* 26, no. 1–2 (2013): 44–45, 50, https://doi.org/10.1080/01960075.2014.913966.

Others avoid the workload generated by forensic methods, and let donors themselves transfer the files they want to include in the digital repository via online platforms. This, however, may exclude and distort metadata and the context that makes it possible to determine how and when files were created, where they were created and stored, and so on.[47]

It makes sense for a library to primarily collect selected text files from well-known authors. Other institutions, such as museums of ethnography or technology, should consider capturing and preserving a broader range of data. As I have shown in this chapter, discarded hard drives have much to tell us about the way we live our lives, about daily routines and rhythms, and about how technology became part of everyday life. Important clues to explore can be found in all kinds of files, but also in the metadata. This category of data is often (but not always) less sensitive than the actual content of files. The file metadata from a single hard drive is not big data; it can easily be processed manually if parts of it need to be redacted.

If a disk image has already been prepared by archivists when the hard drive is donated, researchers do not need to repeat the process. Nevertheless, they need to think like a forensic investigator and need to be aware of the principles of digital storage and how different file systems generate different kinds of metadata. A disk image captures every trace on the storage medium, but not all traces ever left. There is no way of knowing what is overwritten and missing from the record. Still, researchers interested in issues related to routines, time and information management, the domestication of technology, and the history of new media have much to gain from a forensic investigation of a hard drive. An interview relies on an interviewee's self-understanding and what he or she can remember. A disc image represents a different kind of memory, a record of the microscopic details of everyday media use. Ethnography and computer forensics have different roots, but they can work in concert.

47 Katinka Ahlbom (head of the department for manuscripts, maps, and images, National Library of Sweden), personal communication November 21, 2019.

Acknowledgments

The research for this chapter was conducted within the project *Digital Models: Techno-Historical Collections, Digital Humanities & Narratives of Industrialisation*, funded by the Royal Swedish Academy of Letters, History and Antiquities. My gratitude also goes to Jim Robertsson at Humlab, Umeå University, for guiding me through the technicalities of the forensic investigation.

References

Bakardjieva, Maria. *Internet Society: The Internet in Everyday Life*. London: Sage, 2005.

Benardou, Agiatis, Erik Champion, Costis Dallas, and Lorna Hughes, eds. “Introduction: A Critique of Digital Practices and Research Infrastructures.” In *Cultural Heritage Infrastructures in Digital Humanities*, 1–14. Abingdon: Routledge, 2018.

Casey, Eoghan. *Digital Evidence and Computer Crime: Forensic Science, Computers and the Internet*. Burlington: Academic Press, 2011.

Crary, Jonathan. *24/7: Late Capitalism and the End of Sleep*. London, New York: Verso, 2013.

Edgerton, David. *The Shock of the Old: Technology and Global History Since 1900*. New York: Oxford University Press, 2006.

Fuller, Matthew. “Software Studies Methods.” In *The Routledge Companion to Media Studies and Digital Humanities*, edited by Jentery Sayers, 250–257. New York: Routledge, 2018.

Goldman, Ben, and Timothy D. Pyatt. “Security Without Obscurity: Managing Personally Identifiable Information in Born-Digital Archives.” *Library and Archival Security* 26, no. 1–2 (2013): 37–55. https://doi.org/10.1080/01960075.2014.913966.

Hamer, Alyssa. “Ethics of Archival Practice: New Considerations in the Digital Age.” *Archivaria* 85 (Spring 2018): 156–179.

Hyry, Tom. “On Digital Archives: Lessons from the Susan Sontag Hard Drives.” Paper presented at the Society of American Archivists meeting, Cleveland, OH, August 20, 2015. http://nrs.harvard.edu/urn-3:HUL.InstRepos:40918991.

Kirschenbaum, Matthew G. *Mechanisms: New Media and Forensic Imagination*. Cambridge, MA, London: MIT Press, 2008.

Kirschenbaum, Matthew G., Richard Ovenden, and Gabriela Redwine. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Washington, DC: Council on Library and Information Resources, 2010.

Kopytoff, Igor. "The Cultural Biography of Things: Commoditization as Process." In *The Social Life of Things: Commodities in Cultural Perspective*, edited by Arjun Appadurai, 64–91. Cambridge: Cambridge University Press, 1986.

Kävrestad, Joakim. *Fundamentals of Digital Forensics: Theory, Methods and Real-Life Applications*. Cham: Springer, 2018.

Lally, Elaine. *At Home with Computers*. Oxford: Berg, 2002.

Leighton John, Jeremy, Ian Rowlands, Peter Williams, and Katrina Dean. *Digital Lives: Personal Digital Archives for the 21st Century. An Initial Synthesis*. London: British Library 2010.

Manovich, Lev. *Software Takes Command*. New York, London: Bloomsbury, 2013.

Morris, Jeremy Wade. *Selling Digital Music, Formatting Culture*. Oakland, CA: University of California Press, 2015.

Moser, Benjamin. "In the Sontag Archive." *The New Yorker*, January 30, 2014. https://www.newyorker.com/books/page-turner/in-the-sontag-archives.

Ries, Thorsten, and Gábor Palkó. "Born-Digital Archives." *International Journal of Digital Humanities* 1, no. 1 (April 2019): 1–11. https://doi.org/10.1007/s42803-019-00011-x.

Snickars, Pelle. "Hårddisken och samtiden [The Hard Drive Today]." In *The Story of Storage I*, edited by Lars Björk, Jānis Krēsliņš, and Matts Lindström, 44–51. Stockholm: Kungliga biblioteket, 2010.

Sterne, Jonathan. "Out with the Trash: On the Future of New Media." In *Residual Media*, edited by Charles R. Acland, 16–31. Minneapolis, MN: University of Minnesota Press, 2007.

Stoll, Clifford. *The Cuckoo's Egg: Tracking a Spy Through the Maze of Computer Espionage*. New York: Doubleday, 1989.

YouTube Podcasting, the New Orality, and Diversity of Thought: Intermediality, Media History, and Communication Theory as Methodological Approaches

Christer Johansson

New media in the postdigital age are sometimes said to promote political polarization, antagonism, and different kinds of hate speech. This is especially true of social media, like Facebook and Twitter. We are by now all familiar with internet trolls stirring the pot, trying to start shitstorms, and downright online harassment. However, the short and crude interactions on Twitter and in comment sections on different platforms are challenged by a different trend, what we might brand "long-form conversation podcasting," a form of talk show with a host and invited guests.[1] In this chapter I explore two exponents of this trend, the YouTube show *The Joe Rogan Experience* and its Swedish cousin, *Hur kan vi?* (*How Can We?*).[2] *The Joe Rogan Experience* was launched on December 24, 2009, and has by now published more than 1,400

[1] This label is inspired by the website of *The Joe Rogan Experience*, https://www.joerogan.com. For more on podcasting, see: Dario Llinares, Neil Fox, and Richard Berry, eds., *Podcasting New Aural Cultures and Digital Media* (Cham: Springer International Publishing, 2018).

[2] *Hur kan vi?* is directly inspired by *The Joe Rogan Experience*, as is evident in this clip: *Hur kan vi?*, "Därför ska jag tatuera in Joe Rogans ansikte på min rumpa" ("That is why I will tattoo Joe Rogan's face on my butt"), YouTube, December 9, 2019, https://www.youtube.com/watch?v=LLY3cItp2so.

How to cite this book chapter:
Johansson, Christer. "YouTube Podcasting, the New Orality, and Diversity of Thought: Intermediality, Media History, and Communication Theory as Methodological Approaches." In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 253–284. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.k.

episodes. It is published on YouTube three to four times a week (until recently it aired live). *Hur kan vi?* was launched on May 21, 2018, and has published 88 episodes. It is recorded live-on-tape, in real time; the episodes are published once a week, and a couple of times every season the show organizes public live events streaming on YouTube. Both shows have sparked off debate on free speech, and some of the invited guests have been questioned and denounced as controversial and/or undemocratic.

The research question I would like to confront *The Joe Rogan Experience* and *Hur kan vi?* with is a very simple one: what are they? It relates to the digital human sciences' concern with "the study of digital objects and environments and their significance for human beings and society," and more specifically to the "interaction between human beings in digital systems and between human beings and digital entities" part of the definition of the field.[3] I will try to show how this YouTube phenomenon, the long-form conversation podcast, can be analyzed from a digital human sciences perspective, combining three transdisciplinary approaches: (1) an intermedial analysis focusing on the relations between media and modalities, (2) a media-historical analysis of the remediational structures and cultural status of orality, and (3) a communication-theoretical analysis of the current channeling of information. In other words, I will successively describe the podcasts as intermedial objects, their position in media history, and their place in today's medial infrastructure.

Intermediality and Multimodality: A Formal Analysis

The first step in the analysis of the podcasts is to characterize what distinguishes these digital entities as intermedial constellations. I will describe them in terms of media relations and as multimodal artifacts. A number of researchers, influenced by semiotics, such as Werner Wolf, Irina O. Rajewsky, and Hans Lund, working in the field of intermediality studies, are concerned with schematizing media interrelations in typologies based on formal qualities. We can distinguish between "extracompositional" and "intracompositional" intermediality (intermediality in the narrow sense), to

3 See the introduction to this volume.

begin with. Extracompositional intermediality means, for example, transmedial qualities common to all or many representations, such as narrativity or rhythm, or transformations, such as film adaptations of novels.[4] As I try to show later in this chapter, the podcasts could be viewed as transformations of sorts, mediations of spoken dialogue, and remediations of radio as a medium, but for now I concentrate on various aspects of their intracompositional intermediality. Furthermore, within this category, we can distinguish between "intermedial reference" ("using a single semiotic system") and "plurimediality" ("displaying signifiers that appear to belong to more than one semiotic system").[5] Intermedial references, that is, references to an absent medium by way of the first medium's own media specific means, may be implicit (intermedial imitation, such as literature imitating film or music) or explicit (intermedial thematization, such as literature about film or music). Within the category of plurimediality, two, or more, distinct media are present in one object in their own materiality. Here, we can distinguish between intermedial interreference (such as illustrations or picture and title), combination (such as comics, opera, or video art), and fusion (concrete poetry or verbal signs in a pictorial representations).

The Joe Rogan Experience and *Hur kan vi?* are inherently intermedial, that is, a combination of sound, mostly linguistic sounds,

---

4 These categories and distinctions are Werner Wolf's, "The Relevance of Mediality and Intermediality to Academic Studies of English Literature [2008]," in *Selected Essays on Intermediality by Werner Wolf (1992–2014): Theory and Typology, Literature-Music Relations, Transmedial Narratology, Miscellaneous Transmedial Phenomena*, eds. Bernhart Walter and Werner Wolf (Leiden: Brill, 2018), 127–152. Similar categories and distinctions are formulated by Irina O. Rajewsky, "Border Talks: The Problematic Status of Media Borders in the Current Debate about Intermediality," in *Media Borders, Multimodality and Intermediality*, ed. Lars Elleström (Basingstoke: Palgrave Macmillan, 2010), 51–68; Irina O. Rajewsky, "Intermediality, Intertextuality, and Remediation," *Intermediality: History and Theory of the Arts, Literature and Technologies*, no. 6 (2005): 43–64. See also Christer Johansson and Sonya Petersson, "Introduction," in *The Power of the In-Between: Intermediality as a Tool for Aesthetic Analysis and Critical Reflection*, eds. Sonya Petersson et al. (Stockholm: Stockholm University Press, 2018), 1–21.

5 Wolf, "The Relevance of Mediality and Intermediality," 127–152.

moving images, occasional expositions of websites, and the viewers' written commentaries in the YouTube comment sections and in live chatrooms. The podcast episodes are instances of intracompositional, plurimedial intermediality, that is, composites of images and sounds, two media materially present in one configuration. This intracompositional intermediality, moving images and synchronized sound, is of course highly conventional, to the point of being considered a medium in its own right, and could thus be thought of as an instance of intermedial fusion, perhaps the dominant one (except for the internet itself) at this point in time of media history. The occasional presence (in the form of mimetic representations) and discussions (thematizations of sorts) of websites, themselves intermedial interreferences, combinations, and fusions, and the written messages of the comment sections, have a quality of indirectness. The latter functions as intermedial references to other digital objects and to the primary digital object (the present audiovisual podcast episode), respectively, rather than being part of the intracompositional intermediality. The podcast episodes thus not only are audiovisual intermedial combinations but also contain intermedial references to various websites and are the targets of intermedial references by way of written commentaries, thus including the viewers, and their interactivity, in the shows. This description is probably valid for many or most "talk shows" on YouTube and other similar video-sharing websites.

A different but related way to approach media relations and media compounds is, following W. J. T. Mitchell and Lars Elleström, to analyze them by means of the multimodal conception of mediality, less based on formal typologies and more focusing on the role and function of sensory, perceptual, and interpretative interaction with media.[6] According to Mitchell, a medium and its mediation

[6] W. J. T. Mitchell, "There Are No Visual Media," *Journal of Visual Culture* 4 (2005): 257–266; Lars Elleström, ed., "The Modalities of Media: A Model for Understanding Intermedial Relations," in *Media Borders, Multimodality and Intermediality* (Basingstoke: Palgrave Macmillan, 2010), 11–48. For a related, practical, application of "synesthetic prototyping" in digital recreations of ancient soundscapes, see Anna Foka and Viktor Arvidsson, "Experiential Analogies: A Sonic Digital Ekphrasis as a Digital Humanities Project," *Digital Humanities Quarterly* 10, no. 2 (2016): paragraphs 1–50.

always entail some mix of the sensory, perceptual, and semiotic elements. All media are necessarily approached by the senses (sight, hearing, touch, etc.) and prompted by "semiotic operators," such as the Peircean triad of iconic (based on similarity), indexical (based on contiguity), and symbolic (based on conventionality) signs.[7] Mitchell's point is that all media are "mixed media," but not "mixed in the same way."[8] Elleström's multimodal concept follows and elaborates the logic of Mitchell's: four "modalities" are understood as present in all media, but in different ways, to different degrees, and in different combinations. Elleström distinguishes between "the material modality," "the sensorial modality," "the spatiotemporal modality," and the "semiotic modality."[9] The sensorial modality regards the physical and mental acts of perceiving the medial interface, and the material modality includes the "latent corporeality of the medium," whereas the spatiotemporal modality covers "the structuring of the sensorial perception of sense-data of the material interface into experiences and conceptions of space and time."[10] The material interfaces condition sensory inputs, which give rise to perceptions that are structured in space and time and understood as signifying.

So, how should the multimodal mix of the podcasts be described? The material modality, to begin with, has a quality characteristic of many podcasts today: it consists of two different digital versions posted on different internet platforms. *The Joe Rogan Experience* and *Hur kan vi?* are published on a number of platforms: on YouTube as filmed audiovisual shows, including sound and image tracks, and on other platforms, such as Acast, iTunes, and Spotify, as audio files. The materiality of these shows are thus of two different kinds: the one multimedial (consisting of two different basic media, audiotext and moving images), the other monomedial (consisting of one basic media, audiotext), which means that the combination of images and sounds could not, in

7 Mitchell, "There Are No Visual Media," 261; Charles S. Peirce, "Logics as Semiotic: The Theory of Signs," in *Semiotics: An Introductory Anthology*, ed. Robert E. Innis (Bloomington, IN: Indiana University Press, 1985), 1–23.
8 Mitchell, "There Are No Visual Media," 260.
9 Elleström, "The Modalities of Media," 17–24.
10 Elleström, "The Modalities of Media," 18.

this case, be considered an intermedial fusion but a combination where the channels are possible to separate. This situation has obvious implications for the "the sensorial modality." The YouTube versions of the shows engage both hearing and sight, while the audio versions only engage the former sense. This fact, in turn, has implications for the spatiotemporal modality of the podcasts, as the spatial location is implicit in the audio versions, while the dimension of time is central in both versions. From these observations we can conclude that, in the multimodal mix, the soundtrack and hearing, and the temporal mode are superordinate to the image track, sight, and space. Otherwise, it would not make sense to publish the podcast episodes without the image track. An investigation of the semiotic modality seems to confirm this interpretation. Of central importance are the spoken words, conventional signs amplified by facial expressions and body language, that is, a combination of indexical and iconic signs. The most important iconic aspect is, however, a temporal one. The recorded conversations air uncut as they evolve, as temporal continuities, without any deletions. The episodes of *The Joe Rogan Experience* and *Hur kan vi?* could thus be described as a sort of diagrammatic icons, the temporal order and duration of the actual conversation coinciding with the temporal order and duration of the episode.[11] All in all, indexical and iconic sign relations are, it could be argued, in service of the conventional signs of the spoken word.

In conclusion, an intermedial and multimodal analysis indicates, on the one hand, that the podcasts here are typical contemporary YouTube instances of intracompositional, plurimedial combinations (audiotext and moving images) referencing other parts of the internet and being commented on. On the other hand, the analysis indicates that the medial components, partly due to the diversification of the materiality of the episodes, are differently weighted: sound is more important than image, hearing more important than sight, time more important than space, and conventional linguistic signs more important than indexical and iconic sign relations.[12]

11 Peirce, "Logic as Semiotic."

12 For more on YouTube as multimodal text see Phil Benson, *The Discourse of YouTube: Multimodal Text in a Global Context* (New York: Routledge, 2017).

A rather obvious interpretation of the constitution of this multimodal complex is that, among all media, modalities, and modes, the spoken word, oral communication and dialogue, takes pride of place in the multimodal mix. This brings us to another dimension of the intermedial/multimodal analysis, put on hold until now, namely the distinction between mediation and transformation. These terms designate the fact that sometimes, or perhaps most of the time, the content of one medium is another medium. The other medium is in some cases mediated, that is, represented without the loss of essential modal qualities, in other cases transformed, that is, changed as a consequence of differences between the signifier of the transforming medium and the qualities of the medium being transformed.[13] We could argue, I believe, that the podcasts mediate oral dialogue, even though the audio version is a less rich mediation, lacking the visual dimension. When listening to the podcast episodes we are actually listening to the conversations between the hosts and their guests, and when watching the YouTube shows the image tracks contribute, by way of iconic and indexical representations, additional information to the context of the dialogue. The relations between the podcast episodes as representations and the mediated dialogue are, however, I will argue, more complicated, demanding the introduction of yet another concept: remediation, which is always a transformation. Remediation is compatible with the concepts of intermediality and multimodality, as it designates relations between media, but by making use of this concept we leave the formal analysis of intermediality studies, and enter the domain of media-historical explorations.

Media History as Remediation: The Refashioning of the Radio and Televison Talkshow

Inspired by Marshall McLuhan, in the late '90s David Bolter and Richard Grusin coined the term "remediation" to describe a kind of intermedial relationship in which both old and new media, by way of processes of medial refashioning, are involved

[13] Lars Elleström, *Media Transformation: The Transfer of Media Characteristics Among Media* (Basingstoke: Palgrave Macmillian, 2014), 12–35.

in "competition or rivalry," and struggling for cultural status, either through "immediacy" (concealing media) or through "hypermediacy" (foregrounding media). Remediation is thus, Bolter and Grusin argue, a "defining characteristic" of digital media and, at the same time, a fundamental characteristic of all medial practices.[14]

It is my contention that podcast shows such as *The Joe Rogan Experience* and *Hur kan vi?* not only mediate oral dialogue but also remediate and transform the medium of radio and, more specifically, the genre of the radio talk show. When watching *The Joe Rogan Experience* on YouTube, it is obvious that the images show us the interior of something akin to a radio studio, a studio designed for the recording of spoken dialogue. The full shots represent a studio with red brick walls, an American flag on the wall behind Joe, always situated to the left, and a dark velvet curtain on the wall behind the guest(s), always to the right. The studio also contains some gadgets and water bottles on the table between the host and the guest, heavy microphones on tripods, big earphones on the host and the guests, and, in the foreground, a digital mixer on a computer screen. The show mainly consists of shot/reverse shot sequences, following the dialogue, representing the speaker in close-up shots (the microphones held by microphone arms and the headphones being prominent), now and then replaced by full shots showing both host and guest(s), as well as the interior of the studio.

Hur kan vi? is modeled after the patterns of *The Joe Rogan Experience*, and the basic structure is the shot/reverse shot sequences, representing the dialogue, as described above. But the Swedish show also differs from the original. With the exception of some of the early shows, the full shots are lacking, and the headphones are optional: sometimes they are used by all involved, sometimes neither host nor guest have them on, and sometimes the host wears them while the guest does not, and, in the latest version of the show, the headphones are no longer used. In general, the radio studio connotations are less emphasized in *Hur kan vi?*, and the episodes are shot in various premises (in the latest

---

14 Jay David Bolter and Richard Grusin, *Remediation: Understanding New Media* (Cambridge: MIT Press, 1999), 45, 70–72.

version of the show there is a new studio with partly different connotations; for more on this, see below). In both shows, the recording process as such is referred to every now and then, by the host or by both host and/or guests, and the big microphones are adjusted for the sound level to be optimal. So, both podcasts are, to an extent, remediating and sometimes even thematizing radio and the sound recording studio as media. Perhaps there is a predecessor in the *Howard Stern Show*, a radio show that early on combined radio and TV, filming the events in the studio, to this day posting clips from the radio show on YouTube and the SiriusXM app.[15] Joe Rogan has himself identified the radio comedy talk show *Opie and Anthony* as his most important influence, stating that it was the lack of structure and the improvisation that he found interesting.[16]

Remediations are, however, never just recreations of older media and media situations but also transformations of the remediated media patterns, challenging the other medium. Consequently, the YouTube podcasts do things that conventional radio does not. The filming of the conversations and the studio, the combination of image and soundtrack, is of course the most obvious transformation. Another difference between radio talk shows and the podcasts is the fluid time format. In distinction to conventional, linear radio, the podcast episodes lack a standardized duration, and are usually much longer than the typical radio show, from one to over four hours. The duration of the conversations determines the duration of the episodes, not the other way around.

In addition to this, and analogous to the radio case, the podcasts could be viewed as remediations of the genre of the television talk show. We all know the pioneering ones like *The Tonight Show*

15 For more on the *Howard Stern Show*, see the Howard Stern website, https://www.howardstern.com.

16 *Opie and Anthony* is an American radio show cohosted by Gregg "Opie" Hughes and Anthony Cumia that aired from 1995 to 2014, with comedian Jim Norton serving as third mic beginning in 2001. For more information on the show, see Internet Archive website, "Opie and Anthony Radio Show Archive," https://archive.org/details/opieandanthonyarchive?tab=about; Joe Rogan on the *Opie and Anthony* show: "Joe Rogan on the Influence of Opie & Anthony," YouTube, December 6, 2017, https://www.youtube.com/watch?v=ttZZ30BwHbI; https://www.youtube.com/watch?v=yXZn2DkrXig.

(1954–) in the USA, and *Hylands hörna* (1962–83) in Sweden.[17] As in the traditional talk shows, there are hosts and one or more guests in the podcasts and the conversations are focused on the guests, as persons and professionals.[18] That being said, the podcast shows under scrutiny differ more from television talk shows than they resemble them. There are no house bands, no live audiences (except for the *Hur kan vi?* live events), and, more importantly, no interviews and, again, no time limits or predetermined segments. The hosts do not interview their guests; what we experience are spontaneous conversations between (more or less) equals. A case in point is a recent *JRE* episode (#1413) guested by Bill Maher, a fellow comedian, political commentator, and television host for HBO show *Real Time with Bill Maher*.[19] The episode is thus a meeting, and confrontation, between the old format of the political talk show and the new format of the long-form conversation podcast. At the beginning of the episode the two hosts/comedians discuss whether Bill feels constrained by the hour format of his show and he arrives at the conclusion that conversations of the kind they are about to have "lend themselves […] to just letting it happen," and we learn that neither the host nor the guest has prepared for the show. What we then experience is thus an almost two-hour-long, spontaneous, unscripted, conversation.[20]

In the latest version of *Hur kan vi?*, the headphones are gone, and the studio is now designed to be viewed, with a round coffee

[17] For more on the talk show as genre, see Wayne Munson, *All Talk: The Talkshow in Media Culture* (Philadelphia, PA: Temple University Press, 1993).

[18] It could be argued that YouTube is, together with subscription-based streaming services such as Netflix and HBO, about to replace traditional, linear television, and that shows such as *The Joe Rogan Experience* and *Hur kan vi?* are part of that shift. Notice that YouTube is also a remediation of the video player. For more on YouTube and television, see *The YouTube Reader*, eds. Pelle Snickars and Patrick Vonderau (Stockholm: National Library of Sweden, 2009); on YouTube and interactivity, see Jean Burgess and Joshua Benjamin Green, *YouTube: Online Video and Participatory Culture* (Cambridge: Polity, 2009).

[19] For more information on *Real Time*, see "Real Time with Bill Maher," You Tube, https://www.youtube.com/channel/UCy6kyFxaMqGtpE3pQTflK8A.

[20] *The Joe Rogan Experience*, "Episode #1413 – Bill Maher," You Tube, January 17, 2020, https://www.youtube.com/watch?v=-KQGZa773sI.

table with cups and a water bottle, two visible armchairs, and a lot of gadgets in the shelf in the background, showed in wider shots than in the previous seasons. On the first show after the summer break, the host Navid Modiri refers to the new studio, comparing it to a well-known Swedish "experimental" television talk show in the past, *Knesset* at ZTV, which ran between 1995 and 1997.[21] "Punk" television is now explicitly appropriated and remediated, and transformed, following the patterns of *The Joe Rogan Experience* as described above.

To summarize, *The Joe Rogan Experience* and *Hur kan vi?* remediate radio as a medium and the radio talk show as a genre. The shows even include a discreet metalevel concerned with the recording process, the studio, and the recording equipment. At the same time, the image track, showing the recording situation, and the extended, variable duration of the episodes are transformations of the radio medium beyond its conventional limits. The same is, to a lesser extent, true of television as a medium and the televised talk show as a genre. The remediation transforms the traditional television format into a spontaneous, unscripted conversation.

Media History and the (New) Oral Paradigm

Another way to approach the historical aspects of the podcasts here is to relate them to the medium of writing and the media landscapes of the past and the present. The founding scholars of media-historical research, such as Jack Goody, Eric Havelock, Marshall McLuhan, and Walter Ong, investigate the domain of orality, its relation to writing, and, more generally, the relationship between communication, technology, and society.[22] Let us explore

---

[21] "Ztv Knesset," YouTube, https://www.youtube.com/channel/UCFbXdDDRFogVAnPt2cFr-bA/videos.

[22] Marshall McLuhan, *Understanding Media: The Extensions of Man*, critical ed. (Corte Madera: Ginko Press, 2003 [1964]); Jack Goody, *Myth, Ritual and the Oral* (Cambridge: Cambridge University Press, 2010); Eric Alfred Havelock, *The Muse Learns to Write: Reflections on Orality and Literacy from Antiquity to the Present* (New Haven, CT: Yale University Press, 1986); Walter J. Ong, *Orality and Literacy: The Technologizing of the Word* (London: Methuen, 1982). For criticism of these approaches see Jan Assmann, *Cultural Memory and Early Civilization: Writing,*

what this early research on the spoken word and oral culture can tell us about the new orality of long-form conversational podcasts like *The Joe Rogan Experience* and *Hur kan vi?*

Ong argues that the characteristics of the oral world, before writing, are that it is "formulaic," conservative, "close to the human world," "agonistically toned," empathetic, homeostatic, situational, communal, and involves memorization by formula rather than verbatim. The literate world is the opposite to all of these things, as it is abstract, analytic, distancing, objective, and separative.[23] These differences can be derived from the qualities of the media involved: the existence of sound is volatile, always in process; marks on visual, external surfaces, that is, writing, are isolating, dissecting, analytical, and fixed in a way that sound is not. Owing to the basic differences between sound and vision, the advent of literacy leads to the development of a new and different consciousness. Writing enables a shift from a prelogical to a logical mentality: the distinction of myth from history, the development of science, objectivity, critical thought, and abstraction.[24] Similar thoughts are formulated by Goody and Havelock.

According to McLuhan, there is a struggle between orality and literacy, reinforced by printing, and between tribal man and literal man. McLuhan detects a break between "the auditory and the visual experience of man." Only the phonetic alphabet makes such a sharp division in experience possible, "giving to its user an eye for an ear, and freeing him from the tribal trance of resonating word magic and the web of kinship."[25] Man's development from tribal to literate means that he is "emotionally free to separate from the tribe and to become a civilized individual, a man of visual organization who has uniform attitudes, habits, and rights with all other civilized individuals."[26] The Western world is homogenized, through segmentation and repeatability, McLuhan argues.[27]

Remembrance, and Political Imagination, 1st English ed. (Cambridge: Cambridge University Press, 2011).

[23] Ong, *Orality and Literacy*.

[24] McLuhan, *Understanding Media*, 30, 33, 118.

[25] McLuhan, *Understanding Media*, 120.

[26] McLuhan, *Understanding Media*, 118.

[27] McLuhan, *Understanding Media*, 75, 237, 237, 239.

However, this also means that man is fragmented; writing, and printing, extends the human mind but at the same time dissociates senses and faculties.[28] Printing, the mechanization of the word, consolidates this development; there is a typographic explosion of the consciousness of the human mind.[29] Tribalism and the clan are exploded by printing.[30]

This balance between speech and writing is changed by the, in McLuhan's terms, electric technology:

> The alphabet (and its extension into typography) made possible the spread of the power that is knowledge, and shattered the bonds of tribal man, thus exploding him into an agglomeration of individuals. Electric writing and speed pour upon him, instantaneously and continuously, the concerns of all other men. He becomes tribal once more. The human family becomes one tribe again.[31]

Because of its extension of our central nervous system, electric technology (the media of telephone, radio, and TV), McLuhan argues, seems to favor the inclusive and participational spoken word over the specialist written word.[32] McLuhan's description dovetails with Ong's concept of "secondary orality." Secondary orality is dependent on literate culture and the existence of writing, manifested, for example, in the phonograph and radio, and it is a "a more deliberate and self-conscious orality."[33]

Long-form conversation podcasts like *The Joe Rogan Experience* and *Hur kan vi?* could and should thus be analyzed as a form of secondary orality, with deep roots in primary orality. They could be viewed as a continuation and an updated version of the orality characteristic of the electric age of radio and television, as described by McLuhan. The consciousness of the podcasts has, it could be argued, many of the characteristics of orality pointed out by McLuhan. The podcasts are, on the one hand, empathetic,

28 McLuhan, *Understanding Media*, 114, 120.
29 McLuhan, *Understanding Media*, 233.
30 McLuhan, *Understanding Media*, 240.
31 McLuhan, *Understanding Media*, 234.
32 McLuhan, *Understanding Media*, 117; cf. 21, 65, 70, 75, 235–236.
33 Ong, *Orality and Literacy*, 133; cf. Walter J. Ong, *Rhetoric, Romance, and Technology: Studies in the Interaction of Expression and Culture* (Ithaca, NY: Cornell University Press, 1971).

situational, communal, inclusive, and participational; on the other, volatile and processual. And perhaps podcasts, like the ones under scrutiny, finally realize McLuhan's dream of the retribalization of man.

There is a sense of enhanced community to these projects, and perhaps a tribal (in McLuhan's sense of the word) quality. Both *The Joe Rogan Experience* and *Hur kan vi?* are all about the inclusive and participational spoken word, and their respective hosts try to bond with all kinds of people interested in expressing themselves and enter the dialogue. They talk to everyone, from left to right. They invite the politically incorrect: the conspiracy theorists, the controversial, the crude, and the crazy. All and everyone are part of the tribe and worth listening to, seems to be the attitude. And, so, Alex Jones from Infowars, democratic socialist Bernie Sanders, leftist journalist Abby Martin, philosopher and political activist Cornel West, and left-wing political pundit and academic David Pakman, as well as conservative political commentator Ben Shapiro and professor of psychology and libertarian Jordan Peterson participate in different episodes of Joe Rogan's podcast. The same is true of *Hur kan vi?*, inviting alternative journalist and libertarian Chang Frick, conceptual artist and activist Lars Wilks, and right-wing activist Ingrid Carlqvist, as well as investigative journalist Janne Josefsson, left leaning comedian Henrik Schyffert, left-wing politician Amineh Kakabaveh, and the former leader of the Swedish Left Party and spokesperson of Feminist Initiative Gudrun Schyman. *The Joe Rogan Experience* and *Hur kan vi?* could thus be interpreted as cultural projects attempting to recreate the lost tribalism of "written," digitalized culture, trying to involve all kinds of deviant voices in the ongoing dialogue.

More importantly, the podcasts seem to constitute part of a counterculture to important aspects of digital mainstream culture. Much of what McLuhan thought was characteristic of literacy and the culture of writing could be used as a description of the effects of the digital platforms and social media of today, especially fragmentation and segmentation, a sense of isolation and distance. This is true of Facebook and Instagram, blogs, vlogs and Twitter, internet forums and comment sections, Snapchat, and memes, as well as of emails and SMS. These media are all, in one way or another, remediations of text types, written genres and linguistic phenomena (the book, the diary, the telegram, the log, the letter, written commentary, the lexeme, mimesis, and so on) extending the human mind and connecting people. At the same time, the mediated messages are brief, functioning as vehicles of short information and/or self-expression, leaving no room for elaborate conversations. Podcasts like *The Joe Rogan Experience* and *Hur kan vi?* can be construed as attempts to be an alternative to the briefness and fragmentation of much of internet culture, giving precedence to continuity. Both podcasts promote hour-long, improvised conversations. The agents involved in the conversation are present to each other and the sense of presentness is, as we have seen, enhanced through the uncut, live format (either live proper or live-on-tape), and through an *in medias res* quality, as if we as viewers or listeners were thrown into an already ongoing conversation. *The Joe Rogan Experience* usually starts with the host, Joe Rogan, announcing that the live feed has begun ("And boom, we're live"), or with the guest and/or the host already talking over the rotating vignette (an image of Joe Rogan with a third, all-seeing eye), fading out, and giving way to the moving images showing the interior of the studio. The *Hur kan vi?* episodes sometimes start mid-conversation, without the guests being aware that the show has begun.[34]

To these early media historians, I would like to add the recent observations of media philosopher John Durham Peters, who has investigated the function of the written word in our digital era, and how it relates to orality. Peters observes that the "written word is one of the clearest historical continuities between early civilization and the latest innovations." New social media takes us back, Peters argues, to the beginnings of distant interaction. And he continues: "digital media, whatever else they are, are machines that convert everyone into writers."[35] In the digital era we are all potential writers, and the original questions of the relations

[34] *Hur kan vi?*, "Episode #69: Janne Josefsson – Jag dör om jag slutar jobba," YouTube, June 30, 2019, https://www.youtube.com/watch?v=ZwSBoglx6RU.

[35] John Durham Peters, *The Marvelous Clouds: Toward a Philosophy of Elemental Media* (Chicago, IL: University of Chicago Press, 2015), 264,

between flesh and word, presence and absence, are once again updated.[36]

The absence characteristic of writing comes with certain benefits that face-to-face communication does not provide. To begin with, as Peters observes, "[t]ext messaging, Twitter, and Facebook, like earlier forms of writing, suspend the risks of real time."[37] The medium of writing offers us a parallel world to the serial universe of face-to-face talk. In online posts and comments, people have time to work on their profiles and presentations, and can thus appear a little better, a little more interesting. Peters continues: "As Plato complained, writing can serve as a performance-enhancing drug; you always get a second chance. On Facebook, everybody gets to be Oscar Wilde."[38] The interaction on social media also provides a bodiless and/or anonymous presence, allowing for behavior beyond the social contract in everyday situations. According to Peters, online interaction reduces "emotional bandwidth." In text messages, you do not have to decide whether to identify yourself, or behave politely. A text message is, Peters explains, decapitated and legless: "Online interaction changes the recipe by seemingly taking the body out of the equation and scrambling the rules of animal behavior."[39] So-called "trolls" and harassers can get away with sociopathic behaviors online that would be hard to do elsewhere. Peters continues: "The online world breeds styles of maliciousness that seem perversely fitted to the narcissism and solitude of the act of writing, something that obeys laws very different from those of orality. Writing takes its revenge in the age of digital media."[40] Therefore, Peters seem to imply that the social media of today foster behaviors closely connected to, and perhaps made possible by, the medium of writing:

> The Internet is a breeding ground for fantasy bodies, "one big orgy, an endless informational bacchanal" for memes, spam, porn,

265. Peters does not share McLuhan's nostalgic approach to the oral past, 264.

[36] Peters, *The Marvelous Clouds*, 266, 273.

[37] Peters, *The Marvelous Clouds*, 273.

[38] Peters, *The Marvelous Clouds*, 274.

[39] Peters, *The Marvelous Clouds*, 275.

[40] Peters, *The Marvelous Clouds*, 275.

> and their human hosts—a destructive, volatile zone not at all the safe, happy place that the Internet companies paint it to be. In this way, it is like an older habitat, writing.[41]

It is, once again, my contention that the podcasts under scrutiny are developed, more or less consciously, in relation to this new form of writing. Unlike the written social media of our time, the podcasts elaborate, work with, the risks of real time: the clock never stops; you have to perform on the spot, a quality enhanced by the lack of script and the reluctance to edit the episodes. It is exactly writing's possibility to stall time and hide our faces, voices, and identities, so typical of social media like Twitter and Facebook, that the podcasts try to abolish. *The Joe Rogan Experience* and *Hur kan vi?* have the ambition to simulate face-to-face communication in real time, to make the presence of human bodies felt. The risks and social anxieties of conversations, pointed out by Peters, are embraced by the podcast shows. Considered from a historical perspective, the podcasts here could be described as anti-writing, if by writing we mean modern social media's exploration of the medium. Just as electronic media could be understood in relation to writing, long-form conversation podcasts can be understood in relation to social media.

Finally, alongside the tribalism 2.0, or perhaps 3.0, the podcasts seem to promote some of the qualities of writing highlighted by researchers such as Goody, Havelock, and Ong. Writing and literacy are, as we have seen, strongly associated with abstraction, analyticity, and critical thought, elements essential to the conversations of *The Joe Rogan Experience* and *Hur kan vi?* The consciousnesses fashioned by these podcasts are thus very much the products of literacy, or perhaps the perfect illustration of the concept of secondary orality, in the sense of a more deliberate and self-conscious orality. One last comment is appropriate. The broadcasting on YouTube means that the channels of *The Joe Rogan Experience* and *Hur kan vi?* have another quality typical of writing: they are living archives for audiovisual text. Transmission and preservation are colliding.

41 Peters, *The Marvelous Clouds*, 277.

Media as Infrastructure and the Bias of Communication: A Synchronic Analysis

I will end this chapter by applying a communication-theoretical perspective to *The Joe Rogan Experience* and *Hur kan vi?*, a perspective that, paradoxically, transcends the phenomenon of communication proper. My ambition is to sketch a synchronic analysis of the media infrastructure constituting the conditions for both podcasts.

In current media theory, as formulated by W. J. T. Mitchell, Mark B. N. Hansen, and John Durham Peters, among others, "medium" is not only considered a semiotic, representational, and communicative phenomenon but something even more fundamental, permeating both nature and culture, the semiotic as well as the nonsemiotic. This environmental media concept enables a contextualizing interpretation of modes and media as agents and infrastructures.[42] In this context, I would like to single out one aspect of the infrastructural understanding of media, as it is formulated by Peters, Friedrich Kittler, and especially Harold Innis. Kittler coins the concept of the "Schriftmonopol," or monopoly of writing,[43] explained by Peters, referencing Innis, in the following way:

> Historically, priests, scribes, scholars, and bureaucrats have had little incentive to make writing technology more accessible. Barriers to entry, as Innis would remind us, aid monopoly control, and the history of writing is not only a succession of scripts but a struggle between groups for control over the means of communication.[44]

For Innis, the focal point of interest is the analysis of politics and power in relation to media.[45] His most important concept,

[42] W. J. T. Mitchell and Mark B. N. Hansen, eds., "Introduction," in *Critical Terms for Media Studies* (Chicago, IL: University of Chicago Press, 2010), vii–xxii.

[43] Friedrich A. Kittler, *Gramofon. Film. Typewriter* (Berlin: Brinkman and Bose, 1986), 12.

[44] Peters, *The Marvelous Clouds*, 296–297.

[45] Harold A. Innis, *The Bias of Communication* (Toronto: University of Toronto Press, 1951); Harold A. Innis, *Empire and Communications* (Toronto and New York: Dundurn Press, 2007 [1950]); Edward Comor, "Harold Innis and 'The Bias of Communication,'" *Information, Communication & Society* 4, no. 2 (2001): 274–293, 277, 282.

formulated in his seminal works on media and communication theory, *The Bias of Communication* and *Empire and Communications*, is "media bias," or, more precisely, "the 'biases' of core institutions, organizations and technologies—the nodal points through which what we know and how we know are produced and reproduced."[46] Innis's work offers a model for the conceptualization of a struggle involving not only time, space, and the temporal or spatial biases of dominant media but also control over knowledge, wealth, and force. Innis views history as the dialectics of "'monopolies of knowledge' and 'monopolies of power'." Monopolies of knowledge means those interests that control access to information, or those interests that has the dominant influence on patterns of social habits and conceptions.[47] Innis's model constitutes a means of assessing potential developments involving how changing media environments affect power relations.[48] In the following, I will focus on the control over knowledge, wealth (here interpreted as economic infrastructure), and force (here interpreted as a struggle within the political sphere), which I consider important parts of a digital human sciences method for analyzing internet culture.[49]

Knowledge, to begin with. It is obvious that digital media platforms like YouTube have changed the dynamics of knowledge production in countries like the USA and Sweden. Former monopolies of knowledge, controlling dominant ways to perceive and think, have been challenged. The medial infrastructure is in a state of transformation. This is true of traditional mass media and of educational institutions. *The Joe Rogan Experience* and *Hur kan vi?* publish intellectual content independent of these traditional institutions, directly reaching their audiences, large in numbers. The former show has, at this point, more than 7.5 million subscribers on YouTube and many episodes have over a million views; some shows tens of millions of views. *Hur kan vi?* has 23,300 subscribers, most episodes have tens of thousands of views and some episodes over a hundred thousand views. Both hosts, Joe Rogan and Navid Modiri, have pasts in mainstream media,

---

46 Comor, "Harold Innis," 276.
47 Comor, "Harold Innis," 281.
48 Comor, "Harold Innis," 284.
49 Comor, "Harold Innis," 284.

Joe Rogan (besides being a stand-up comedian) as a sports commentator and television host, Navid Modiri as a radio and television host, and criticism directed at established knowledge institutions is a theme running through both podcasts. In *Hur kan vi?*, discussions of the problematic status and bias of Swedish public service radio and television and its relation to freedom of speech and democracy are, for example, central topics in episodes with author and satirist Jens Ganman, stand-up comedian and podcast host Aron Flam, and investigative journalist Janne Josefsson.[50] *The Joe Rogan Experience* has become part of the so-called Intellectual Dark Web,[51] a group of public personalities at some point being silenced, marginalized, and even purged from institutions hostile to deviant thoughts, in the media and academia.[52] Dominant figures in this loosely composed group are neuroscientist Sam Harris; the evolutionary biologists Bret Weinstein and Heather Heying; Eric Weinstein, a mathematician and managing director of Thiel Capital; Jordan Peterson, the psychologist and bestselling author; the conservative commentators Ben Shapiro and Douglas Murray; and the feminists Ayaan Hirsi Ali and Christina Hoff Sommers. The core members have little in common politically; their preferences belong all over the political spectrum. The question of free speech, political correctness, and the dangers of identity politics are frequent topics when one or several IDW members visit the *Joe Rogan Experience* podcast. Whatever the topic, the IDW members agree on the importance of conversation, on the possibility of disagreeing without hostility, and of favoring facts before feelings. The podcast has also had James A. Lindsay and Peter Boghossian as guests, the men behind the "grievance studies affair," highlighting what they considered to be poor scholarship

[50] *Hur kan vi?*, "Episode #8: Aron Flam – 'Vi är slavar för staten,'" YouTube, July 8, 2018, https://www.youtube.com/watch?v=UarNWZgzcXc; *Hur kan vi?*, "Episode #11: Jens Ganman – 'Vi lever i en värld där media vrider på verkligheten,'" YouTube, July 22, 2018, https://www.youtube.com/watch?v=KZB7SOoJ3wE; *Hur kan vi?*, "Episode #69: Janne Josefsson – Jag dör om jag slutar jobba," YouTube, June 30, 2019, https://www.youtube.com/watch?v=ZwSBoglx6RU.

[51] "Joe Rogan on Being in the Intellectual Dark Web," YouTube, July 26, 2018, https://www.youtube.com/watch?v=spb8bYIfP1g.

[52] For more information, see Intellectual Dark Web website, https://intellectualdarkweb.site/vanguards-of-the-intellectual-dark-web.

in several academic fields. Their project consisted of submitting bogus academic papers to academic journals in cultural, queer, race, gender, fat, and sexuality studies to determine if they would pass through peer review and be accepted for publication (which several of these papers were).[53]

Both podcasts have been debated on Twitter and in other social media, as well as in traditional media. The *Hur kan vi?* podcast has been contested in traditional Swedish media, especially after the appearance of right-wing extremist Ingrid Carlqvist on January 27, 2019, Holocaust Memorial Day.[54] The day of the release was, according to Modiri, a mistake. The episode, however, became the subject of vigorous debate. Navid Modiri was criticized, in Swedish newspapers like *Arbetet*, *Expressen*, and *Sydvenskan*, for being much too uncritical of the controversial guest, and for trivializing the Holocaust.[55] In interviews in the Swedish magazines *Resumé* and *Dagens Media*, and the newspaper *Svenska Dagbladet*, Modiri says the main goal of the podcast episode is the same as the podcast as a whole: to contribute to a more open climate of discussion where a wider range of opinions could be expressed. This means listening to antagonists, challenging opinions, and talking to people we find annoying. The show's goal is to make people with different opinions talk to each other, as is evident during the *Hur kan vi?* live events. All dialogue and

[53] *The Joe Rogan Experience*, "Episode #1191: Peter Boghossian & James Lindsay," YouTube, October 30, 2018, https://www.youtube.com/watch?v=AZZNvT1vaJg; "The Grievance Studies Affair," YouTube, last updated April 23, 2019, https://www.youtube.com/playlist?list=PLLHyNSlsz44-mMSAXVWXjwRc8KOlA8Vrg.

[54] *Hur kan vi?*, "Episode #44: Ingrid Carlqvist – 'Hijaben används för att män ska få veta vilka kvinnor som får våldtas,'" YouTube, January 27, 2019, https://www.youtube.com/watch?v=yLLj2YMwiRM.

[55] Kolbjörn Guwallius, "'Modiris podd om Förintelsen är ett fullständigt haveri,'" *Arbetet*, January 29, 2019, https://arbetet.se/2019/01/29/modiris-podd-om-forintelsen-ar-ett-fullstandigt-haveri; Dan Korn, "Det fria samtalet har en gräns, Navid Modiri," *Expressen*, February 14, 2019, https://www.expressen.se/kultur/ide/det-fria-samtalet-har-en-grans-navid-modiri; Ida Ölmedal, "Navid Modiri och högerextremismen," *Sydsvenskan*, February 3, 2019, https://www.sydsvenskan.se/2019-02-03/navid-modiri-och-hogerextremismen; Andreas Ekström, "Navid Modiris podd är ansvarslös," *Sydsvenskan*, January 30, 2019, https://www.sydsvenskan.se/2019-01-30/navid-modiris-podd-ar-ansvarslos.

discussions should be acceptable, we should be able to talk to everybody about every- and anything, even Norwegian terrorist Anders Behring Breivik. This is what democracy and freedom of speech are all about, and a prerequisite for a collaborative society.[56] Modiri also stresses that *Hur kan vi?* is something new and different, and that the podcast team is trying to understand what it is. It is not public service, and not traditional journalism, but instead some kind of popular adult education.

This debate is basically about how media should function, and about the difference between new media forms such as the podcasts here and more traditional media outlets. Two different models, directly related to Innis's ideas on monopolies of knowledge, are played out against each other: on the one hand, the conversation activists, trying to fight polarization and radicalization with inclusion; on the other hand, the traditional media strategy using gatekeepers/publishers to decide what debates are worth having.

The discourse surrounding *The Joe Rogan Experience* is similar to the one surrounding *Hur kan vi?* The most important debate concerns the lack of a gatekeeping function. Joe is accused of not being critical enough when interviewing guests such as Elon Musk or Milo Yiannopoulos; he is not behaving like an investigative, critical journalist but as a conversation partner, and he talks to everybody he finds interesting, including people not welcome in mainstream media, right-wing profiles, and tricksters of different kinds, often challenging political correctness and using satire to make fun of contemporary phenomena. Rogan is a free speech absolutist, and he, and his guests, are sometimes drinking and smoking weed during the shows, which is likely to weaken the

[56] Andreas Rågsjö Thorell, "Navid Modiri svarar på kritiken: 'Ett olyckligt och stort misstag,'" *Resumé*, January 31, 2019, https://www.resume.se/kommunikation/media/navid-modiri-svarar-pa-kritiken-ett-olyckligt-och-stort-misstag; John Söderberg, "Navid Modiri: 'Jag står för det öppna samtalet utan kompromiss,'" *Dagens Media*, February 9, 2019, https://www.dagensmedia.se/medier/radio-podd/navid-modiri-jag-star-for-det-oppna-samtalet-utan-kompromiss; Sam Sundberg, "'Ett klumpigt misstag att lägga ut podden den dagen,'" *Svenska Dagbladet*, April 7, 2019, https://www.svd.se/navid-modiri-skulle-se-ett-varde-i-att-bjuda-in-breivik.

gatekeeping even further. It is up to the viewers and listeners to separate the good from the bad, serious discourse from satire.[57] Accordingly, the debate is about the current power struggle between two different media systems, alternative media without traditional gatekeeping functions, and traditional media dependent on publishing decisions and selectivity.

As far as the economic side of things are concerned, the podcasts are part of a new market, quite different from both the Swedish and the American media markets, and many of the guests could be considered stars in the heaven of alternative media. *Hur kan vi?* was launched by way of crowdfunding on Kickstarter and is supported via Patreon and Swish, two channels for the handling of donations. Consequently, closely related to the Swedish debate on gatekeeping is another debate, in new and traditional media, on this new system for financial support.[58] The conversation form

[57] See, e.g., Theodore Kupfer, "Joe Rogan's Boundary-Free Arena," *National Review*, April 13, 2018, https://www.nationalreview.com/2018/04/joe-rogan-podcast-free-speech-zone-non-mainstream-thought; Michael Blight, "'A Conversation isn't a Co-Sign:' An Investigation of Joe Rogan and Deplatforming," *Media Commons*, July 31, 2019, http://mediacommons.org/imr/content/%E2%80%9C-conversation-isnt-co-sign%E2%80%9D-investigation-joe-rogan-and-deplatforming; Dani Di Placido, "Twitter Doesn't Seem to Understand Joe Rogan's Podcast," *Forbes*, January 24, 2020, https://www.forbes.com/sites/danidiplacido/2020/01/24/twitter-doesnt-seem-to-understand-joe-rogans-podcast/#4b89fd1471bd; Devin Gordon, "Why Is Joe Rogan So Popular?," *The Atlantic*, August 19, 2019, https://www.theatlantic.com/entertainment/archive/2019/08/my-joe-rogan-experience/594802; http://www.mediafiledc.com/time-end-joe-rogan-experience. *The Joe Rogan Experience* has also been criticized for inviting too many men, for spreading fake news, for being shallow, and for not being honest about his political conviction. See, e.g., Therese Larsson Hultin, "Miljoner män följer Rogan: 'Man får en annan bild.'" *Svenska Dagbladet*, January 22, 2020, https://www.svd.se/miljoner-man-foljer-rogan-man-far-en-annan-bild.

[58] Sundberg, "'Ett klumpigt misstag att lägga ut podden den dagen'"; Brit Stakston, "Så trivialiserar Navid Modiri en högerextrem propagandist," *Dagens Nyheter*, January 31, 2019, https://www.dn.se/kultur-noje/kulturdebatt/sa-trivialiserar-navid-modiri-en-hogerextrem-propagandist; Marcus Joons, "Granskade Lamotte: 'Sedan tog det hus i helvete,'" *Svenska Dagbladet*, January 11, 2020, https://www.svd.se/har-stalls-svensk-sanning-om-journalistik-pa-huvudet.

is described as a cheap and effective way to sell controversial opinions directly to the audience, via Kickstarter and Swish, and the soundness of fans directly supporting their favorite alternative media and thus indirectly influencing their content is questioned.

The Joe Rogan Experience is funded through ads and commercial endorsements on YouTube and via iTunes. Both YouTube and Patreon, as private, highly influential companies, have been criticized for endangering free speech through so-called deplatforming and demonetizing of certain channels and outlets with supposed offensive content.[59] These controversies, and the dangers of the monopoly of Google and YouTube and its effects on free speech, as well as the hate speech policy of Twitter, have been topics of discussion on both podcasts. *The Joe Rogan Experience* and *Hur kan vi?* are thus part of a new media economy as well as channels for discussions of this economy and its dangers of fostering new and different monopolies of knowledge and power. You can almost experience, when watching and listening to these podcasts, the theories of Innis and Peters, on the bias of media and media as infrastructures, being negotiated before your eyes and ears.

Joe Rogan's podcast has recently been under attack in a context related to political power. In August 2019, Bernie Sanders guested *The Joe Rogan Experience* (like other candidates, Andrew Yang and Tulsi Gabbard, before him) for an hour-long discussion on politics and media.[60] When asked, on a recent episode, who he was planning to vote for in the Democratic primaries, Joe said about Bernie Sanders: "I believe in him, I like him—I like him a lot," nodding to the time he had interviewed Sanders on his show last year. "He's been insanely consistent his entire life."[61] The Sanders campaign used this and tweeted out a video of Rogan's endorsement. The reactions were excessive. Sanders was recommended to reconsider the video, and Rogan was accused of attacking transgender people, gay men, women, people

59 "Joe Rogan on the Patreon Controversy," YouTube, December 19, 2018, https://www.youtube.com/watch?v=2wsQPFyZghc.

60 *The Joe Rogan Experience*, "Episode #1330 – Bernie Sanders," YouTube, August 6, 2019, https://www.youtube.com/watch?v=2O-iLk1G_ng.

61 "Joe Rogan Says He's Voting for Bernie," YouTube, January 23, 2020, https://www.youtube.com/watch?v=ve7ccl3YrHU.

of color, and countless marginalized groups, as well as being accused of promoting Islamophobia and for being a bigot. For the frequent viewer and listener of the podcast, it is obvious that the epithets and accusations hurled at Rogan are untrue; they seem to be the result of a misreading (willful or not) of the Joe Rogan discourse and a misunderstanding of the new media logics characteristic of podcasting. Some journalists defended Rogan; some piled on.[62] Some commentators highlighted an important argument: Joe Rogan is so influential that any serious presidential candidate should accept his endorsement. *The Joe Rogan Experience* has become part of the medial infrastructure of political power in the USA.[63]

The Joe Rogan Experience and *Hur kan vi?* are part of an emerging medial infrastructure, including new controlling agencies, new financial services, and new channels of knowledge and politics. The podcasts are also forums for ongoing conversations about these issues, and the shaping of a new cultural self-understanding.

---

[62] Matt Stevens, "Why a Joe Rogan Endorsement Could Help (or Backfire on) Bernie Sanders," *New York Times*, published January 24, 2020, updated January 26, 2020, https://www.nytimes.com/2020/01/24/us/politics/bernie-sanders-endorsement-joe-rogan.html; Bobby Azarian, "Some Media Outlets Are Gaslighting Us about Joe Rogan," *Psychology Today*, January 25, 2020, https://www.psychologytoday.com/us/blog/mind-in-the-machine/202001/some-media-outlets-are-gaslighting-us-about-joe-rogan; Dylan Matthews, "The Joe Rogan Controversy Revealed Something Important about the American Left," *Vox*, January 27, 2020, https://www.vox.com/future-perfect/2020/1/27/21081876/joe-rogan-bernie-sanders-henry-kissinger; Tom Slater, "Bernie Sanders, Joe Rogan and the Folly of Purity Politics," *Spiked*, January 27, 2020, https://www.spiked-online.com/2020/01/27/bernie-sanders-joe-rogan-and-the-folly-of-purity-politics; Simon Chandler, "Joe Rogan Triggers Internet Outrage Because He Endorsed Bernie Sanders," *CCN*, January 24, 2020, https://www.ccn.com/joe-rogan-triggers-internet-outrage-because-he-endorsed-bernie-sanders.

[63] Joe Rogan's reaction to the controversy: "Joe Rogan Responds to Bernie Sanders Endorsement Controversy," YouTube, January 31, 2020, https://www.youtube.com/watch?v=P-KjcOQPVeI. See also "Panel Reacts to SJW Freakout of Rogan Bernie Support," YouTube, January 24, 2020, https://www.youtube.com/watch?v=fZwTEbDEQF4.

Conclusion

In this chapter, I tried to demonstrate how a combination of media scientific analytical approaches can be applied to a current internet phenomenon, the long-form conversation podcast exemplified by *The Joe Rogan Experience* and *Hur kan vi?* From the angles of a formal, intermedial analysis, via a media-historical analysis focused on remediation and orality, to a communication-theoretical analysis emphasizing the power dynamics of the infrastructures of media, what these podcasts are all about is the spoken word as a vehicle for and a symbol of freedom of speech.

References

Assmann, Jan. *Cultural Memory and Early Civilization: Writing, Remembrance, and Political Imagination*. 1st English ed. Cambridge: Cambridge University Press, 2011.

Azarian, Bobby. "Some Media Outlets Are Gaslighting Us about Joe Rogan." *Psychology Today*, January 25, 2020. https://www.psychologytoday.com/us/blog/mind-in-the-machine/202001/some-media-outlets-are-gaslighting-us-about-joe-rogan.

Benson, Phil. *The Discourse of YouTube: Multimodal Text in a Global Context*. New York: Routledge, 2017.

Blight, Michael. "'A Conversation isn't a Co-Sign:' An Investigation of Joe Rogan and Deplatforming." *Media Commons*, July 31, 2019. http://mediacommons.org/imr/content/%E2%80%9C-conversation-isnt-co-sign%E2%80%9D-investigation-joe-rogan-and-deplatforming.

Bolter, Jay David, and Richard Grusin. *Remediation: Understanding New Media*. Cambridge: MIT Press, 1999.

Burgess, Jean, and Joshua Benjamin Green. *YouTube: Online Video and Participatory Culture*. Cambridge: Polity, 2009.

Chandler, Simon. "Joe Rogan Triggers Internet Outrage Because He Endorsed Bernie Sanders." *CCN*, January 24, 2020. https://www.ccn.com/joe-rogan-triggers-internet-outrage-because-he-endorsed-bernie-sanders.

Comor, Edward. "Harold Innis and 'The Bias of Communication.'" *Information, Communication & Society* 4, no. 2 (2001): 274–294.

Di Placido, Dani. "Twitter Doesn't Seem to Understand Joe Rogan's Podcast." *Forbes*, January 24, 2020. https://www.forbes.com/sites/danidiplacido/2020/01/24/twitter-doesnt-seem-to-understand-joe-rogans-podcast/#315539971bde.

Durham Peters, John. *The Marvelous Clouds: Toward a Philosophy of Elemental Media*. Chicago, IL: University of Chicago Press, 2015.

Ekström, Andreas. "Navid Modiris podd är ansvarslös." *Sydsvenskan*, January 30, 2019. https://www.sydsvenskan.se/2019-01-30/navid-modiris-podd-ar-ansvarslos.

Elleström, Lars, ed. "The Modalities of Media: A Model for Understanding Intermedial Relations." In *Media Borders, Multimodality and Intermediality*, 11–48. Basingstoke: Palgrave Macmillan, 2010.

Elleström, Lars. *Media Transformation: The Transfer of Media Characteristics Among Media*. Basingstoke: Palgrave Macmillan, 2014.

Finnegan, Ruth. *Literacy and Orality: Studies in the Technology of Communication*. Oxford: Basil Blackwell, 1988.

Foka, Anna, and Viktor Arvidsson. "Experiential Analogies: A Sonic Digital Ekphrasis as a Digital Humanities Project." *Digital Humanities Quarterly* 10, no. 2 (2016): paragraphs 1–50.

Goody, Jack. *Myth, Ritual and the Oral*. Cambridge: Cambridge University Press, 2010.

Goody, Jack, and Ian Watt. "The Consequences of Literacy." *Comparative Studies in Society and History* 5, no. 3 (April 1963): 304–345.

Gordon, Devin. "Why Is Joe Rogan So Popular?" *The Atlantic*, August 19,2019. https://www.theatlantic.com/entertainment/archive/2019/08/my-joe-rogan-experience/594802.

Guwallius, Kolbjörn. "'Modiris podd om Förintelsen är ett fullständigt haveri.'" *Arbetet*, January 29, 2019. https://arbetet.se/2019/01/29/modiris-podd-om-forintelsen-ar-ett-fullstandigt-haveri.

Havelock, Eric Alfred. *The Muse Learns to Write: Reflections on Orality and Literacy from Antiquity to the Present*. New Haven, CT: Yale University Press, 1986.

Howard Stern website. https://www.howardstern.com.

Hur kan vi? "Därför ska jag tatuera in Joe Rogans ansikte på min rumpa." YouTube, December 9, 2019. https://www.youtube.com/watch?v=LLY3cItp2so.

Hur kan vi? "Episode #8: Aron Flam – 'Vi är slavar för staten.'" YouTube, July 8, 2018. https://www.youtube.com/watch?v=UarNWZgzcXc.

Hur kan vi? "Episode #11: Jens Ganman – 'Vi lever i en värld där media vrider på verkligheten.'" YouTube, July 22, 2018. https://www.youtube.com/watch?v=KZB7SOoJ3wE.

Hur kan vi? "Episode #44: Ingrid Carlqvist – 'Hijaben används för att män ska få veta vilka kvinnor som får våldtas.'" YouTube, January 27, 2019. https://www.youtube.com/watch?v=yLLj2YMwiRM.

Hur kan vi? "Episode #69: Janne Josefsson – Jag dör om jag slutar jobba." YouTube, June 30, 2019. https://www.youtube.com/watch?v=ZwSBoglx6RU.

Innis, Harold A. *The Bias of Communication*. Toronto: University of Toronto Press, 1951.

Innis, Harold A. *Empire and Communications*. Toronto and New York: Dundurn Press, 2007 [1950].

Intellectual Dark Web website. https://intellectualdarkweb.site/vanguards-of-the-intellectual-dark-web.

Internet Archive website. "Opie and Anthony Radio Show Archive." https://archive.org/details/opieandanthonyarchive?tab=about.

"Joe Rogan on Being in the Intellectual Dark Web." YouTube, July 26, 2018. https://www.youtube.com/watch?v=spb8bYIfP1g.

"Joe Rogan on the Influence of Opie & Anthony." YouTube, December 6, 2017. https://www.youtube.com/watch?v=ttZZ30BwHbI.

"Joe Rogan on the Patreon Controversy." YouTube, December 19, 2018. https://www.youtube.com/watch?v=2wsQPFyZghc.

"Joe Rogan Responds to Bernie Sanders Endorsement Controversy." YouTube, January 31, 2020, https://www.youtube.com/watch?v=P-KjcOQPVeI.

"Joe Rogan Says He's Voting for Bernie." YouTube, January 23, 2020. https://www.youtube.com/watch?v=ve7ccl3YrHU.

Johansson, Christer, and Sonya Petersson. "Introduction." In *The Power of the In-Between: Intermediality as a Tool for Aesthetic Analysis and Critical Reflection*, edited by Sonya Petersson, Christer Johansson, Magdalena Holdar, and Sara Callahan, 1–21. Stockholm: Stockholm University Press, 2018.

Joons, Marcus. "Granskade Lamotte: 'Sedan tog det hus i helvete.'" *Svenska Dagbladet*, January 11, 2020. https://www.svd.se/har-stalls-svensk-sanning-om-journalistik-pa-huvudet.

Kittler, Friedrich A. *Gramofon. Film. Typewriter*. Berlin: Brinkman and Bose, 1986.

Korn, Dan. "Det fria samtalet har en gräns, Navid Modiri." *Expressen*, February 14, 2019. https://www.expressen.se/kultur/ide/det-fria-samtalet-har-en-grans-navid-modiri.

Kupfer, Theodore. "Joe Rogan's Boundary-Free Arena." *National Review*, April 13, 2018. https://www.nationalreview.com/2018/04/joe-rogan-podcast-free-speech-zone-non-mainstream-thought.

Larsson Hultin, Therese. "Miljoner män följer Rogan: 'Man får en annan bild.'" *Svenska Dagbladet*, January 22, 2020. https://www.svd.se/miljoner-man-foljer-rogan-man-far-en-annan-bild.

Llinares, Dario, Neil Fox, and Richard Berry, eds. *Podcasting New Aural Cultures and Digital Media*. Cham: Springer International Publishing, 2018.

Lund, Hans, ed. *Intermedialitet: Ord, bild och ton i samspel*. Lund: Studentlitteratur, 2002.

Matthews, Dylan. "The Joe Rogan Controversy Revealed Something Important about the American Left." *Vox*, January 27, 2020. https://www.vox.com/future-perfect/2020/1/27/21081876/joe-rogan-bernie-sanders-henry-kissinger.

McLuhan, Marshall. *Understanding Media: The Extensions of Man*. Critical ed. Corte Madera: Ginko Press, 2003 [1964].

Media Borders, Multimodality and Intermediality, edited by Lars Elleström. Basingstoke: Palgrave Macmillan, 2010.

Mitchell, W. J. T., and Mark B. N. Hansen, eds. “Introduction.” In *Critical Terms for Media Studies*, vi–xxii. Chicago, IL: University of Chicago Press, 2010.

Mitchell, W. J. T. “There Are No Visual Media.” *Journal of Visual Culture* 4 (2005): 257–266.

Munson, Wayne. *All Talk: The Talkshow in Media Culture*. Philadelphia, PA: Temple University Press, 1993.

Ölmedal, Ida. “Navid Modiri och högerextremismen.” *Sydsvenskan*, February 3, 2019. https://www.sydsvenskan.se/2019-02-03/navid-modiri-och-hogerextremismen.

Ong, Walter J. *Rhetoric, Romance, and Technology: Studies in the Interaction of Expression and Culture*. Ithaca, NY: Cornell University Press, 1971.

Ong, Walter J. *Orality and Literacy: The Technologizing of the Word*. London: Methuen, 1982.

“Panel Reacts to SJW Freakout of Rogan Bernie Support.” YouTube, January 24, 2020. https://www.youtube.com/watch?v=fZwTEbDEQF4.

Peirce, Charles S. “Logics as Semiotic: The Theory of Signs.” In *Semiotics: An Introductory Anthology*, edited by Robert E. Innis, 1–23. Bloomington, IN: Indiana University Press, 1985.

Rågsjö Thorell, Andreas. “Navid Modiri svarar på kritiken: ‘Ett olyckligt och stort misstag.’” *Resumé*, January 31, 2019. https://www.resume.se/kommunikation/media/navid-modiri-svarar-pa-kritiken-ett-olyckligt-och-stort-misstag.

Rajewsky, Irina O. “Intermediality, Intertextuality, and Remediation.” *Intermediality: History and Theory of the Arts, Literature and Technologies*, no. 6 (2005): 43–64.

Rajewsky, Irina O. “Border Talks: The Problematic Status of Media Borders in the Current Debate about Intermediality.” In *Media Borders, Multimodality and Intermediality*, edited by Lars Elleström, 51–68. Basingstoke: Palgrave Macmillan, 2010.

“Real Time with Bill Maher.” You Tube. https://www.youtube.com/channel/UCy6kyFxaMqGtpE3pQTflK8A.

Slater, Tom. "Bernie Sanders, Joe Rogan and the Folly of Purity Politics." *Spiked*, January 27, 2020. https://www.spiked-online.com/2020/01/27/bernie-sanders-joe-rogan-and-the-folly-of-purity-politics.

Söderberg, John. "Navid Modiri: 'Jag står för det öppna samtalet utan kompromiss.'" *Dagens Media*, February 9, 2019. https://www.dagensmedia.se/medier/radio-podd/navid-modiri-jag-star-for-det-oppna-samtalet-utan-kompromiss.

Stakston, Brit. "Så trivialiserar Navid Modiri en högerextrem propagandist." *Dagens Nyheter*, January 31, 2019. https://www.dn.se/kultur-noje/kulturdebatt/sa-trivialiserar-navid-modiri-en-hoger extrem-propagandist.

Stevens, Matt. "Why a Joe Rogan Endorsement Could Help (or Backfire on) Bernie Sanders." *New York Times*, January 24, 2020, updated January 26, 2020. https://www.nytimes.com/2020/01/24/us/politics/bernie-sanders-endorsement-joe-rogan.html.

Stone, Alice. "Why Is 'The Joe Rogan Experience' Podcast so Popular?" *Showbiz CheatSheet*, February 5, 2020. https://www.cheatsheet.com/entertainment/why-is-the-joe-rogan-experience-podcast-so-popular.html.

Sundberg, Sam. "'Ett klumpigt misstag att lägga ut podden den dagen.'" *Svenska Dagbladet*, April 7, 2019. https://www.svd.se/navid-modiri-skulle-se-ett-varde-i-att-bjuda-in-breivik.

"The Grievance Studies Affair." YouTube. Last updated April 23, 2019. https://www.youtube.com/playlist?list=PLLHyNSlsz44-mMSAXVWXjwRc8KOlA8Vrg.

The Joe Rogan Experience. "Episode #1191: Peter Boghossian & James Lindsay." YouTube, October 30, 2018. https://www.youtube.com/watch?v=AZZNvT1vaJg.

The Joe Rogan Experience. "Episode #1330 – Bernie Sanders." YouTube, August 6, 2019. https://www.youtube.com/watch?v=2O-iLk1G_ng.

The Joe Rogan Experience. "Episode #1413 – Bill Maher." YouTube, January 17, 2020. https://www.youtube.com/watch?v=-KQGZa773sI.

The Joe Rogan Experience website. https://www.joerogan.com.

The YouTube Reader, edited by Pelle Snickars and Patrick Vonderau. Stockholm: National Library of Sweden, 2009.

Wolf, Werner. *The Musicalization of Fiction: A Study in the Theory and History of Intermediality*. Amsterdam: Rodopi, 1999.

Wolf, Werner. "The Relevance of Mediality and Intermediality to Academic Studies of English Literature [2008]." In *Selected Essays on Intermediality by Werner Wolf (1992–2014): Theory and Typology, Literature-Music Relations, Transmedial Narratology, Miscellaneous Transmedial Phenomena*, edited by Bernhart Walter and Werner Wolf, 127–152. Leiden: Brill, 2018.

"Ztv Knesset." YouTube. https://www.youtube.com/channel/UCFbXdDDRFogVAnPt2cFr-bA/videos.

YouTube Help. "Report inappropriate content." Accessed February 26, 2020. https://support.google.com/youtube/answer/2802027?co=GENIE.Platform%3DAndroid&hl=en

Mining Art History: Bulk Converting Nonstandard PDFs to Text to Determine the Frequency of Citations and Key Terms in Humanities Articles

Amanda Wasielewski and Anna Dahlgren

Introduction

Text mining and other computational methods are not widely used in art historical scholarship.[1] One rationale for the lack of text mining in art history is that the field concerns itself with the study of visual and material objects rather than text-based ones. Although artists themselves have not produced huge volumes of text to study, art historians and critics *have* produced a large corpus of writing on art. This means that a computational methodology like text mining, which remains untested in art historiography, might provide insights into the state of the field of art history.[2] "Text mining" consists of extracting, sorting, and discovering patterns within a given set of text (typically a very large

[1] Matthew P. Long and Roger C. Schonfeld, "Preparing for the Future of Research Services for Art History: Recommendations from the Ithaka S+R Report," *Art Documentation: Journal of the Art Libraries Society of North America* 33, no. 2 (2014): 192–205.

[2] As far as we have been able to ascertain, computational methods have not yet been applied to art historiography. As Paul B. Jaskot notes, text mining has not been central to other types of art historical research either. See Paul B. Jaskot, "Digital Art History as the Social History of

How to cite this book chapter:
Wasielewski, Amanda, and Anna Dahlgren. "Mining Art History: Bulk Converting Nonstandard PDFs to Text to Determine the Frequency of Citations and Key Terms in Humanities Articles." In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 285–305. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.l.

corpus) through automated/computational means. By processing bodies of text that are too large to sort through manually, such as a decade's worth of journal articles or a bibliography of books on a particular topic, text mining can uncover disciplinary trends and popular reference points unique to that dataset that provide a picture of the influences and biases contained therein.

In order to text mine an article in Adobe PDF format, it is necessary to export clean, plain text from the document. Most PDF articles today are already composed of searchable text (OCR) or have been created digitally and therefore have recognized text/characters. It is possible, therefore, to copy and paste out the text from articles manually when working with just one or a few articles. Manual extraction of text is, however, impractical for a large corpus. Text mining, like most computational methodologies, works best with the largest set of data possible. Indeed, large datasets *call for* the use of computational methodologies such as text mining because of how time-consuming it is to analyze large quantities of text manually.

The objective of this study is to develop a methodology for determining the most frequently appearing terms in a given set of articles, which is nearly impossible to do manually—even for one article. More specifically, the text mining method we outline addresses challenges that are particular to humanities journals such as those in art history, which do not adhere to a standardized format.[3] As noted, text mining is a new tool for art historians, and the majority of "digital art history" projects focus on image

Art: Towards the Disciplinary Relevance of Digital Methods," *Visual Resources* 35, no. 1–2 (April 3, 2019): 24.

[3] We are aware of that a large amount of scholarly writing in art history is published in monographs and edited volumes. These books are not easy to text mine at the present time, as many of them are either not available as e-books or not collected in a comprehensive way in any given online database. One reason for this is that the high cost of image permissions paired with centrality of images to the discipline makes mass digitization a difficult undertaking. See Long and Schonfeld, "Preparing for the Future," 203; Maureen Whalen, "What's Wrong with This Picture? An Examination of Art Historians' Attitudes About Electronic Publishing Opportunities and the Consequences of Their Continuing Love Affair with Print," *Art Documentation: Journal of the Art Libraries Society of North America* 28, no. 2 (Fall 2009): 13–22.

collections rather than text or historiography.[4] Since there is a shortage of research that addresses this particular area—that is, methodologies for text mining art history publications—the following chapter constitutes a unique contribution to the field of art history and has wider implications for humanities and social sciences digital scholarship in general.

Background

Journal databases for art history scholarship commonly store articles in PDF, which preserves the look and feel of printed journals in digital form. PDFs maintain the layout, pagination, and images for journal articles. The latter is particularly important for art history because we use images not merely as illustration but as essential components of our scholarly arguments.[5] Additionally, citation standards across disciplines demand page citations, which other online and text-based formats do not provide. For example, ebook formats are often problematic for citation because many of them are only available in file formats that are designed to vary the size of the text and the pagination according to the device and the readers' preference. Despite the advantages researchers gain by viewing journal articles in PDF format, the fact that articles may only be available in this format is a significant barrier for text mining.

Text mining in academic journals falls into four main categories: abstract mining, full text mining, metadata mining, and citation mining.[6] Text mining techniques for scientific articles in the

---

4 See articles published in the *International Journal for Digital Art History*, on the website International Journal for Digital Art History, https://dahj.org.

5 Keith Moxey, "Visual Studies and the Iconic Turn," *Journal of Visual Culture* 7, no. 2 (2008): 139.

6 There are many examples one could cite from each of these categories of research, e.g.: Sharon Block and David Newman, "What, Where, When, and Sometimes Why: Data Mining Two Decades of Women's History Abstracts," *Journal of Women's History* 23, no. 1 (March 10, 2011): 81–109; David Westergaard et al., "Text Mining of 15 Million Full-Text Scientific Articles," BioRxiv: The Preprint Server for Biology, published July 11, 2017, https://doi.org/10.1101/162099; Yang Yang et al., "The Researcher Social Network: A Social Network Based on Metadata of

STEM (science, technology, engineering, and mathematics) fields are well developed, but there is far less research that addresses text mining in humanities scholarship.[7] Journals and databases in the natural and social sciences have aided the process of text mining by providing inbuilt tools for analytics as well as structured text formats for articles. According to Giovanni Colavizza and Matteo Romanello,

> Citation mining is by now a "solved problem" with respect to STEM literature. Despite some drawbacks and margins for improvement, the crawlers behind mainstream indexes such as Google Scholar, Web of Science, Scopus, and Dimensions are largely capable of mining most citations accurately. The main problem which is left open is the skewness in literature coverage and mining

Scientific Publications," *Proceedings of WebSci '09: Society On-Line*, Athens, Greece, Mars 18–20, 2009, https://eprints.soton.ac.uk/267156; Iana Atanassova, Marc Bertin, and Philipp Mayr, "Mining Scientific Papers for Bibliometrics: A (Very) Brief Survey of Methods and Tools," paper presented at the 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 2015, arXiv: Preprint Archive:1505.01393, https://arxiv.org/abs/1505.01393v1.

7 Aside from the many STEM field studies that have looked specifically at *academic literature* and/or *journal articles* in a particular discipline such as biomedical science, some of which are cited below, there have also been studies that looked at the literature of humanities disciplines (if not art history), which include: Block and Newman, "What, Where, When, and Sometimes Why"; David Mimno, "Computational Historiography: Data Mining in a Century of Classics Journals," *Journal on Computing and Cultural Heritage* 5, no. 1 (April 2012): 3:1–3:19; Andrew Goldstone and Ted Underwood, "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us," *New Literary History* 45, no. 3 (November 7, 2014): 359–84; Allen Beye Riddell, "How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models," *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, eds. Matt Erlin and Lynne Tatlock (Rochester, NY: Camden House, 2014), 91–114; Benjamin M. Miller, "The Making of Knowledge-Makers in Composition: A Distant Reading of Dissertations" (Graduate Center, CUNY, 2015); Elisabeth Günther and Emese Domahidi, "What Communication Scholars Write About: An Analysis of 80 Years of Research in High-Impact Journals," *International Journal of Communication* 11 (July 26, 2017): 21; Lino Wehrheim, "Economic History Goes Digital: Topic Modeling the Journal of Economic History," *Cliometrica* 13, no. 1 (January 1, 2019): 83–125.

performance over different disciplines, with those within the humanities usually faring worse than most.[8]

STEM and social science researchers have addressed issues of text mining in journal articles using both journal indexes and PDF articles.[9] In one study of 15 million full-text articles in biomedical science, researchers found that, in addition to utilizing available structured text formats for articles, they *also* needed to address the issue of PDF to text/structured text conversion so that they could cover the vast array of articles included in their study.[10]

In the humanities, text/XML extraction from PDF is essential to forming a workable dataset. There are a number of tools that have been developed specifically to recognize the sections of a PDF journal article and convert it into structured format for text mining. While these tools were, for the most part, designed to parse natural and social science literature, many of them can also be adapted for use within humanities PDFs until more appropriate tools are developed. The techniques that have been created by

---

8 Giovanni Colavizza and Matteo Romanello, "Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead," *Journal of European Periodical Studies* 4, no. 1 (June 30, 2019): 37.

9 For example, see: Chiquito J. Crasto et al., "Text Mining Neuroscience Journal Articles to Populate Neuroscience Databases," *Neuroinformatics* 1, no. 3 (September 1, 2003): 215–237; Stephan Dahl, "Current Themes in Social Marketing Research: Text-Mining the Past Five Years," *Social Marketing Quarterly* 16, no. 2 (June 1, 2010): 128–136; Allan Peter Davis et al., "A CTD–Pfizer Collaboration: Manual Curation of 88 000 Scientific Articles Text Mined for Drug–Disease and Drug–Phenotype Interactions," *Database: The Journal of Biological Databases and Curation*, 2013, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3842776; Tai-Quan Peng et al., "Mapping the Landscape of Internet Studies: Text Mining of Social Science Journal Articles 2000–2009," *New Media & Society* 15, no. 5 (August 1, 2013): 644–664; P. K. Jayasekara and Abu K. S., "Text Mining of Highly Cited Publications in Data Mining," in *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)* (Red Hook, NY: Curran Associates Inc., 2018), 128–130.

10 Westergaard et al., "Text Mining of 15 Million Full-Text Scientific Articles"; Lindsay McKenzie, "Want to Analyze Millions of Scientific Papers All at Once? Here's the Best Way to Do It," Science AAAS website, July 21, 2017, https://www.sciencemag.org/news/2017/07/want-analyze-millions-scientific-papers-all-once-here-s-best-way-do-it.

researchers include PDFX,[11] GROBID,[12] ParsCit,[13] PDF-extract,[14] LA-PDFText,[15] Biblo (for citations, humanities-focused),[16] and CERMINE.[17] We chose to use CERMINE for the PDF extractions in this study. Our reasons for doing so are detailed in the methodology below.

While the majority of art history journals are not available in formats other than PDF, JSTOR—which contains a large number of humanities publications—has a service that provides "Data for Research" or DfR.[18] So studies that skip the complex step of PDF extraction are possible, given that JSTOR has the articles a researcher would like to study.[19] This is another area that can be developed across journal repositories to aid humanities scholars in understanding the literature in their disciplines.

Once the journal articles are in a format that is ready to process, the researcher must select the methodology for text mining.

[11] Alexandru Constantin, Steve Pettifer, and Andrei Voronkov, "PDFX: Fully-Automated PDF-to-XML Conversion of Scientific Literature," in *Proceedings of the 2013 ACM Symposium on Document Engineering*, DocEng '13 (New York: ACM, 2013), 177–180.

[12] Patrice Lopez, "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications," in *International Conference on Theory and Practice of Digital Libraries* (Cham: Springer, 2009), 473–474.

[13] Isaac G. Councill, C. Lee Giles, and Min-Yen Kan, "ParsCit: An Open-Source CRF Reference String Parsing Package," in *LREC* 8 (2008): 661–667.

[14] The PDF-extract project has been "retired" and the creators now recommend using CERMINE. See "Pdfextract," CrossRef website, https://www.crossref.org/labs/pdfextract.

[15] Cartic Ramakrishnan et al., "Layout-Aware Text Extraction from Full-Text PDF of Scientific Articles," *Source Code for Biology and Medicine* 7 (May 28, 2012): 7.

[16] Young-Min Kim et al., "Automatic Annotation of Bibliographical References in Digital Humanities Books, Articles and Blogs," in *Proceedings of the 4th ACM Workshop on Online Books, Complementary Social Media and Crowdsourcing* (New York: ACM, 2011), 41–48.

[17] Dominika Tkaczyk et al., "CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature," *International Journal on Document Analysis and Recognition (IJDAR)* 18, no. 4 (December 1, 2015): 317–335.

[18] "JSTOR Data for Research," JSTOR website, https://www.jstor.org/dfr.

[19] See, e.g., Riddell, "How to Read 22,198 Journal Articles," 91–114.

For this study, we chose to use R (including the tm [text mining] package) because it offered flexibility and customizability that other methods did not offer. There are also text mining tools such as Voyant that do not require programming experience and can be operated from the browser window online, but these do not allow for the level of functionality that programmatic means such as R allow.[20] For example, one important step in text mining is cleaning the text of noise—punctuation, numbers, *et cetera*—which is easily and efficiently done in R. These cleaning steps can be customized according to the dataset at hand. It is also possible to analyze and extract particular segments of the data as well as create a wide variety of visualizations as well as export the data in a custom format.

Methodology

To prepare sample sets of articles to test, the PDFs must be downloaded from online scholarly databases. Attaining permission to bulk download PDFs can require lengthy permission negotiations between publishers, university libraries, and researchers. In assembling the corpuses for this study, therefore, 373 articles from the journal *Art History* and 215 from *Art Journal* from the last decade were assembled by manually downloading the articles from the Wiley and Taylor & Francis portals, respectively.[21] This is a relatively small set of data compared to some of the studies that have been done on academic articles previously, but nevertheless large enough to track significant trends within a journal.[22] It should be noted, however, that it would not be ideal to

[20] Voyant Tools website, "Voyant: See Through your Text," https://voyant-tools.org.

[21] While every page/section from these journals can be downloaded as PDFs, the text corpus we produced was limited to the editorial letters and articles in these journals. For *Art Journal*, six types of texts or PDFs were not included: the title page, information on the editorial board, table of contents, funding information, reviews and artist projects. For *Art History*, two types of texts were not included: abstracts/authors' biographies and reviews.

[22] See, e.g., Westergaard et al., "Text Mining of 15 Million Full-Text Scientific Articles"; McKenzie, "Want to Analyze Millions of Scientific Papers All at Once?"

manually download articles if the sample were much larger than this because it would be too time-consuming.

Once the articles have been downloaded, Adobe Acrobat has several export options for PDFs including plain text, rich text, and XML. Using the two corpuses of articles from *Art History* and *Art Journal*, it was clear that these are inadequate methods by which to export text from journal articles. Straight export in Acrobat preserves database headings, journal volume/issue headings, and other information that is repeated on every page of the article, which can skew text mining results. Furthermore, the XML is not exported with reliable individual tags to clearly denote the separate sections of the article. It is for this reason that a number of research groups have developed protocols, such as CERMINE and the others cited above, that are specifically designed to capture the data from scientific journal articles. In particular, these methods are useful for demarcating headings, body text, and bibliographic information, as well as recognizing columns and blocks of text.

Converting PDF to text and XML with CERMINE

In order to convert the sample of PDFs to text and XML, we used an open-source Java protocol called CERMINE (Content ExtRactor and MINEr).[23] Compared to other PDF to XML protocols, CERMINE is able to capture specific features of bibliographic information, including author, year, volume, *et cetera*.[24] Applying CERMINE to the two sample corpuses produced much more reliable and useful text/XML from PDFs than could be achieved by exporting full text from PDFs in Acrobat.[25] Since it is one of the newer protocols in this area, it has addressed some of

23 Tkaczyk et al., "CERMINE."

24 Tkaczyk et al., "CERMINE," 319.

25 It should be noted that one study, by the researchers who created CERMINE, found that GROBID outperformed CERMINE "out of the box" for citations. Nevertheless, we found CERMINE more straightforward to use as non-experts in computer science. See Dominika Tkaczyk et al., "Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers," in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18 (New York: ACM, 2018), 99–108.

the issues and omissions in the design of other similar protocols.[26] However, CERMINE works best for articles that have a reference list at the end of the article rather than just a "Notes" section (either endnotes or footnotes), as is common in art history. When faced with notes that contain more than just citation information, that is, commentary or added information rather than simply bibliographic sources, the XML tags for article title, authors' given names and surname, year, volume, issue, *et cetera*, are sometimes garbled. Nevertheless, CERMINE was able to reliably capture the surnames of authors from the endnotes of 373 articles in *Art History* when they contained references rather than commentary.

Since notes in *Art History* appear at the end of the article, CERMINE was effective at demarcating the bibliography section of the article accurately. *Art Journal*, however, provided a number of issues that had to do with the layout of the journal. In *Art Journal*, notes appear along the left side of the page next to where they are cited. In Figure 1, which was taken as a sample from the summer 2005 issue of *Art Journal*, the notes are highlighted in dark gray for both the first page and an interior page of an article. This format has been used by *Art Journal* since summer 1998, so a corpus of articles from the last 10 years all have this format. Since all the notes are along the margin on the left side of the page, we were able to solve this formatting issue by batch cropping our corpus of 215 documents from *Art Journal* in Acrobat to create a corpus of PDFs that only contain the left margin of the page. This corpus of left margins could then be run through CERMINE to export plain text and XML from the PDFs. Likewise, the reverse operation was executed in which just the notes were cropped off the page to leave the main body text of the article. In cases such as this, where formatting is unusual but consistent throughout, this work-around produces satisfactory results.

If a larger sample of the journal from the 1960s to the present had been used, however, we would have had to contend with a number of formatting changes that would have complicated the operation. Figure 2 shows the format of the journal from autumn 1965, for example, with the notes highlighted in gray for the first page and several interior pages. Here, notes appear under only the

[26] Tkaczyk et al., "CERMINE," 318–319.

Figure 1: A sample page from the summer 2005 issue of *Art Journal* (footnotes in dark gray). Copyright: Authors. License: CC BY 4.0.

Figure 2: A sample page from the autumn 1965 issue of *Art Journal* (footnotes in dark gray). Copyright: Authors. License: CC BY 4.0.

second column alone, only the first column, under both columns, or under neither column. It would be very difficult to reliably crop just the notes from these pages without including any of the body text of the article. Although CERMINE *does* recognize references that are interspersed in the text, it places them in situ in the XML (i.e., among the <p> tags), so references may be garbled in exporting the information rather than cleanly demarcated, as they are when they appear at the end of the article. These differences in the layout mean that, even though the text mining itself can be automated, every volume that goes into the dataset has to be analyzed separately (based on the particularities of its layout) in order to get appropriate and reliable text extractions.

Moving through the years that the journal has been published, one would have had to contend with a number of additional changes in format. An example from winter 1975–76 (Figure 3) shows a two-column format with notes at the end. CERMINE is relatively good at handling columns in articles and has checks in place for determining the layout of the columns in the article.[27] This is useful to note, as *Art Journal* had a three-column format in an example from spring 1985 (Figure 4), and notes were placed at the end of the article. By winter 1995, the journal had returned to a two-column format (Figure 5). In 1998, as noted, the journal transitioned to its current format, which places the references at the left margin of the page. It is, therefore, important to bear in mind that the unique style formats of humanities journals vary not only in comparison to each other but also in comparison to their earlier or later manifestations.

Analyzing the plain text

In the given sample of art history journal articles, we used CERMINE to export both plain text and XML. Although the plain text files created by CERMINE exclude many (if not all) of the abovementioned auxiliary PDF headers that were left in the text files exported from Adobe Acrobat, they still contain a lot of text, which creates unwanted noise in the final text mining operations. For example, copyright information for images and figure tags may be undesirable in a corpus collected for text mining since particular image collections and repositories that provide art images for publication will repeatedly occur. In the case of *Art History*, the headings for the publication itself were successfully removed, but, owing to the unusual format of *Art Journal*, the plain text files created by CERMINE for this journal (before page cropping) still contain unwanted headings such as the issue (e.g., "Winter 2011"). Given these limitations, extracting particular

[27] Tkaczyk et al., "CERMINE," 320–321.


The Artistic Collaboration of Pablo Picasso and Julio Gonzalez

Figure 3: A sample page from the winter 1975–76 issue of *Art Journal* (endnotes in dark gray). Copyright: Authors. License: CC BY 4.0.


Edouard Manet and "Civil War"

Figure 4: A sample page from the spring 1985 issue of *Art Journal* (endnotes in dark gray). Copyright: Authors. License: CC BY 4.0.

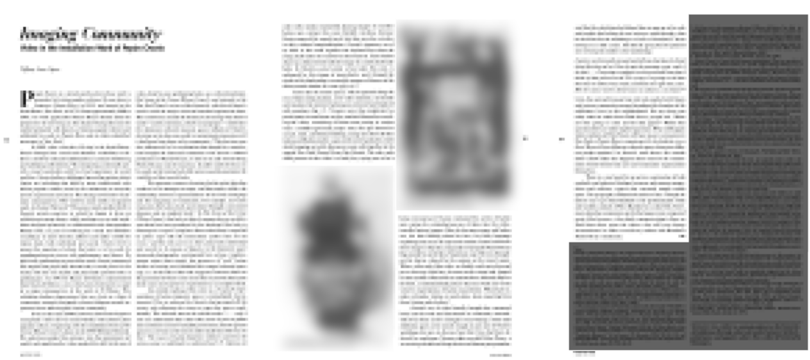


Figure 5: A sample page from the winter 1995 issue of *Art Journal* (endnotes in dark gray). Copyright: Authors. License: CC BY 4.0.

sections from XML files, used alone or in combination with each other, provides more flexibility and clarity in the data.

Extracting XML tags with Python

Certain XML tags in the documents created by CERMINE were of particular interest for text mining the journals in this study. Pulling the <p> (paragraph) tags generates a text document that contains only the body text of a given article minus headings, bibliography, and any other information in the article. This can be used to look at term frequency and create word clouds that capture the main themes of the journal. Additionally, pulling all the text contained in the subchildren of the <back> tag creates a text output of the entire bibliography of the article. This is useful in determining the most frequently cited authors. The <surname> tag was also pulled from the XML in order to compare frequently cited authors to the full bibliography text. While all of the surnames were not accurately captured in the <surname> tag, as mentioned, pulling just this tag creates a very clean list of authors to work with and provides a good basis of comparison to the results of the full bibliography, which need to be cleansed of nonauthor names to a much larger extent before text mining.

In order to pull XML tags, we created two different scripts using Python that could pull the three tags listed above: <p>, <back> (plus children/subchildren), and <surname>. One of the scripts takes the given corpus of journal articles and compiles the selected tag from all those articles into one single plain text file. The other script creates a separate text file with the data from the given tag for each of the journal articles. These were later used to determine overall frequency of terms or authors from a given corpus versus frequency of term or author across the documents. For example, Author X might be a very important reference for three articles out of 10. If these three articles reference her 10 times each or a total of 30 times, she would be placed high on the list of frequently cited scholars in the corpus of 10 documents. If Author Y, however, is mentioned in all 10 of the articles but only one time for each, her frequency of total citations would be lower than Author X (a total of 10). Given this, frequency alone is not the only indicator of influence or popularity. One could argue that Author X is more influential, as she is mentioned in all the articles, even if she is not mentioned *as much*. Therefore, this methodology looks at *both* total frequency and frequency across the documents. So, Author X would be counted for three documents and Author Y for 10, putting Author Y higher on the list of most-cited authors across all the articles.

Text mining with R

Once the desired data from particular XML tags have been pulled from the files, preparations for text mining can begin. For this purpose, we used the tm (text mining) package in R. The first step after loading the given corpus of text files was to clean the data. Some of the text exported from the PDF did not come out as clean as we would have liked. For example, some words contained spaces between vowels, such as the word "figure," which sometimes appeared in the XML as "fi gure." In most cases, this kind of export error was corrected *ad hoc* with find/replace within the text. Some words were also mashed together without spaces in the export. These would later be combined with instances of the words that came out cleanly in the frequency lists when they were found in the resultant CSV file.

In order to clean the text, we used several different cleaning protocols in R/tm that are designed to remove unwanted characters and words. Often, within a given corpus, there are combinations of words that have to be excluded from the bibliography, as they could skew the results. For example, "Journal of the Warburg and Courtauld Institutes" and "Warburg Institute" were removed from bibliographic references in this study, as they would have distorted the frequency for citations of the art historian Aby Warburg, after whom the institute takes its name. Another facet of the language found in art history journals is that there are plenty of non-ASCII characters from non-English languages, such as accented characters. In order to simplify the export of these characters in the given data, we converted accented characters to ASCII equivalents. So, for example, ö became o. After attending to these special cases, which depend on the contents of the corpus at hand, we followed a number of common steps for text cleaning. These included: removing punctuation and numbers, making text lowercase, removing stopwords, removing whitespace, and stemming the text in order to combine words with similar roots such as "paintings" and "painting."

In the case of looking at bibliography text, choice of stopwords became an issue. Since we were only interested primarily in the names found in the references, the standard English stopword list was not comprehensive enough. Stopwords are the words that are common to a given language but not interesting or important for text mining, for example "the," "and," or "for." R has its own stopwords function and lists where the language (in this case, English) can be specified to remove these commonly used but unimportant words from the frequency output. To analyze the bibliography, however, we created our own custom stopwords lists of the top 20,000 most frequently used words in the English language.[28] Applying these lists to the text gave a much cleaner picture of the authors cited. After cleaning the text, the next step was to create a document term matrix (dtm) from the corpus and export a CSV spreadsheet of the terms in decreasing order of frequency. This produces the total frequency for the single combined text file of the given corpus of documents.

In order to find the spread of a given term or author across the documents, however, a slightly different method was used. For this, another corpus was created from a directory containing all the individual text files of each exported tag and then cleaned using the same steps as above. Once the text was cleaned, it was exported as a term document matrix (tdm) and written to a CSV file. Each column was then counted if there was a value > 0 (indicating the author or term had been cited at least once) and those totals were used to create a list of the top-cited authors or terms by number of documents in which they appear.

Findings and Discussion

This study found that CERMINE can be used on art history journal articles to export body text and bibliographic information in a parseable XML format, which can then be used to perform text

[28] This list was compiled from preexisting word lists that are readily available online. The inbuilt stopwords exclusion function in R, however, has a much smaller list of words such as those listed above.

mining in R (tm package) to determine the frequency of citations and terms for a given set of journal articles. While the <surname> tags exported in CERMINE were less reliable than extraction of either the full bibliography or the <p> tags, it still provided a useful, clean set of data to work with in combination with the other extractions. The text mining protocol in R is well established and can be used as a flexible means for a variety of text mining functions beyond simple frequency of terms. Before data can be put into this tried and tested text mining method, the input data needs to be relatively clean. This study shows that there is a need for a protocol like CERMINE that is specifically designed for footnotes and endnotes that include both bibliographic information and commentary for the study of art history and other humanities journal articles. The highly varied formats of such journals, however, make the task very difficult. Text mining of frequently cited authors and terms in journal articles can then be used to provide an overall picture of the themes and biases of a particular journal over a given time period.

The overall goal of dealing with a large corpus of text is to automate as many steps as possible, so the steps of this methodology that had to be conducted manually point toward areas for further development when it comes to text mining humanities PDFs.[29] The first such area is in downloading the original corpus of PDFs. This is easily solved for a larger corpus, as some journal databases are now willing to allow researchers access to articles for text mining studies.[30] Given some of the inadequacies in the quality of text exported and the manual text cleaning that was necessary to correct errant spaces and words that were mashed together in the converted text produced by CERMINE, this study would have certainly benefited from delivery of text in plain text format. When PDF text—even OCR or born-digital articles—are captured from the document, the quirks of spacing do not always translate, as anyone who has ever tried to copy/paste text from a

[29] There are existing out-of-the-box software packages to do text analysis like NVivo, but acquiring the actual *text* of journal articles in a usable and machine-readable form remains the main barrier here.

[30] This varies by database and country where the researcher is based and is by no means a resolved issue for research such as this.

PDF will have learned. If journal articles were available as structured text or even just plain text, these issues would be irrelevant.

There were also issues that arose within the final frequency lists of terms and authors. One of the major issues encountered was that, in focusing on surnames, the output of common surnames such as “Smith” and “Jones” were conflated with one another. In further studies, a means by which specific Smiths are associated with their given name or specific references would help to solve this issue. In the literature, an author named Smith might be referred to by surname several times, with their given name only mentioned in the first citation. This makes connecting all the *same* Smiths more difficult. Another major issue in the text cleaning process was that stemming was not always very successful. The particular jargon of art history, that is, the nominalization and adjectivization of certain words and authors, is not accounted for in standard stemming protocols. For example, the name of the art historian Aby Warburg is sometimes adjectivized to the descriptor “Warburgian” or the verb “perform” may be nominalized as “performativity,” following scholar Judith Butler.[31] In other words, one of complications in dealing with text mining of art history scholarship is the complexity of the language used and its deviation from standard English. The nominalized or adjectival forms of certain terms are, however, rather easy to locate through searching for the root in the CSV output files.

Despite these issues, the methodology described in this study produced results that were rich with insight and often defied expectations or preconceptions we had about particular journals or the concerns of the field in general. The frequency lists of terms and authors can be used in the future to compare and analyze differences and similarities across sample sets.[32] Possible avenues of further research using this methodology include comparisons of the content in a selection of different journals or analysis of the content in one particular journal over time.

---

31 Judith Butler, *Bodies That Matter: On the Discursive Limits of “Sex”* (New York: Routledge, 1993); Judith Butler, *Gender Trouble: Feminism and the Subversion of Identity* (New York: Routledge, 1999).

32 For this aim we will make these datasets open access on our research project’s website metadataculture.se in the future.

Acknowledgments

The research for this chapter was conducted within the project *Sharing the Visual Heritage* (metadataculture.se) at Stockholm University, funded by the Swedish Research Council.

References

Atanassova, Iana, Marc Bertin, and Philipp Mayr. "Mining Scientific Papers for Bibliometrics: A (Very) Brief Survey of Methods and Tools." Paper presented at the 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 2015. arXiv: Preprint Archive:1505.01393. https://arxiv.org/abs/1505.01393v1.

Block, Sharon, and David Newman. "What, Where, When, and Sometimes Why: Data Mining Two Decades of Women's History Abstracts." *Journal of Women's History* 23, no. 1 (March 10, 2011): 81–109.

Butler, Judith. *Bodies That Matter: On the Discursive Limits of "Sex."* New York: Routledge, 1993.

Butler, Judith. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge, 1999.

Colavizza, Giovanni, and Matteo Romanello. "Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead." *Journal of European Periodical Studies* 4, no. 1 (June 30, 2019): 36–53.

Constantin, Alexandru, Steve Pettifer, and Andrei Voronkov. "PDFX: Fully-Automated PDF-to-XML Conversion of Scientific Literature." In *Proceedings of the 2013 ACM Symposium on Document Engineering*, 177–180. DocEng '13. New York: ACM, 2013.

Councill, Isaac G., C. Lee Giles, and Min-Yen Kan. "ParsCit: An Open-Source CRF Reference String Parsing Package." *LREC* 8 (2008): 661–667.

Crasto, Chiquito J., Luis N. Marenco, Michele Migliore, Buqing Mao, Prakash M. Nadkarni, Perry Miller, and Gordon M. Shepherd. "Text Mining Neuroscience Journal Articles to Populate Neuroscience Databases." *Neuroinformatics* 1, no. 3 (September 1, 2003): 215–237.

Dahl, Stephan. "Current Themes in Social Marketing Research: Text-Mining the Past Five Years." *Social Marketing Quarterly* 16, no. 2 (June 1, 2010): 128–136.

Davis, Allan Peter, Thomas C. Wiegers, Phoebe M. Roberts, Benjamin L. King, Jean M. Lay, Kelley Lennon-Hopkins, Daniela Sciaky, et al. "A CTD–Pfizer Collaboration: Manual Curation of 88 000 Scientific Articles Text Mined for Drug–Disease and Drug–Phenotype Interactions." *Database: The Journal of Biological Databases and Curation* (2013). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3842776.

Goldstone, Andrew, and Ted Underwood. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45, no. 3 (November 7, 2014): 359–384.

Günther, Elisabeth, and Emese Domahidi. "What Communication Scholars Write About: An Analysis of 80 Years of Research in High-Impact Journals." *International Journal of Communication* 11 (July 26, 2017): 21.

International Journal for Digital Art History website. https://dahj.org.

Jaskot, Paul B. "Digital Art History as the Social History of Art: Towards the Disciplinary Relevance of Digital Methods." *Visual Resources* 35, no. 1–2 (April 3, 2019): 21–33.

Jayasekara, P. K., and Abu K. S. "Text Mining of Highly Cited Publications in Data Mining." In *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)*, 128–130. Red Hook, NY: Curran Associates inc., 2018.

JSTOR website. "JSTOR Data for Research." https://www.jstor.org/dfr.

Kim, Young-Min, Patrice Bellot, Elodie Faath, and Marin Dacos. "Automatic Annotation of Bibliographical References in Digital Humanities Books, Articles and Blogs." In *Proceedings of the 4th ACM Workshop on Online Books, Complementary Social Media and Crowdsourcing*, 41–48. New York: ACM, 2011.

Long, Matthew P., and Roger C. Schonfeld. "Preparing for the Future of Research Services for Art History: Recommendations from the Ithaka S+R Report." *Art Documentation: Journal of the Art Libraries Society of North America* 33, no. 2 (2014): 192–205.

Lopez, Patrice. "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications." In *International Conference on Theory and Practice of Digital Libraries*, 473–474. Cham: Springer, 2009.

McKenzie, Lindsay. "Want to Analyze Millions of Scientific Papers All at Once? Here's the Best Way to Do It." Science AAAS website, July 21, 2017. https://www.sciencemag.org/news/2017/07/want-analyze-millions-scientific-papers-all-once-here-s-best-way-do-it.

Miller, Benjamin M. "The Making of Knowledge-Makers in Composition: A Distant Reading of Dissertations." Graduate Center, CUNY, 2015.

Mimno, David. "Computational Historiography: Data Mining in a Century of Classics Journals." *Journal on Computing and Culteral Heritage* 5, no. 1 (April 2012): 3:1–3:19.

Moxey, Keith. "Visual Studies and the Iconic Turn." *Journal of Visual Culture* 7, no. 2 (2008): 131–146.

CrossRef website. "Pdfextract." https://www.crossref.org/labs/pdfextract.

Peng, Tai-Quan, Lun Zhang, Zhi-Jin Zhong, and Jonathan J. H. Zhu. "Mapping the Landscape of Internet Studies: Text Mining of Social Science Journal Articles 2000–2009." *New Media & Society* 15, no. 5 (August 1, 2013): 644–664.

Ramakrishnan, Cartic, Abhishek Patnia, Eduard Hovy, and Gully Burns. "Layout-Aware Text Extraction from Full-Text PDF of Scientific Articles." *Source Code for Biology and Medicine* 7 (May 28, 2012): 7.

Riddell, Allen Beye. "How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models." In *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, edited by Matt Erlin and Lynne Tatlock, 91–114. Rochester, NY: Camden House, 2014.

Tkaczyk, Dominika, Andrew Collins, Paraic Sheridan, and Joeran Beel. "Machine Learning vs. Rules and Out-of-the-Box vs.

Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers." In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 99–108. JCDL '18. New York: ACM, 2018.

Tkaczyk, Dominika, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. "CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature." *International Journal on Document Analysis and Recognition (IJDAR)* 18, no. 4 (December 1, 2015): 317–335.

Voyant Tools website. "Voyant: See Through your Text." https://voyant-tools.org.

Wehrheim, Lino. "Economic History Goes Digital: Topic Modeling the Journal of Economic History." *Cliometrica* 13, no. 1 (January 1, 2019): 83–125.

Westergaard, David, Hans-Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, and Søren Brunak. "Text Mining of 15 Million Full-Text Scientific Articles." BioRxiv: The Preprint Server for Biology, July 11, 2017. https://doi.org/10.1101/162099.

Whalen, Maureen. "What's Wrong with This Picture? An Examination of Art Historians' Attitudes About Electronic Publishing Opportunities and the Consequences of Their Continuing Love Affair with Print." *Art Documentation: Journal of the Art Libraries Society of North America* 28, no. 2 (Fall 2009): 13–22.

Yang, Yang, Ching Man Au Yeung, Mark J. Weal, and Hugh Davis. "The Researcher Social Network: A Social Network Based on Metadata of Scientific Publications." *Proceedings of WebSci '09: Society On-Line*, Athens, Greece. Mars 18–20, 2009. https://eprints.soton.ac.uk/267156.

About the Authors

Jonas Andersson Schwarz is an associate professor in media and communications, Södertörn University (Stockholm, Sweden), whose primary research interest lies in epistemological and axiological aspects of digital media infrastructure. Coming from a background in file-sharing and piracy studies, he has throughout the last decade primarily studied digital platforms and attendant challenges related to governance and policy, particularly in terms of their role for civil society, knowledge-making, and the data-driven media economy.
ORCID: https://orcid.org/0000-0002-7353-4172

Teresa Cerratto Pargman is an associate professor of human–machine interaction at the Department of Systems and Computer Sciences (DSV), Stockholm University. She holds a PhD in cognitive psychology granted by Université Paris 8, France. Teresa's research work seeks to explain the interplay between material (technological), semiotic, and cultural aspects that are constitutive of contemporary sociotechnical practices, particularly those emerging in the educational sector. She coordinates the research area "Technology-Enhanced Learning" at DSV.
ORCID: https://orcid.org/0000-0001-6389-0467

Anna Dahlgren is professor of art history at Stockholm University. She is the principal investigator of the project *Sharing the Visual Heritage* (Swedish Research Council, 2019–23), which focuses on visual cultural heritage and digital interfaces. During 2019–21 she is coordinating the Swedish Research Council's funding program DIGARV. She has written extensively on different aspects of photography and visual culture including print culture, historiography, the digital turn, archives, and museum practices. She is the author of *Travelling Images: Looking Across the Borderlands of Art, Media and Visual Culture* (2018).
ORCID: https://orcid.org/0000-0001-7772-273

Björn Ekström is a PhD candidate in library and information science at the University of Borås, writing a dissertation on information practices in citizen science projects. His research interests include information practices, scholarly communication, and digital methods.
ORCID: https://orcid.org/0000-0003-4187-7004

Uno Fors is professor of computer and systems sciences, and head of department for Computer and Systems Sciences at Stockholm University. Fors is also convener of the digital human sciences committee at Stockholm University. Fors's research focuses on digital health, technology-enhanced learning, and ethical and privacy issues of digital systems.
ORCID: https://orcid.org/0000-0002-3166-1640

Stanley Greenstein (LLD) is an assistant professor in law and information technology at the Department of Law, Stockholm University. His main area of interest is the interaction between technology and society. In this regard, his teaching, research, and practical participation in project work has centered on the topic of artificial intelligence (AI) and its ethical and societal implications, of which privacy, data protection, and information security considerations are central.
ORCID: https://orcid.org/0000-0002-0694-768X

David Gunnarsson Lorentzen is a senior lecturer and has a PhD in library and information science from the University of Borås, focusing on digital methods and social media. His research interests also cover machine learning and algorithms in relation to man and society.
ORCID: https://orcid.org/0000-0003-0659-4754

Johan Jarlbrink is an associate professor in media history and a senior lecturer in media studies, Umeå University. His research has mainly focused on the history of journalism, media technologies, and information management. He is currently involved in the project *Welfare State Analytics: Text Mining and Modeling Swedish Politics, Media & Culture, 1945–1989*.
ORCID: https://orcid.org/0000-0002-1167-046x

Christer Johansson (1965–2020) is a research officer and professor at the Department of Culture and Aesthetics, Stockholm University. He received his PhD in literary studies (2008) and is currently working on the research projects *Concepts of Mediality: A Metatheoretical Investigation of Intermedial Studies* and *The Materialized Word: Medium and Meaning in Eyvind Johnson's Prose Fiction*.
ORCID: https://orcid.org/0000-0002-5919-9902

Cecilia Magnusson Sjöberg is professor of law and informatics at Stockholm University. Her research areas include IT law, privacy and information security, and legal system design and management. In addition to research projects nationally and internationally addressing the interplay between law and modern technologies, she is engaged by the Swedish government chairing and participating in public inquiries about, for example, legal aspects of e-government and artificial intelligence. Magnusson Sjöberg's research activities also include a wide variety of EU projects in which she contributes to legal and ethical advisory boards.
ORCID: https://orcid.org/0000-0002-6795-9510

Cormac McGrath is an associate professor of education at the Department of Education (IPD), Stockholm University. He holds a PhD in education from Karolinska Institutet, Stockholm, Sweden. Cormac's research interests extend to educational design and change management in higher education settings. Moreover, Cormac is interested in the impact of emerging technologies on student learning as well as university teachers' and institutions' practices in higher education.
ORCID: https://orcid.org/0000-0002-8215-3646

Julia Pennlert is a senior lecturer in library and information science at the University of Borås and has a PhD in comparative literature from the University of Umeå, focusing on reading and reading promotion activities. Her research interest covers reading in the contemporary media landscape, digital literature, and digital humanities. She is also the editor of a Swedish research anthology on digital humanities (2017).
ORCID: https://orcid.org/0000-0003-3110-1376

Sonya Petersson is an assistant professor in art history and visual studies at the Department of Culture and Aesthetics, Stockholm University, and coordinator of the digital human science committee at Stockholm University. She received her PhD in art history 2014 and has since then conducted the postdoctoral project *Graphic Illustration: Concepts and Combined Mediality from the Point of View of Mechanical Reproduction* (2016–19). Her research interests include text and image studies, intermedial studies, and analog/digital image reproduction.
ORCID: https://orcid.org/0000-0003-1335-1080

Jonas Stier is a professor of social work at Mälardalen University, Sweden. From an intercultural and social psychological stance, his research mainly revolves around identity, intercultural interaction, and diversity in arenas such as preschools, schools, and social services. Stier has a long experience of research and education collaboration with civil society, government, and industry. He is also one of the founders of the addAi.org—a policy initiative and an arena for discussions on artificial intelligence.
ORCID: https://orcid.org/0000-0001-7832-2155

Karolina Uggla is a postdoctoral researcher and assistant professor in information design and part of the Information Design Research Group at the School of Innovation, Design and Engineering at Mälardalen University, Eskilstuna, Sweden. She holds a PhD (2015) in art history from Stockholm University and has a long-standing interest in intersections between art, visual culture, and science. Uggla's current research interests are visual management: design thinking methods for organizational change, visualization of time in production planning, and digital timelines of history in museum education.
ORCID: https://orcid.org/0000-0003-1676-4366

Amanda Wasielewski is a postdoctoral fellow in art history at Stockholm University. She is a participant in the research project *Sharing the Visual Heritage*, which focuses on visual cultural heritage and digital interfaces. She is the author of *Made in Brooklyn:*

Artists, Hipsters, Makers, Gentrifiers (2018) and has taught social media and internet studies at the University of Amsterdam, architectural history at the Spitzer School of Architecture, and modern art history at Lehman College in New York.
ORCID: https://orcid.org/0000-0002-3034

Index

D

G

H

I

T

U

V

The ongoing digitization of culture and society and the ongoing production of new digital objects in culture and society require new ways of investigation, new theoretical avenues, and new multidisciplinary frameworks. In order to meet these requirements, this collection of eleven studies digs into questions concerning, for example: the epistemology of data produced and shared on social media platforms; the need of new legal concepts that regulate the increasing use of artificial intelligence in society; and the need of combinatory methods to research new media objects such as podcasts, web art, and online journals in relation to their historical, social, institutional, and political effects and contexts.

The studies in this book introduce the new research field "digital human sciences," which include the humanities, the social sciences, and law. From their different disciplinary outlooks, the authors share the aim of discussing and developing methods and approaches for investigating digital society, digital culture, and digital media objects.

ISBN 978-91-7635-147-5

9 789176 351475

STOCKHOLM
UNIVERSITY PRESS