

Article

Advancing Historical Research Through AI and Data-Centric Approaches

Wolfgang Thomas Göderle ^{1,2,3,*} , Malte Rehbein ^{2,3,†}  and Markus Gerstmeier ^{2,†} ¹ Department of Digital Humanities, University of Graz, 8010 Graz, Austria² Department of Computational Humanities, University of Passau, 94032 Passau, Germany; malte.rehbein@uni-passau.de (M.R.)³ Department of Coevolution of Land Use and Urbanisation, Max Planck Institute of Geoanthropology, 07745 Jena, Germany

* Correspondence: wolfgang.goederle@uni-graz.at

† These authors contributed equally to this work.

Abstract

The rapid digitization of large source collections in the humanities over the last three decades has comprehensively transformed the discipline. The accessibility of primary sources has improved drastically, the pre-processing of research data has been revolutionized in some areas, and new transdisciplinary approaches have emerged and become possible. However, while digital and computational historians have produced extensive reflection on these developments, the theoretical grounding of this transformation has not been fully integrated into mainstream historical methodology: most critically, the concept of ‘information’, central to computer science and computational methods, has not yet been systematically received as a technical category within the discipline’s methodological canon. In this contribution, we employ a concept from Science and Technology Studies—Bruno Latour’s ‘circulating reference’—to analyze and render describable the processes of historical research within a digitized research environment. Through three case studies—AI-supported segmentation of Habsburg cadastral maps (1817–1861), computational analysis of the *Hof- und Staatsschematismus* (1702–1918), and the datafication of the Munich Special Court archive inventory (1933–1945, 1975–1977)—we demonstrate how and at which specific points historical research benefits from this framework, and what new insights it enables.

Keywords: digital source criticism; computational history; computer vision; digital hermeneutics; Historical Information Science



Academic Editor: Doris Gruber

Received: 13 February 2026

Revised: 2 June 2026

Accepted: 11 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Massive digitization efforts in the past two decades have fundamentally transformed the conditions of historical research. The increasing availability of machine learning for automating indexing tasks, particularly layout detection and OCR/HTR, as well as the widespread accessibility of generative AI, will continue to reshape research conditions (Eberle et al. 2024; Siebold and Valleriani 2022). Whether these technical innovations will transform research practices on the broadest level remains to be seen; the development of digital methods in digital and computational humanities and the discourse about Digital History have not yet led to their methodological innovations being translated into historical research at large (Hiltmann 2024). However, we argue that the most profound impact goes beyond mere accessibility of sources or computational power.

Digitization and AI are catalyzing an epistemological shift that reveals a fundamental tension between traditional conceptualizations of historical sources and emerging data-centric approaches. For over a century, historical methodology has conceptualized sources not as ‘information containers’ but through a rich vocabulary of epistemological concepts: in the German tradition as *Quellen*, distinguished into *Überreste* (unintentional remains) and *Tradition* (intentional testimonies) (Bernheim 1908; Droysen 1868); in the Anglo-American world as ‘evidence’ requiring authentication and interpretation (Collingwood 1946); and in French historiography as *documents* and *témoignages* demanding rigorous *critique externe* and *critique interne* (Langlois and Seignobos 1898). Notably absent from this vocabulary was any systematic conceptualization of sources as containing ‘information’ in a technical sense. Sources were understood as fragments requiring reconstruction, testimonies demanding evaluation, or traces needing interpretive contextualization, but not as information repositories awaiting extraction.

The emergence of ‘information’ as a category for understanding historical sources is a comparatively recent development, introduced by Manfred Thaller in the context of Historical Information Science (Thaller 2017). It has since appeared primarily in contexts of digitization and computational analysis, introducing a conceptual framework borrowed from information science rather than emerging from historical methodology itself. This shift is not merely terminological. When sources undergo digital preprocessing—through OCR, database structuring, georeferencing, or AI-powered analysis—they are transformed into ‘information objects.’ More significantly, these processes reveal a granular layer beneath the documentary surface: data. Digital tools, particularly AI-powered instruments, allow us to decompose what traditional methodology understood as holistic traces or testimonies into constituent data elements that can be manipulated, recombined, and analyzed at scale.

This transformation raises fundamental questions: What happens when we reconceptualize sources from traces requiring interpretive reconstruction to information awaiting extraction? What is lost, and what new analytical possibilities emerge, when the hermeneutical tradition meets the computational paradigm of data processing? Can computational methods be productively applied to historical research when the discipline’s theoretical foundations provide no systematic framework for conceptualizing information, data, or their relationship to historical knowledge (Hiltmann 2024)?

This article examines the implications of this conceptual shift, extensively discussed in specialized Digital History scholarship but not yet systematically integrated into mainstream historical methodology, for historical epistemology. We argue that historians are increasingly adopting computational methods without the theoretical grounding necessary to justify their epistemological claims. Drawing on the methodological traditions of German *Quellenkritik*, Anglo-American empiricism, and French critical approaches, we first demonstrate that, while digital and computational historians have produced rich reflection on the epistemological challenges of datafication, the concept of ‘information’ as a technical category has not yet been integrated into the methodological canon of the discipline at large, which continues to privilege selective, exemplary source work justified through hermeneutical interpretation. We then propose that frameworks from Science and Technology Studies, particularly Bruno Latour’s concept of ‘circulating reference,’ offer productive pathways for bridging the gap between traditional source criticism and computational analysis.

Our analysis proceeds in five parts. First, we examine how traditional historical methodology has conceptualized sources, tracing the gap between the discipline’s sophisticated hermeneutic apparatus and the technical concept of information as it operates in computational practice, and the methodological legitimacy historically accorded to selective source work. Second, we explore how Latour’s work on knowledge construction

might provide theoretical foundations for understanding the relationship between data, information, and historical knowledge. Third, fourth, and fifth, we demonstrate these theoretical considerations through three case studies: AI-supported segmentation of Habsburg cadastral maps from the long 19th century; computational analysis of the *Schematismus*, a comprehensive directory of officeholders spanning 1700–1918; and the datafication of the archive inventory of the National Socialist Munich Special Court. Through this investigation, we aim to contribute to what several scholars have called ‘digital hermeneutics’: an approach that maintains the critical rigor of traditional historical methodology while embracing the transformative potential of digital technologies (Fickers et al. 2022; Siebold and Valleriani 2022).

Methodologically, our approach is deliberately case-driven rather than pipeline-driven. Each of the three case studies applies the transformation-chain framework developed in Section 3 to a distinct source type—cartographic, textual-administrative, and archival-secondary—using different computational methods. This design is not incidental but constitutive of our argument: we contend that there is no single ‘correct’ digital methodology for historical research and that the epistemological configuration of each pipeline is necessarily shaped by the material properties of the source, the historical question, and the constraints of the computational apparatus. The case studies are therefore not parallel applications of a unified method but three independent tests of the framework’s analytical reach across heterogeneous materials.

2. History as Data Science? Information, Data, and the Epistemology of Historical Sources

The increasing application of computational methods to historical research raises a fundamental question: how should the discipline conceptualize ‘information,’ how does it relate to historical sources, and what distinguishes ‘data’ from ‘knowledge’? While digital historians have addressed these questions extensively in specialized contexts, mainstream historical methodology has not yet produced a systematic theoretical integration of *information* as a technical category in the sense in which information science, computer science, and the computational disciplines understand it. This is not for want of reflection on the digital transformation of historical scholarship; rather, it reflects deeper methodological commitments that have shaped historical practice for over a century, and the relative isolation in which the most relevant theoretical work has developed.

The past two decades have produced a substantial international literature on the digital transformation of historical research, spanning German-, English-, and French-language scholarship. Foundational English-language treatments such as Cohen and Rosenzweig’s *Digital History* (Cohen and Rosenzweig 2005) introduced a generation of historians to the conditions of digital scholarship; Milligan (2019) has charted the methodological consequences of source abundance; and Crymble (2021) has historicized the relationship between technology and historical practice. In the German-language tradition, Föhr (2019) has systematically reworked source criticism for digital objects, while in French scholarship, Rygiel (2017), Clavert and Noiret (2013), and the ongoing collaborative project by Clavert and Muller (2017–2026) have examined the transformation of historical practice under digital conditions. Major collaborative initiatives have consolidated this work into shared infrastructures: the *Ranke.2* teaching platform at the C²DH Luxembourg (Scagliola et al. 2018) provides a pedagogical framework for digital source criticism, while the German-language *Living Handbook ‘Digitale Quellenkritik’* (Deicke et al. 2024) offers a continually updated, community-driven reference work on the methodological challenges raised by digitized and born-digital sources. The current generation of textbooks that explicitly address digital and computational approaches—such as *Digital Humanities: Eine Einführung* (Jannidis et al.

2017), *Digital Humanities in den Geschichtswissenschaften* (Antenhofer et al. 2024), and *Digital History: Konzepte, Methoden und Kritiken* (Döring et al. 2022)—has further established the practical and methodological foundations for computational work with historical sources, articulating many of the epistemological challenges that arise from this transformation. Building on these contributions, the present article seeks to advance one specific aspect of the conversation: the theoretical integration of information and data as technical concepts into mainstream historical epistemology. This is a task these works variously identify as a desideratum but, given their broader scope, leave to more specialized investigations.

When we examine the foundational textbooks that have shaped generations of historians—such as *Grundlagen des Studiums der Geschichte* by Boshof, Düwell, and Kloft (Boshof et al. 1973), *Werkzeug des Historikers* by Ahasver von Brandt (von Brandt 2012), *The Pursuit of History* by John Tosh (Tosh 2021), *From Reliable Sources* by Howell and Prevenier (Howell and Prevenier 2001), or *The Historian's Craft* by Marc Bloch (Bloch 1953)—a consistent pattern emerges. While these works demonstrate sophisticated approaches to source criticism and interpretation, none develops a systematic theory of what constitutes ‘information’ in historical sources or how such information relates to the data from which it is constructed or the knowledge it produces. Even the *Historische Hilfswissenschaften*—such as palaeography, diplomatics, sigillography—which came closest to formalizing the analysis of source properties, operated with categories of authenticity, provenance, and material form rather than with information as a technical category. Traditional historical methodology treats ‘information’ as recoverable content within sources, requiring proper methodological application through *Quellenkritik* for extraction and interpretation. *Quellenkritik* provides sophisticated tools for establishing authenticity and context; hermeneutics guides interpretation; but neither tradition requires historians to theorize what ‘information’ itself might be in a technical or formal sense. Information is simply assumed to be directly accessible through established hermeneutic procedures.

This theoretical gap becomes striking in comparison with adjacent fields. In Science and Technology Studies, scholars like Bruno Latour have systematically theorized the construction and circulation of scientific knowledge. In information science, concepts like data granularity, knowledge representation, and the relationships between these categories have been elaborated into comprehensive theoretical frameworks. The history of science, too, has produced sustained reflection on the historicity and constructed nature of data itself: Strasser and Edwards (2017) have offered a particularly incisive analysis of how the contemporary celebration of ‘big data’ presupposes—but rarely interrogates—the question-driven character of any data collection, while the broader *Osiris* volume in which their essay appears places this insight into a long historiographical perspective. Within historical scholarship itself, Manfred Thaller’s pioneering work since the 1980s on ‘Historical Information Science’ explicitly theorized information structures in historical sources (Thaller 2017), and subsequent overviews have documented this subdiscipline’s development (Boonstra et al. 2004). However, these contributions remained confined to a specialized community and were never integrated into mainstream historical methodology. While the digital and computational humanities have produced rich theoretical reflection on source criticism, hermeneutics, and the materiality of digital objects, the conceptual triad of *data*, *information*, and *knowledge* as technical categories—categories with specific operational meanings in computational systems—has not yet been systematically integrated into the methodological canon of the historical discipline as a whole.

This integration gap reflects the discipline’s long-standing methodological privilege of selective source work, the epistemological legitimacy of drawing historical insights from carefully chosen examples without systematic assessment of their representativeness. Traditional methodology has operated with what might be called a ‘principle of

exemplary interpretation': the assumption that individual sources, properly contextualized through *Quellenkritik*, can yield valid insights without requiring assessment of their statistical representativeness.

This approach reflects the discipline's hermeneutical foundations. The German tradition explicitly distinguished between *verstehen* (understanding) and *erklären* (explaining), with the former requiring interpretive depth rather than quantitative breadth. Anglo-American historiographical tradition similarly privileged the historian's judgment in selecting sources over systematic sampling. French historiography, from the *école méthodique* through the Annales School, emphasized critical evaluation of documents but not the systematic theorization of information as such. Even Marc Bloch argued that historians must 'select and sort' facts as a necessary form of abstraction, without developing a systematic theory of how this selection relates to representativeness or information completeness.

This deficit produces a persistent methodological dichotomy between the micro-perspective—deep, context-rich, but statistically vulnerable—and the macro-perspective—broad, structural, but often detached from primary sources. We argue that the integration of AI and data-centric approaches provides the epistemological infrastructure to reconcile these perspectives. By subjecting massive corpora to computational analysis (referencing the 'Totum' rather than the 'Pars'), we no longer rely on the *pars pro toto* fallacy alone. Instead, data-centric history enables what Armaselu and Fickers, building on a wider conversation in Digital History, have termed a 'scalable hermeneutics' (Armaselu and Fickers 2024): the ability to toggle between the distant reading of structural patterns and the close reading of the individual anomaly. In this framework, the general truth is derived from the comprehensive data layer, which allows us to identify the specific case as either representative or exceptional with statistical confidence.

Even quantitatively oriented traditions engaged with representativeness only selectively. The Annales School's emphasis on serial sources and *longue durée* introduced quantitative methods but within specific research agendas rather than as a general methodological requirement. The cliometric revolution, pioneered by scholars like Robert Fogel in the 1960s, developed sophisticated statistical methods but remained a specialized subfield (Fogel and Engerman 1974). Its methodological innovations were never translated into broader historical methodology; most historians continued to work with selective sources justified through hermeneutical interpretation (Hudson 2000).

The historical discipline's reliance on selective, exemplary source work—justified through hermeneutical interpretation—has historically masked this theoretical lacuna. When reading a limited number of documents *pars pro toto*, the trained historian's tacit knowledge seamlessly processes textual traces into historical meaning without ever needing to define information as a formal category. The result is that mainstream historical methodology has produced a substantial body of reflection on narrative construction and historical explanation, but no canonical theory of information commensurate with computational practice. However, this implicit approach becomes unsustainable in the context of the digital transformation (Fickers 2020; Hiltmann 2024). As datafication transforms historical sources into structured datasets amenable to computational analysis at scale, the methodological privilege of selective source work is challenged. When cadastral records spanning entire territories or complete parish registers become computationally accessible in their totality, the methodological focus necessarily shifts from exemplary selection to statistical representativeness. It is precisely this shift in scale that forces an epistemological reckoning: Computational systems, whether databases or machine learning algorithms, cannot rely on hermeneutic intuition. They require explicit, formal definitions of what constitutes a data point, how metadata is structured, and how information is extracted and encoded. Thus, the scale of digital research does not merely change the *amount* of

sources we process; it exposes the lack of a theoretical framework for the very ‘data’ we are extracting, turning what was once a pragmatic necessity of selective reading into an urgent epistemological problem.

In attempting to bridge this gap, many digital historians have turned to theoretical frameworks from information science. The DIKW (Data–Information–Knowledge–Wisdom) hierarchy has proven particularly attractive, appearing frequently in digital humanities discourse as an apparently ready-made solution. Despite being what Rowley (2007) calls a ‘fundamental, widely recognized and taken-for-granted model,’ the DIKW hierarchy has been criticized as ‘strongly depend[ing] on a subjective viewpoint’ (Rehbein 2021, p. 161) or ‘unsound and methodologically undesirable’ (Frické 2009). Applied to historical sources, these problems become even more pronounced: sources resist hierarchical ordering, simultaneously embodying data points, conveying information, representing knowledge claims, and requiring wisdom for interpretation.

The uncritical adoption of such frameworks risks obscuring what makes historical sources distinctive: their fragmentary nature, their embeddedness in specific contexts, their resistance to standardization, and their demand for interpretive rather than extractive approaches. Yet the digital transformation also offers unprecedented opportunities. The challenge lies in developing theoretical frameworks that preserve the hermeneutical depth of traditional source criticism while embracing the analytical possibilities of computational methods. We argue that such frameworks exist but must be sought not in information science’s hierarchical models but in Science and Technology Studies’ more nuanced accounts of how data, inscriptions, and knowledge circulate and transform.

3. Bridging the Gap: Circulating Reference and Historical Sources

The integration gap identified in the previous section—the lack of a systematic theoretical framework that ties *information* as a technical category into mainstream historical methodology—cannot be filled by simply importing concepts from information science. The DIKW hierarchy, as we have seen, imposes a linear structure fundamentally at odds with the iterative, hermeneutical nature of historical interpretation (Drucker 2011). The digital humanities, however, have long worked on this problem from a different angle. Drucker’s distinction between *data* (the given) and *capta* (the taken) (Drucker 2011) has denaturalized the seeming objectivity of digital datasets; the program of *digital hermeneutics* developed at the C²DH and elsewhere (Fickers et al. 2022; Scagliola et al. 2018) has insisted that digital workflows demand the same critical scrutiny as any analog source; and Armaselu and Fickers’ notion of *scalable hermeneutics* (Armaselu and Fickers 2024) has begun to articulate how the move between micro and macro can be theorized as a hermeneutic operation in its own right.

Building on this rich conversation, we propose to add one specific analytical instrument: Bruno Latour’s concept of ‘circulating reference’ (Latour 1999a), deployed alongside Hans-Jörg Rheinberger’s account of *experimental systems* (Rheinberger 1997). Latour’s framework offers a fine-grained vocabulary for tracking what is lost and what is gained at each step of a transformation chain, and it foregrounds *bidirectional traceability* as an operational criterion for epistemically sound digital workflows. Rheinberger’s account of how ‘epistemic things’ and ‘technical objects’ co-evolve in the laboratory provides the complementary insight that our research questions and our computational instruments are not given in advance but emerge together through iterative practice. Used jointly, these two frameworks let us describe AI-augmented historical workflows at a granularity that more abstract critiques of datafication cannot reach.

We adopt these frameworks pragmatically, as heuristics for analyzing digitization processes, rather than endorsing the broader ontological commitments of either Actor–

Network Theory or Rheinberger's wider science studies program. Latour's ANT in particular has faced substantial critique regarding human agency (Spiegel 2007), epistemological relativism (Bloor 1999), and ahistoricism. The pragmatic deployment of *circulating reference* as an analytical heuristic, however, has a tradition of its own in historical research (Goederle 2015), and it is in this restricted sense that we use it here.

3.1. Latour's Circulating Reference: From Matter to Form

In his ethnographic study of soil scientists working in the Amazon forest of Boa Vista, Latour (1999b) traces how observations of physical phenomena—the forest–savanna transition—become scientific knowledge through a series of transformations.¹ His account begins not with data or information but with the material world itself: soil, plants, earthworms.

Latour concluded from his participant observation that—at each stage of the scientist's research work—something is lost and something is gained: When botanist Edileusa extracts plant specimens from the forest, she loses the living ecosystem but gains portable, preservable samples. When these specimens are dried, pressed, and labeled, they lose their three-dimensionality and color variation but gain stability and compatibility with taxonomic systems. When pedologist René takes soil samples using standardized protocols, he loses the continuous variation of the soil profile but gains discrete, comparable data points. When these samples are encoded using the Munsell color system, they lose their material presence but gain universal commensurability—any soil scientist anywhere can understand '10YR3/2.' This principle of simultaneous loss and gain through transformation has parallels in historical epistemology. Daston and Galison (2007) trace how mechanical reproduction technologies in 19th-century science—photography, lithography—similarly involved trade-offs between interpretive richness and reproducible standardization, which they term 'mechanical objectivity.' The transformation chains we observe in Digital History represent a contemporary instantiation of this ongoing tension between the singular and the mobile, the interpretive and the systematic.

Crucially, Latour emphasizes that reference is not a correspondence between words and world or between language and reality, but rather a property of the entire chain of transformations. Truth and accuracy depend not on eliminating mediations but on maintaining their traceability. The pedocomparator (the device that allows soil samples to be systematically compared and encoded) does not distort reality—it constructs a reality that can be known, manipulated, and communicated. Rather than questioning whether the world exists independently of representation, Latour's concept of circulating reference shifts the focus to the practical chains of translation and the series of mediations required to establish a reliable connection to reality Latour (1999b). This perspective directly challenges the representationalist epistemology underlying much traditional source criticism, which seeks to minimize mediation between historian and past. Latour's framework instead valorizes mediation as being constitutive of knowledge-making, a position that resonates closely with Rheinberger's account of how scientific knowledge is produced not by stripping away instruments and protocols but by building experimental systems in which objects become answerable to questions (Rheinberger 1997).

Reference circulates in both directions. The expedition can return 'upstream' from the final scientific diagram to the forest site because each transformation preserved essential features while making them mobile. The numbered tags nailed to trees, the coordinates inscribed in notebooks, the soil samples in plastic bags—these create a reversible path that allows movement back and forth between matter and form, between the forest and the laboratory, and between observation and inscription. This bidirectionality is crucial for historical research: unlike purely extractive data models, where 'raw data' flows unidirectionally into increasingly abstract representations, critiqued by Gitelman (2013),

Latour's chain preserves what we might call 'epistemic reversibility,' the possibility of return. Computational transformations remain valid only insofar as they maintain traceable connections back to source materials, allowing historians to move between aggregate patterns and individual documents.

3.2. *Tracing the Hybrid Chain: Transformations in a Blended Workflow*

Latour's conceptual framework is particularly valuable because it allows us to dissect the historical research process without falling into false dichotomies between 'analog' and 'digital' scholarship. Most contemporary research takes place in a blended setting, where traditional hermeneutic practices are inextricably interwoven with digital tools. Latour's model of the chain of reference enables us to pinpoint exactly *where* and *how* digital technologies intervene in this workflow.

Rather than treating digitization as a rupture, a division between 'traditional' and 'digital' history, Latour's framework reveals continuity in practices of transformation, inscription, and mobility. This perspective aligns with recent critiques of digital exceptionalism in the humanities (Hitchcock 2013; Putnam 2016), which emphasize that computational methods represent intensifications and accelerations of existing scholarly practices rather than fundamental breaks. As Putnam (2016) demonstrates, full-text database searching systematically biases research toward sources that were already privileged—digitized first, English-language, institutionally preserved—creating what she calls 'shadows' that reshape historical practice.

Yet this observation, while important, should not be mistaken for a uniquely digital phenomenon. Traditional historical practice has always been shaped by comparable selection biases: the privileging of printed editions over archival manuscripts, the authority of editorial committees determining what merits publication, and the institutional hierarchies governing access to collections. In this sense, the 'shadows' Putnam identifies in digital repositories mirror long-standing patterns in which scholarly gatekeepers have determined which sources circulate widely and which remain buried in archives. What digital transformation does change, however, is the *scale* at which these biases can be both reproduced and challenged. AI-driven approaches to mass digitization and computational analysis hold the potential to democratize access to previously marginalized sources and to disrupt established canons—not by eliminating selection, but by making its mechanisms visible and contestable (Rehbein 2024).

Latour's model makes visible that these digital interventions are not neutral supports. Every time we insert a digital link into the chain, be it a search algorithm or a text recognition model, we alter the trajectory of the reference. The aim of this article is to utilize this framework to identify these exact insertion points and to subject the implicit decisions made there (often by software developers rather than historians) to the same critical scrutiny we apply to the sources themselves. This attention to infrastructure echoes the work in Science and Technology Studies on how technical systems encode particular epistemologies and value commitments (Bowker and Star 2000). Classification systems, database schemas, and algorithmic parameters are never merely technical: They embody decisions about what counts as similar, what features matter, what can be ignored.

3.3. *The Data Layer as Intermediate Transformation*

Latour's framework further illuminates the specific nature of the steps labeled 'Digital Intervention' above. It helps us to understand the emergence of a 'data layer', a phenomenon often misunderstood in practice. When we apply AI-powered segmentation to cadastral maps, for instance, we are not simply 'mining' pre-existing information. Instead, we are extending the chain of reference by inserting a new, highly interpretative transfor-

mation. This reframing builds directly on Drucker's well-known argument that humanities materials should be understood as *capta* (taken) rather than *data* (given) (Drucker 2011): where Drucker usefully denaturalizes the apparent objectivity of digital datasets, the transformation chain framework allows us to specify, at the level of operational steps, *how* and *by whom* the taking is done, and which features of the original source are preserved or lost in the process. This operational specification depends, in turn, on recognizing that digital objects are never flat: as Fickers (2021) has shown, every digital representation has a 'layered materiality' in which hardware, software, file formats, and metadata schemata co-determine what can be read, retrieved, and transformed. The transformation chain we trace below thus operates not on 'the source' as such but on a specific configuration of these layers, each of which is itself the result of prior technical and editorial decisions. The process of generating this layer can be broken down into specific operations of abstraction:

1. Input (Visual representation): The digitized map is treated as a matrix of pixels. At this stage, it preserves the visual complexity and the 'noise' of the original archival document.
2. Operation (Algorithmic segmentation): An AI model predicts pixel classes (e.g., 'building' vs. 'background'). This is the act of *discretization*—forcing unambiguous categories onto a continuous or ambiguous reality.
3. Output (Data layer): The probability maps are converted into vector polygons. These are immutable mobiles: distinct, countable objects that can be moved into databases, separated entirely from the visual context of the map.

It is crucial to underscore that this generation of data is a fundamentally contingent process. The resulting dataset is never the inevitable extract of the source, but always one of many possible outcomes dependent on variable technical parameters. For instance, a shift in the optical resolution (DPI) during the initial digitization fundamentally alters the 'materiality' available to the machine. A map scanned at 300 DPI may yield a generalized set of building footprints, while the same map scanned at 600 DPI might allow the algorithm to distinguish between solid walls and garden fences. Similarly, choices regarding color depth or contrast thresholds in the pre-processing stage dictate what the algorithm 'sees' and what remains invisible. Thus, the data we analyze are not fixed properties of the past, but artifacts of a specific configuration of the technical apparatus. This contingency of data generation aligns with arguments in critical data studies emphasizing the performative rather than descriptive nature of datasets (Gitelman 2013; Schöch 2013), and it resonates more broadly with the long-standing insight in the history of science that 'data' are never simply given but are always answers to specific questions (Strasser and Edwards 2017). Data do not describe a pre-existing reality so much as enact particular ways of engaging with sources. The 300 DPI scan of a cadastral map is not 'more accurate' than the 600 DPI scan in any absolute sense—each resolution enacts a different reality, preserving different features, enabling different questions. The choice of resolution is thus not merely technical but fundamentally epistemological.

This operational sequence is never neutral; it is shaped by a negotiation process between historiographical intent and technical feasibility. Historians and computer scientists must jointly define the 'ontology of the model': How many nuances can be preserved? Should the algorithm distinguish between distinct types of buildings (e.g., residential vs. utility), or is a generic 'structure' label sufficient? In the Computational Humanities, these epistemological questions are frequently determined by resource constraints ('pragmatic modeling'). The cost of creating ground truth (training data) and the limitations of accessible computing power often dictate the final accuracy. Thus, the resulting data layer always represents a compromise between the ideal historical question and the material reality of the project's resources. Consequently, these decisions define the horizon of the research outcomes; they catalyze the specific form of knowledge that can be produced.

Although factors like annotation budgets or algorithmic architectures may initially appear extrinsic to the intellectual process of historical inquiry, they are effectively reconfigured here. They cease to be mere logistical details and become constitutive elements of the methodological apparatus itself, dictating the granularity and the reliability of the historical narrative that follows.

What we call ‘data’—the polygons representing buildings, roads, forests, and fields, or the pixels classified as water—are not atomic units that existed ‘in’ the map waiting to be extracted. Rather, these are generated through a process of discretization; they are created through a transformation that loses certain features (the aesthetic qualities of the original cartography, the variations in how different surveyors represented similar features) while gaining others (machine-readability, statistical analyzability, systematic comparability across map sheets). This reframing has profound implications. It dissolves the false dichotomy between ‘information extraction’ (which implies information pre-exists in sources) and ‘data creation’ (which implies data is invented by researchers). Instead, we recognize that all data are created through transformations, but these transformations are not arbitrary; they preserve certain features while making them mobile, comparable, and manipulable. The question is not whether data are ‘really there’ or ‘constructed,’ but whether the transformation chain is traceable, whether it preserves features relevant to our research questions, and whether it allows us to move back and forth between different levels of abstraction.

3.4. From Source to Data: The Chain of Circulating Reference

Latour’s framework allows us to dismantle the static hierarchy of the DIKW pyramid in favor of a dynamic model of circulation. The ‘data layer’ described above is not a distinct foundational stratum upon which information and knowledge sit; rather, it represents a specific node in a bi-directional chain of transformations. This reconceptualization has precedent in information science itself: [Frické \(2009\)](#) demonstrates that the DIKW hierarchy lacks coherent definitions, logical structure, or empirical support. More fundamentally, the hierarchy’s linear progression contradicts actual knowledge-making practices in all fields. Rather than abandoning these categories entirely due to DIKW’s flaws, scholars like [Rehbein \(2021\)](#) advocate for a model of multi-directional interdependence. Within historical inquiry, this implies that the comprehension of any single element is fundamentally predicated upon its relationship to the others. Scientific knowledge, as [Rheinberger \(1997\)](#) shows through laboratory ethnography, emerges through iterative cycles where ‘epistemic things’ (objects under investigation) and ‘technical objects’ (instruments and methods) co-evolve without clear hierarchy.

This bidirectionality is the hallmark of digital scholarship. We zoom from aggregate statistics to individual map sheets. We compare segmented layers with original rasters to validate model performance. However, this movement is no longer solely under the control of the historian. It is mediated by the technical dispositive itself. The ability to ‘zoom back in’ in order to traverse the chain upstream depends on interface designs, database architectures, and storage decisions made by computer scientists and engineers. If a data pipeline discards the link to the original image filename to save storage, the chain is broken; the reference ceases to circulate.

Therefore, the agency in this epistemological process is distributed. It is not just the historian interpreting sources but a collaborative interplay involving the historian’s questions, the engineer’s architectural decisions, the algorithm’s pattern recognition, and the hardware’s limitations. This distribution of agency has been productively theorized in Digital History through Kemman’s adaptation of the ‘trading zone’ concept ([Kemman 2021](#)), in which historians and computational experts negotiate shared practices and infrastruc-

tures without fully merging their disciplinary identities—a framework we return to in our concluding discussion.

For the limited operational purposes of this article, and as a working consequence of the framework developed above, we propose the following provisional re-descriptions of the three key terms within a transformation-chain perspective:

- **Data** are not atomic facts, but highly processed, mobile transformations optimized for manipulation.
- **Information** is not ‘contextualized data,’ but the active maintenance of the link between the abstraction (data) and its origin (source).
- **Knowledge** implies the mastery of the entire transformation chain, thus understanding not just what the data signify, but how to retrace the steps of their production and how to account for the decisions of all actors (human and non-human) involved in the process.

We offer these as analytical heuristics tied to our case studies rather than as universal definitions. They resonate with Rheinberger’s (1997) concept of ‘experimental systems’ in which knowledge emerges not from extracting information from objects but from constructing systems that make objects answerable to questions. For historical research, digital transformation chains constitute experimental systems in this sense—arrangements of sources, tools, protocols, and questions that co-produce historical knowledge. The historian’s task is then not to ‘extract information’ from sources but to design and maintain transformation chains that preserve epistemic reversibility while enabling new forms of engagement.

3.5. Implications for Data-Driven Historical Research

We argue that this theoretical reframing has immediate practical implications for data-driven historical research. It suggests that the value of computational methods lies not in ‘extracting information’ from sources but in creating new transformations that make certain features more visible, comparable, and analyzable while preserving traceability back to original sources. This aligns with arguments in digital humanities for what might be called ‘explicit hermeneutics’ (Capurro 2010)—making visible the interpretive decisions encoded in computational workflows. Berry and Fagerjord (2017) similarly argue for ‘critical digital humanities’ that interrogate rather than naturalize computational methods. Our contribution extends these calls by providing a concrete framework (transformation chains) for documenting and analyzing interpretive decisions.

When we use CNNs to segment cadastral maps, we are not revealing ‘the data in the maps’ but creating a new inscription a vector layer that preserves spatial relationships and feature classifications while losing artistic style and individual cartographic choices. The methodological question is not ‘Is this accurate?’ (in the sense of perfect correspondence), but rather, ‘Does this transformation preserve features relevant to our research questions? Can we trace the transformation chain? Can we move bidirectionally between segmented layers and original maps?’

Similarly, when we use transformer models to extract structured data from historical directories like the *Schematismus*, we create inscriptions (database records) that preserve hierarchical relationships and temporal patterns while losing the layout, typography, and organizational logic of the original volumes. The research value depends not on eliminating these losses but on making them explicit and on maintaining the ability to return to original sources when questions arise that require features lost in transformation.

This approach also suggests new forms of methodological rigor for Digital History. Rather than pursuing impossible goals of ‘complete’ or ‘objective’ digitization, we should focus on

1. Transparency about transformations: Explicitly documenting what is preserved and lost at each stage;
2. Traceability: Maintaining reversible links between transformed data and source materials;
3. Multiplicity: Creating multiple transformations of the same sources to preserve different features;
4. Bidirectionality: Developing workflows that allow movement both toward abstraction (statistics, models) and toward specificity (individual sources, contexts).

These principles find support in recent methodological reflections on computational historiography. [Guldi and Armitage \(2014\)](#), despite their controversial advocacy for ‘longue durée’ Digital History, correctly emphasize that computational scale must be balanced with interpretive depth. More critically, [Putnam \(2016\)](#) demonstrates how digitization’s ‘shadows’—systematic biases in what gets digitized and how—require historians to maintain awareness of transformation chains, understanding not just what digital sources reveal but what they obscure. Recent work on AI-based historical research, particularly the epistemological framework proposed by [Oberbichler and Petz \(2025a, 2025b\)](#), complements these insights by offering systematic criteria for source criticism, model criticism, and prompt and output evaluation when working with large language models. Our framework provides conceptual tools for navigating these tensions: by understanding digitization as transformation rather than reproduction, historians can critically assess both the possibilities and limitations of computational methods.

4. Case Study 1: The *Franzische Kataster*

The *Franzische Kataster* (1817–1861) represents one of the most extensive cadastral mapping projects of the entire nineteenth century ([Kain and Baigent 1992](#); [Scharf 2017](#)). Initiated by Emperor Franz I through the Land Tax Patent of 1817, this survey originally covered the Monarchy’s Cisleithanian territories—a total of 300,082 square kilometers. The resulting inventory consists of 164,357 hand-drawn, colored sheets (approx. 53 × 66 cm) documenting roughly 50 million parcels across 30,000 municipalities ([BEV—Bundesamt für Eich- und Vermessungswesen 2017](#); [Feucht 2008](#)). Present-day Austria alone is represented on 53,212 sheets ([Schlögl 2008](#)).

4.1. The Cadastre as Research Resource

The *Franzische Kataster* constitutes an invaluable resource for understanding land use, settlement patterns, and property relationships at the threshold of industrialization. Its fine scale—typically 1:2880, increasing to 1:1440 or 1:720 in populated areas—and comprehensive coverage make it theoretically possible to reconstruct spatial and socio-economic conditions with unprecedented detail. The cadastre systematically records not only topographic features but also property boundaries, building structures, infrastructure, and detailed land use categories, down to individual trees of economic significance ([Feucht 2008](#), pp. 7–8).

4.2. Georeferencing as First Transformation

Since the early 2000s, the Styrian Provincial Archive (*Steiermärkisches Landesarchiv*) has digitized and georeferenced the *Franzische Kataster* for Styria ([Baumann et al. 2025](#)). As outlined in Section 3, this marks a crucial transformation in the chain of reference: physical sheets—already abstractions of territorial realities—are converted into raster images and aligned with modern coordinate systems (see Figure 1).



Figure 1. Scan of an original sheet of the *Franziszeische Kataster*. The original image is held by the Styrian Provincial Archives, Graz; the scan was provided by the province's Department A17 (Statistics and Geoinformation). They have also kindly made the data available to us.

Georeferencing enables previously impossible systematic comparisons (Knowles 2008). With mean errors below 2.9 m (Baumann et al. 2025), the historical cadastre can now be overlaid with LiDAR models, modern orthophotos, and contemporary cadastral data. This integration allows for precise analyses of settlement morphology, landscape transformation, and infrastructural development against the backdrop of 200 years of change.

Critically, this transforms the cadastre's epistemic status. By shifting the maps from their original Habsburg spatial framework to a modern digital inscription, content becomes mobile across temporal and technical contexts. While involving loss (materiality, texture), the process grants digital manipulability and quantitative comparability. Following Latour, the chain extends as follows: territorial reality → field survey → measurement protocols → cartographic representation → digitization → georeferencing → spatially referenced raster images. Each step makes the data more mobile and amenable to computation.

4.3. AI-Supported Segmentation: Decomposing Information into Data

Building on the georeferenced foundation, we apply a UNet++ convolutional neural network (Zhou et al. 2018) to automatically segment buildings, roads, water bodies, and forests. This process transforms what traditional methodology views as documentary information into discrete, computationally manipulable data (see Figure 2).

Operating on 512×512 -pixel tiles, the model was trained on eleven cadastral municipalities in the Murtal district (province of Styria, Austria) using a combination of manual and model-assisted annotation. Despite source inconsistencies—such as varying drawing styles, color bleeding, and changing representational conventions—the model achieves high accuracy, with an overall Intersection over Union of 96.28% and F1 scores above 95% for all feature classes (Baumann et al. 2025).

Theoretically, this process is not merely 'mining' pre-existing information but generating data through systematic discretization. Consider buildings: The original maps depict heterogeneous hand-drawn polygons varying in color, line quality, and architectural detail. The model reduces this cartographic complexity to a binary classification—pixel X is 'building' or 'not building.' These classified pixels are then converted into vector polygons with precise coordinates and measurable properties.

This transformation involves distinct trade-offs. Lost are esthetic features, individual surveyor styles, and subtle cues regarding construction materials or perspective. Gained are systematic comparability, quantitative manipulability, and scalability. Every building is identified using consistent criteria, turning it into countable objects amenable to GIS operations and allowing the same analytical procedure to be applied across entire territories.



Figure 2. Georeferenced scan of a sheet of the *Franziszeische Kataster* after automatic segmentation; recognized buildings are pink, forests are green, rivers are red, and roads are light blue.

4.4. Scaling Analysis: From Exemplary to Comprehensive

The most significant epistemological implication of AI-supported segmentation concerns scale. Traditional research on the *Franziszeische Kataster* was necessarily selective—limited to specific villages or valleys due to the enormous labor required for manual extraction. This limitation reflected not merely pragmatism but an epistemological foundation that privileged micro-regional analysis and hermeneutical depth over statistical coverage, often leaving the representativeness of case studies unaddressed (Hudson 2000).

AI-supported segmentation fundamentally alters this constraint. Once a model achieves adequate accuracy, it can be applied systematically across entire territories. For Styria, approximately 60% of the provincial territory—where georeferencing is complete—can now be analyzed regarding building distribution, road networks, and land use for the 1817–1861 period (Baumann et al. 2025). This enables research into historical provincial settlement density, correlations between forest coverage and topography, and infrastructure patterns that were previously inaccessible.

Moreover, comprehensive coverage allows for unprecedented temporal comparison. By overlaying historical layers with contemporary spatial data (e.g., modern building footprints or LiDAR-derived forest cover), transformation patterns emerge that remain invisible to selective case studies. In this sense, the integrated research environment functions as a ‘historian’s macroscope’ (Graham et al. 2022): a computational framework allowing researchers to toggle between the large-scale detection of structural patterns and the close inspection of individual cases. This approach aligns with calls for ‘scalable reading’ (Manovich 2020) or ‘blended reading’ (Stulpe and Lemke 2016), combining systematic breadth with local depth to investigate settlement expansion, infrastructure persistence, and environmental change.

4.5. Insights into Historical Source Production

An unexpected benefit of AI-supported segmentation lies in revealing the internal logics and construction processes of the source material. During model training, segmentation errors frequently highlighted discontinuities at map sheet boundaries, where forest edges, roads, or streams were represented differently by adjacent surveyors. These inconsistencies reveal that the cadastre was not a uniform, standardized representation but a composite of individual judgments and local adaptations—aligning with STS findings on the interpretive flexibility inherent in standardization projects (Bowker and Star 2000).

These stylistic variations open a further analytical dimension: computational style analysis. Systematic differences in line weight, color palette, and feature representation constitute ‘cartographic fingerprints’ that may enable authorship attribution for individual surveyors. Beyond attribution, classification discrepancies often reflect material ambiguities in the built environment. For instance, inconsistencies in distinguishing wood from stone structures may capture vernacular practices that resisted the cadastre’s classificatory grid. Reading these discrepancies not as errors but as evidence transforms the map into a dynamic source for material culture.

Buildings provide particularly revealing examples. Where the model struggled to distinguish structures in dense clusters, it often mirrored difficulties faced by the original cartographers. Instances of mid-sheet recoloring suggest changes in surveying teams during the production process—insights into the workflow that are difficult to gain through traditional close reading. Similarly, overlaying segmented buildings with modern footprints yields empirical data on measurement precision, demonstrating high accuracy in flat terrain but noticeable triangulation difficulties in mountainous or forested regions.

4.6. Theoretical Analysis: Data, Information, and the Transformation Chain

How do we theoretically conceptualize AI-supported segmentation? Returning to our framework (Section 3), this process constitutes a new transformation in Latour’s chain of circulating reference. The specific chain for our cadastral research proceeds as follows:

- **Territorial reality:** The physical landscape (villages, fields, and forests) observed by surveyors.
- **Field survey:** Measurement and field book recording. *Loss:* Continuous landscape becomes discrete points. *Gain:* Quantifiable, transportable data.
- **Cartographic production:** Translation into graphic representation. *Loss:* 3D terrain becomes a 2D map; precision reduced to resolution. *Gain:* Visual comprehensibility and preservation.
- **Digitization and Georeferencing:** Scanning and coordinate alignment. *Loss:* Materiality and original spatial framework. *Gain:* Digital reproducibility and GIS integration.
- **AI segmentation:** Pixel classification via CNN. *Loss:* Cartographic style, esthetics, and ambiguity. *Gain:* Systematic classification, scalability, and statistical analyzability.
- **Vector conversion:** Creation of geometric polygons. *Loss:* Pixel-level granularity. *Gain:* Topological relationships and precise measurements.
- **Spatial analysis:** GIS operations. *Loss:* Individual feature uniqueness. *Gain:* Pattern recognition and hypothesis testing.

This chain exemplifies Latour’s ‘preservation through displacement’ (Latour 1999a, pp. 69–70). It demonstrates that ‘data’ are not intrinsic to the maps but depend on the technical parameters of transformation. Furthermore, the chain is bidirectional. Reference circulates from territory to database and back; statistical patterns in the data can prompt close reading of specific map sheets, akin to zooming in with a ‘macroscope.’ This validates the concept of ‘distant reading’ (Moretti 2013) as not a replacement for, but a complement to, hermeneutic engagement.

Crucially, this framework dissolves the dichotomy between ‘information extraction’ and ‘data creation.’ AI segmentation is neither pure extraction (finding pre-existing data) nor arbitrary creation. It is a systematic transformation that translates spatial relationships into computational forms. The success of this pipeline rested on aligning historical necessity with algorithmic logic, illustrated by three technical configurations:

1. **Resolution as Context (The Receptive Field):** We downsampled images not merely for efficiency but to expand the CNN’s ‘receptive field.’ This allowed the network to perceive sufficient spatial context to distinguish complex entities like buildings, rather than just analyzing texture at full resolution.
2. **Contextual Redundancy (Overlapping Inference):** To counter edge errors, we implemented a ‘quadruple inference’ strategy. Each map section was classified four times with shifting input windows. The final result is a consensus, mimicking a rigorous source critique where observations are verified from multiple angles.
3. **The Ontology of ‘Streets’ (Hermeneutics of Classes):** Initial segmentation of streets failed due to heterogeneity. Historical analysis revealed that maps depicted ‘common spaces’ (*Gemeinflächen*) rather than modern infrastructure. Adjusting the target class from a visual/modern category to a historical/legal one allowed the algorithm to converge.

These examples demonstrate that the resulting data are actively constructed through a dialogue between domain knowledge and machine vision. This iterative alignment parallels Rheinberger’s (1997) co-evolution of ‘epistemic things’ and ‘technical objects,’ where the research question and the instrument develop in unison.

5. Case Study 2: Structuring the Social Body—The Hof- und Staatsschematismus

While the *Franzische Kataster* organized the physical space of the Monarchy, our second case study turns to the organization of its social and administrative body. The *Hof- und Staatsschematismus* serves as the primary source: a directory of individuals employed at the Habsburg court and state, offering a longitudinal dataset of titles, offices, and salaries from 1702 to 1918 (Noflatscher 2004).

Comprising 153 editions and over 150,000 pages available as PDF scans (Österreichische Nationalbibliothek 2011), this series requires a methodological shift from image segmentation to layout analysis. The transformation from source to data here is not about separating foreground from background but about decoding the visual grammar of bureaucracy to understand structural hierarchies over time (see Figure 3).

5.1. From Courtly Representation to Administrative Statistics

The genesis of the *Schematismus* reflects the transformation of the Habsburg state itself. Rooted in early modern court almanacs, the publication originally served a ceremonial function: demonstrating the structure and glamor of the Court. However, as the administrative apparatus expanded, the publication’s function shifted from representation to statistics and control (Bauer 2002).

This shift is materially visible in the explosive growth of the volumes, from a mere 180 pages in 1702 to 1907 pages by 1918. This increase signals the transition from listing dignitaries to managing workforce ‘mass data.’ For the digital historian, this creates a dual challenge: the sheer volume prohibits manual transcription, while the historical variability of the text structure complicates automated extraction, requiring robust solutions for layout analysis (Muehlberger et al. 2019).

Heer. — Heeres-Anstalten.			329
Heeres-Anstalten.			
I. Heeres-Museum.			
(Wien, X., Artillerie-Arsenal.)			
Protector.			
Seine kais. u. kön. Hoheit Erzherzog Friedrich, G. d. Inf., etc. etc.			
Mitglieder des Kuratoriums.			
Se. Exz. Hans Graf Wilczek, EKO-R. 1., LO-Kd., G.W., O., u. KO-GK. Ebr. von Wien, Korneuburg u. Wildalpen, E. Mitgl. der Ak. der Wiss. in Wien, E. Präsid. der k. k. geogr. Ges., E. Mitgl. der Ak. der bild. Künste, der Genossenschaft der bild. Künstler u. des österr. Mus. für Kunst u. Ind., Präsid. des Rudolfiner, Protector u. Gründer der Wiener freiw. Rettungs-ges., Gh. R. u. Käm., erbl. Mitgl. des Herrenh.; Obmann des Sub-Komitees.			
Se. Exz. Eduard Gaston Graf Püttlich v. Pottanegg, EKO-R. 2., O., papstl. GO-GK., r. AO-GK., srh. SO-GK., schmh. lip. HO-EK., Th., Ph. & I. can. Dr., Gh. R. u. Käm., Sr. papstl. Heil. Haus- u. Referendar - Prälat, Tit. Bischof von Damiana, Groß-Kapitular des Deutschen Ritter-Ordens, etc. etc.			
Se. Exz. Vinzenz Graf Baillet de Latour, EKO-R. 1., Gh. R. u. Käm., Min. a. D., Mitgl. des Kunstrates des Museums für K. u. U., Kurator des österr. Mus. für Kunst u. Ind., Mitgl. des österr. archäolog. Inst. in Wien u. Mitgl. des Herrenh.			
Weckbecker Wilhelm, Freih. v., H. R. in Sr. Maj. Oberstkämmereramt.			
Wolovich v. Belobreska Emil, G. d. Inf. Schwab Johann, F. M. L.			
Krobatin Alexander, Rit. v., F. M. L. Se. Durchlt. Alois Fürst Schönburg-Hartenstein, Gh. R. u. Käm., Oberst, Vize-Präsid. des Herrenhauses.			
John Wilhelm, FJO-R., O., Ph. Dr., Art. Ob. Ing. III. Kl.			
Moll Alfred, J. Dr., Art. Ing.			
II. Militär-Erziehungs- und Bildungs-Anstalten.			
A. Militär-Akademien.			
Theresianische Militär-Akademie zu Wien.			
Für die Infanterie, Jäger u. Kavallerie. (Errichtet 1782 als adeliges Kadetten-Korps, 1788 Akademie genannt.)			
Giesl v. Gieslingen Artur, Freih. v., SJO-R., LO-R., EKO-R. 3., MVK., D3., O., pr. SJO. 2. i. Br., pr. KO-R. 2. m. St., s. AO-KL 2. pr. FO-KL 2. m. KO-Kd., old. HVO-EK., srh. TO. 3., srh. WAO. 5., F. M. L., Kmdt.			
Technische Militär-Akademie in Mödling.			
Für die Artillerie und die technischen Truppen.			
(1717 als Ingenieur-Schule errichtet; 1766 zur Akademie erhoben.)			
Dermus Georg, Rit. v., EKO-R. 3., FJO-R., MVK., D3., O., pr. KO-R. 3., bayr. MVO-R. 1., G. M., Kmdt.			
B. Infanterie-Kadetten-Schulen.			
(Die Kadettenschulen wurden 1869 errichtet, 1889 in die Militär-Erziehungs- und Bildungs-Anstalten eingereiht.)			
Zu Budapest.			
Möga Viktor, MVK., O., D3., O., Major, Kmdt.			
Zu Innsbruck.			
Eitz Emil, v., MVK., D3., O., Major, Kmdt.			
Zu Kamenitz (Kamenica) bei Peterwardein.			
Rostok Robert, MVK., D3., O., papstl. EK., Major, Kmdt.			
Zu Karlsstadt (Karlova).			
Kurelec v. Boine-mir Adolf, MVK., D3., O., Major, Kmdt.			
Zu Kaschau (Kassa).			
Kirschhofer August, MVK., O., D3., O., Major, Kmdt.			
Zu Königsfeld in Mähren.			
Eckelt Hugo, D3., O., Major, Kmdt.			
Zu Lemberg.			
Wolny Julius, MVK., D3., O., Major, Kmdt.			
Zu Liebenau bei Graz.			
Latzer Franz, MVK., O., D3., O., Oberstl., Kmdt.			
Zu Kobzów bei Krakau.			
Latnik Franz, MVK., D3., O., Major, Kmdt.			
Zu Marburg.			
Mayrhofer v. Grünbühl Josef, FJO-R., MVK., O., D3., O., srh. KO. 4., Oberstl., Kmdt.			
Zu Hermannstadt (Nagyseben).			
Hausner Oskar, O., D3., O., Oberstl., Kmdt.			
Zu Preßburg (Pozsony).			
(Mit einer Abteilung für die Sanitäts-Truppe.)			
Rehwald Karl, MVK., D3., O., Major, Kmdt.			
Zu Prag.			
(Mit einer Abteilung für die Train-Truppe.)			
Hauska v. Zbrankov Ottokar, MVK., D3., O., papstl. E. K., Oberstl., Kmdt.			
Zu Temesvár.			
Sztékay Ludwig, MVK., D3., O., Major, Kmdt.			
Zu Triest.			
Maag Eduard, MVK., O., D3., O., Major, Kmdt.			
Zu Wien.			
(XIII/3, Hütteldorferstraße.)			
Tomanek Edl. v. Beyerfels Friedrich, MVK., D3., O., Major,			
C. Kavallerie-Kadetten-Schule zu Mähr. Weißkirchen.			
Burka Anton, MVK., O., D3., O., Oberstl., Kmdt.			
D. Artillerie-Kadetten-Schulen.			
Zu Traiskirchen.			
Hausner Eduard, FJO-Off. D3., O., pr. KO-R. 3., Oberst, Kmdt.			
E. Pionier-Kadetten-Schule zu Hainburg.			
Weldinger Theodor, MVK., O., O., Major, Kmdt.			
F. Militär-Unter-Realschulen.			
Zu Fischau bei Wiener-Neustadt.			
Sedlakowitz Edl. v. Lanzekampf Csar, FJO-R., O., D3., O., Oberstl., Kmdt.			
Zu Kőszeg (Güns).			
Jerzábek Ferdinand, O., D3., O., Major, Kmdt.			
Zu Maros-Vásárhely.			
Chrestele Friedrich, O., D3., O., Hptm.			
Zu St. Pölten.			
Bannagh Theodor, FJO-R., D3., O., Oberstl., Kmdt.			
Zu Straß in Steiermark.			
Eypeltauer Rudolf, O., D3., O., Major, Kmdt.			
Zu Ems.			
Smuty Karl, FJO-R., O., D3., O., papstl. EK., Major, Kmdt.			
G. Militär-Ober-Realschulen.			
Zu Mähr. Weißkirchen.			
Merbeller Wilhelm, EKO-R. 3., MVK., O., D3., O., Oberst, Kmdt.			
Zu Eisenstadt (Kismarton).			
Zsivan v. Belenetz Alexander, O., D3., O., Oberstl., Kmdt.			
H. Erziehungs-Institut für verwaiste Offiziers-Söhne in Hirtenberg.			
(Errichtet 1888.)			
Kittner Marie, EKO. 2., GVK. m. K., O., Ob. Vorsteherin, Leiterin.			
III. Fachbildungs-Anstalten.			
1. Kriegsschule in Wien.			
(VL, Drechselsengasse 4.)			
(Errichtet 1882.)			
Puhalla v. Brög Paul, LO-R., EKO-R. 2., MVK., D3., O., sp. MVO-GK., s. AO-Kl. 1., wt. FO-Kl. 1., FML., Kmdt.			
Mianow Radoslav, O., D3., O., Major, Adj.			
Lehr-Personal.			
Pfeffer Rudolf, MVK., O., O., pr. KO-R. 3., gr. EO-Ed., Oberst.			
Rudel Rudolf, O., O., tosc. MVO-R.,			
Kopetzky Johann, MVK., D3., O., O.,			
Schilhawsky v. Bahnbrück Josef, Rit., O., O.,			
s. AO-R.			
Müller Rudolf, O., O.,			
Bertolas Richard, O., O.,			
Korzer Karl, O., O.,			
Domus Ferdinand, O., O.,			
Sündermann Ludwig, MVK., O., O.,			
Mayerhofer v. Vedropolje Eberhart, MVK., O., O.,			
Straub Johann, O., O.,			
Huber Josef, O., O.,			
Walzel Erich, v., O., O.,			
Paul Viktor, MVK., O., O.,			
Langer v. Langreide Hermann, Rit., O., O.,			
Hptl.			
Randa Maximilian, O., O.,			
Reichlin-Wedegg Friedrich, Freih. v., O., O., Ritter.			

Figure 3. Sample page of the *Schematismus*.

5.2. The Layout as an Institutional Archive

The primary digitization challenge lies less in text recognition (OCR) than in layout segmentation (Fleischhacker et al. 2025). The page layout is not arbitrary; it acts as a semantic carrier, visually encoding hierarchy and departmental affiliation (see Figure 4). The *Schematismus* can be roughly divided into three separate phases concerning different layouts:

1. Early Phase (1702–1774): Characterized by single, crowded columns with minimal typographical differentiation between headings and content, making semantic structure difficult to parse.
2. Transitional Phase (1775–1848): Aspect ratios normalize and centered headings appear to separate organizational units, yet the layout remains fluid, experimenting with varying column structures.
3. Consolidated Phase (1868–1918): Following the constitutional changes of 1848/1867, the layout stabilizes into a standardized three-column format, marking the ‘industrialization’ of the administrative record.

[illegible]

Figure 4. *Schematismus* page after layout detection.

5.3. Defining the Scope: A Strategic Reduction

Applying the logic of ‘contingent data generation’ (Section 3), we restricted the project’s temporal scope to handle the source’s heterogeneity. Automated layout analysis relies on recurring patterns; a model trained on 1750s single-column tables fails on 1900s three-column tables. Consequently, the project focused on the 1876–1918 series, representing the longest continuous period of consolidated layout structure (Fleischhacker et al. 2025)².

By accepting a reduction in historical breadth to achieve extraction accuracy, we underscore a crucial constraint: the ‘digital dataset’ is inextricably linked to the visual uniformity of the analog source. Where the source’s visual logic breaks down, the data pipeline effectively halts.

5.4. The Synthetic Bridge: Manufacturing Data to Read Data

A fundamental bottleneck in historical Deep Learning is the scarcity of ‘Ground Truth.’ Training a model to recognize that the complex layout of the *Schematismus* requires thousands of annotated pages—a task exceeding project resources.

Our solution involved a significant shift: instead of annotating existing sources, we manufactured them (see Figure 5). Using LuaTeX, we generated 3766 pages of synthetic training data (Fleischhacker et al. 2025). While content was randomized, these pages strictly adhered to historical layout rules (fonts and indentations). This adapts synthetic data strategies (Journet et al. 2017) to specific administrative conventions.

This approach inverts the hermeneutic process. The algorithm learns from a *formalized model* created by researchers, not the source itself. The ‘digital eye’ is trained on a Platonic ideal—perfectly aligned and noise-free—to recognize the imperfect scanned reality. This parallels [Daston and Galison’s \(2007\)](#) concept of ‘truth-to-nature,’ where idealized types are constructed to categorize messy empirical specimens.

FHO. MalO. MalO. DRO. SphO. EHO. DHO. ⑤ gdt. MVR. ⑤ ⑤. Abz Z. gdt. VK. mx. GO. ntl. NOO.	Klasse. Kriegskreuz für Zivilverdienste I.	Vom Jahre 1904. Rewitzer Esther. EH. FHO. ⑤ grb. MO. Kooh Jannik. Anseh. GK. DRO. GVK. bayr. MO.
Vom Jahre 1890. Hiale Maximilian. H. u. G. Kam. GBD. EHO. ⑤ Lausberger Franz. A. K. bev. Sek. Arzt. A. B. Dose. GBD. OFK. R. G. GOF. m. St. m. Schw. EKO 1. EKO 3. EKO 1. FHO. GVIO. EKO 1. EKO 3. EKO 1. ⑤. O. MVR. ⑤. nss. ZVO. anh. AO. pap. stl. ChO. H. R. d. AO. nss. AO. it. AO. VVO. ⑤. pr. H. pr. AO. blr. ZVO.	Klasse. Silbernes Verdienstkreuz Vom Jahre 1856.	Schneider Robert. oster. Geburtst. Sekt. Teleg. H. Mtm. V. GOF. GOF. GOF. SphO. EKO 3. SphO. EKO 1. DHO. EKO 3. EKO 1. MVR. ⑤. G. GTM. ⑤. ⑤. ⑤. 1894 M. hess. PhO. grb. HO. Friedrich David. OFK. GVIO. MKDHO.
Eisernes Verdienstkreuz. Tesfay Kilrom. R. Prof. Sekt. Fabrik. Jov. H. 1. d. L. Sun. R. Armes. Kind. Leg. gld. GK. gld. m. Schw. EKO 1. Br. EKO 3. EKO 1. MTHO. EKO 3. FHO. FHO. EHO. ⑤. ABWV. gdt. MVR. grb. MO. sta. GO. norw. LO. anh. AO. schw. St. pr. BAO. wt. FO. H. KO. Sina. Ali. m. L. EKO 3. DZ.	Klasse. Großkreuz des Leopold-Ordens. Vom Jahre 1884.	Erinnerungszeichen an Feldmarschall Erzherzog Albrecht. Geistliches Verdienstkreuz II. Klasse. Schneemann Dominik. I. Br. MalO. MVR.
Vom Jahre 1903. Kremeth Gertrude. GKord. EKO 3. ⑤. SilarO.	Klasse. Kistler Jörn. Kav. DevK. SIKO. 1864 M.	Sefelin Reinhard. europ. Dir. Rev. m. d. T. P. Mtm. I. Off. Arb. Dr. Gymn. Magnaten. Regul. m. Schw. ER. FHO. FHO. ⑤. ⑤. ⑤. HGO. erb. TO. ntl. NOO. parm. KSGO.
Orden vom Goldenen Vlies.	Silbernes Verdienstkreuz mit der Krone. Kistler Jörn. Kav. DevK. SIKO. 1864 M.	Sträßer Dominik. Ther. Ak. Kord. MalO. ⑤. pr. RAO.
Winchester Yusuf. theol. Dr. d. ges. Hilde. m. E. m. K. m. E. MalO. LO. EKO 2. DRO. GVIO. EKO 3. EKO 3. EKO 1. SphO. FHO. ⑤. DI. TLV. hess. LO. grb. HO. H. KO. mekl. GO. bss. SKO. parm. GO.	Geistliches Verdienstkreuz I. Klasse. Kommandeurskreuz des Sankt Stephans-Ordens.	Pohling Kai. ev. Tit. Res. L. Sch. Magistr. R. V. I. Br. GK. LO. MTHO. MTHO. ⑤. ⑤. GTM. MVR. ⑤. ⑤. EHO. erb. TO.
Kuzmanovic Predrag. m. d. T. Handelun- m. Schw. I. Br. d. St. GOF. GOF. DHO. EKO 1. EKO 2. EKO 3. SIKO. EHO. 1864 M. SVK. TLV. ⑤. ⑤. ⑤. ⑤. ⑤. MKDHO. ⑤. DZ. gdt. VK. haw. KO. r. SIO. rm. SIO. pp. ChO. ntl. GO.	Geistliches Verdienstkreuz I. Klasse. Kommandeurskreuz des Sankt Stephans-Ordens.	Bruckmann Nina. m. St. EKO 2. STV. I. ⑤. Vom Jahre 1880.
Derek Saleh. Hdt. Dr. d. Bdktr. Vsd. Rev. H. u. Min. H. Rav. Kord. GBD. GEL. EK. OFK. Kord. SphO. EHO. EKO 1. FHO. SphO. MalO. LO. SphO. SphO. ⑤. ⑤. DI. DI. MKDHO. ⑤. ⑤. DI. ABWV. bayr. MO. ntl. LO.	Geistliches Verdienstkreuz I. Klasse. Kommandeurskreuz des Sankt Stephans-Ordens.	Zeka Ismail. d. H. Landwirtschaft. Ges. Abg. Gbd. GBD. DHO. FHO. 1. GVIO. DHO. FHO. EHO. EHO. ⑤. EHO. EHO. ⑤. STM. I. ⑤. gdt. VK. ⑤. ⑤. sim. KO. lip. Hdt. gr. DZ. papstl. SO. pr. BAO. oth. HAO.
Hawes Twain. mtl. L. R. Press. m. d. T. GEL. m. K. EK. Kord. GK. DonK. GBD. EHO. ⑤. EKO 2. SIKO. GVIO. ⑤. GTM. STM. I. ⑤. ⑤. ABWV. Abz Z. ⑤. ⑤. sp. AO. mx. GO. sp. MNO.	Eisernes Verdienstkreuz. Salbioni Sara. EB. SIKO. ⑤. mekl. KO.	Wette Dominik. Kr. Hptm. Ak. d. bld. K. Inf. Ges. Exp. Off. Dr. d. Bdktr. m. E. Rd. m. Schw. Kr. KL. MalO. GVIO. MalO. EKO 2. GVIO. SIKO. ⑤. SIKO. EKO 3. DZ. Abz Z. MVR. ⑤. anh. AO. bayr. MVO. bayr. MO. IX. EKO. mont. DO. schw. WO. tun. NTO. mon. KO. H. KO. papstl. EK.
Discher Michael. Th. Dr. EKO. MalO. Abz Z. ott. NIO.	Eisernes Verdienstkreuz. Salbioni Sara. EB. SIKO. ⑤. mekl. KO.	Frank Joachim. wss. H. u. G. Kam. Gew. Gutsverw. math. Verw. R. EB. M. EZ. I. Br. M. EB. FHO. LO. MalO. MTHO. MalO. SVK. ⑤. 1864 M. pr. SJO. pap. stl. GO.
Stahl Tim. Se. Gedr. Mod. R. kath. Bdktr. Daz. i. d. prov. m. Schw. m. Schl. EHO. SIKO. DHO. EKO 2. MTHO. EKO. 3. SIKO. MalO. EKO 3. LO. ⑤. ⑤. EZIKW. DZ. GVK. GTM. slum. FO. Chais. Cham. Flow. r. WIO. erb. MO. fz. MIO. bayr. LO. schw. KO. bd. TrO. SmaO. bld. ZVO.	Eisernes Verdienstkreuz. Salbioni Sara. EB. SIKO. ⑤. mekl. KO.	Gedenkenzeichen an Kaiser und König Franz Joseph I. Lehmann Anja. korr. DonK. SphO. ⑤. H. es. AO.
Insigener Elisabeth-Theresien-Stiftung.	Eisernes Verdienstkreuz. Salbioni Sara. EB. SIKO. ⑤. mekl. KO.	Reno Hanna. m. T. u. R. L. Sch. Ther. Ak. Preth. Ak. d. Wss. L. Sun. R. m. K. DonK. GK. DZ. OFK. GOF. GK. R. Kr. E4. SIKO. FHO. LO. MTHO. MalO. MVR. TLV. bulg. AO. papstl. EHO. bulg. AO. mekl. KO. bss. RO. hess. LO.
Verwundetenmedaille.	Eisernes Verdienstkreuz. Salbioni Sara. EB. SIKO. ⑤. mekl. KO.	Gedenkenzeichen an Kaiser und König Franz Joseph I. Lehmann Anja. korr. DonK. SphO. ⑤. H. es. AO.
Seereise-Denkmünze 1892/93. Miltosh Lars. Adv. Leg. Gew. Statth. R. Prskt. Inf. Hts. Hsaren. H. Kam. Bez. Htsmschft. Adj. Kd. EH. SphO. MTHO. SphO. SphO. TLV. DI. ⑤. Abz Z. EZIKW. ⑤. MVR. ⑤. GVK. MKDHO. pr. TSO. bss. RO. Off. de l'Ac. grb. HO.	Eisernes Verdienstkreuz. Salbioni Sara. EB. SIKO. ⑤. mekl. KO.	Gedenkenzeichen an Kaiser und König Franz Joseph I. Lehmann Anja. korr. DonK. SphO. ⑤. H. es. AO.
Kriegskreuz für Zivilverdienste I.	Eisernes Verdienstkreuz. Salbioni Sara. EB. SIKO. ⑤. mekl. KO.	Gedenkenzeichen an Kaiser und König Franz Joseph I. Lehmann Anja. korr. DonK. SphO. ⑤. H. es. AO.

Figure 5. Synthetically generated page of the *Schematismus*.

5.5. The Architecture of Extraction: Layout Analysis and OCR

The extraction pipeline consists of two stages, each requiring technical adjustments that reflect the source's materiality (Fleischhacker et al. 2025).

5.5.1. Stage 1: Layout Detection (Faster R-CNN)

We employed a Faster R-CNN model (Ren et al. 2017) to identify structural blocks. Standard input resolutions (typ. 1333×800 pixels) proved insufficient, as delicate table separators and spacing vanished. Consequently, we upscaled inputs to 1988×1405 pixels (Fleischhacker et al. 2025). This stands in direct contrast to the map segmentation strategy (Section 4): while maps required *reduced* resolution to gain context, the text layout demanded *increased* resolution to preserve structural syntax.

5.5.2. Stage 2: Optical Character Recognition (Tesseract)

For text recognition, we fine-tuned the Tesseract engine. Standard German models failed to capture 19th-century Habsburg typographic idiosyncrasies (e.g., specific ligatures). By creating a custom ‘unicharset’ and training on our synthetic data (Fleischhacker et al. 2025), we transformed a generic reader into a specialized ‘bureaucratic reader.’ This aligns with recent work emphasizing the necessity of domain-specific training for historical OCR (Kiessling 2019).

5.6. Negotiating the Logic of the List

The pipeline's final configuration resulted from negotiating historical complexity and algorithmic feasibility, illustrated by three key decisions:

1. **The Ontology of the ‘Curly’ (Handling Hierarchy):** The curly bracket (*Geschweifte Klammer*) is a pervasive visual element in the *Schematismus*, grouping officials under shared institutional headings. Technically, however, it is merely a graphical symbol.

To preserve its hierarchical function for computational processing, we defined a dedicated ‘Curly’ class within a three-tiered layout ontology: at the lowest level, *Paragraphs* capture individual text blocks; *Headers* operate across all three levels to mark organizational units; and *Curly*, as a container class at the intermediate level, groups *Paragraphs* and *Headers* into semantically coherent units. This decision elevated a graphical punctuation mark to a structural entity, encoding the insight that in bureaucratic documents, visual syntax carries as much institutional meaning as the text itself.

2. **The Simulation of Variance (Synthetic Data Strategy):** Using synthetic data required encoding our understanding into generator code (e.g., distinguishing indentation types). Consequently, the algorithm’s subsequent ‘discovery’ of structures is tautological: it finds what it was told to look for. However, the strictly rule-based synthetic datasets are augmented during training with a comparatively small proportion (between 10 and 20%) of annotated original pages, which proves sufficient to enable the model to handle the diversity and heterogeneity of the historical source material with confidence. This augmentation effectively fine-tunes the model from an idealized formal representation to the heterogeneous reality of the historical source material. The extraction’s ‘objectivity’ is thus bound by the historian’s initial formalization of layout rules yet tempered by the model’s empirical exposure to the irregularities of the actual documents—a productive tension between idealized structure and archival reality.
3. **Resolution as an Epistemological Dialectic:** Comparing the case studies reveals a dialectic of resolution. We downsampled maps (Section 4) to capture context (‘forest over trees’), but upsampled text (Section 5) to capture structural ‘fences’ like separators. There is no ‘correct’ digital resolution; technical parameters depend on the hermeneutic goal—land use requires blur, administrative hierarchy requires sharpness. Moreover, the choice of model architecture itself imposes additional constraints. In our preliminary experiments with YOLO for layout detection, we found that it operated optimally within different parameter ranges than the Faster R-CNN architecture ultimately adopted, particularly regarding input dimensions and aspect ratios. Depending on the model, it can prove worthwhile to invest considerable effort in pre-processing—adjusting crop dimensions to achieve different aspect ratios, or up- and downscaling resolution—before the source material ever enters the detection pipeline. The ‘correct’ resolution is thus doubly contingent: determined not only by the historical question but also by the computational affordances of the chosen architecture.

The resulting dataset is not a raw copy but a structured interpretation, capturing the 19th-century ‘bureaucratic gaze.’ This exemplifies the transformation of ‘capta’ into data (Drucker 2011), reinforcing that data is bound to its generation context (Gitelman 2013). Strictly speaking, the data did not exist until the algorithm—trained on our premises—generated it. Using this dataset therefore requires a specific ‘digital source criticism’ (Fickers 2012): treating the pipeline not as a neutral conduit but as a transformative layer shaping historical evidence.

6. Case Study 3: The Datafication of Everyday Resistance in Nazi Germany—A Digital ‘Upcycling’ of an Archive Inventory of 10,000 + Criminal Proceedings of the Munich Special Court

6.1. Special Penal Judiciary in National Socialist German Reich

The Special Courts in 20th-century German History were no invention of the National Socialists but had been introduced in 1932 by the first German democracy—at the time, not least as an instrument against violent attacks on the Republic, that is, against Nazis in particular. However, the quality and quantity with which the Nazi regime implemented

the special criminal jurisdiction in its system of repression and persecution right from the beginning of the ‘seizure of power’ illustrate the multi-layered relationship between historically grown institutions of the modern, democratic constitutional and legal state on the one hand and Nazi ideology on the other. The Special Courts primarily dealt with ‘offenses’ that had been criminalized by the Nazi regime for ideological reasons, especially during the war, but also particularly serious crimes such as murder.

6.2. ‘Upcycling’ of the Original Court Files in Terms of ‘Modern Social History,’ and Computational History 1975–1977 and 2021–2026

The original court records of more than 10,000 cases heard by the Munich Special Court between 1933 and 1945 are almost entirely preserved in the Bavarian State Archives. In the 1970s, in the context of reappraising the Nazi era and the historiographical turn of the ‘modern social history,’ key data on the defendants and their crimes was systematically recorded by hand from all these files in a structured form ([Generaldirektion der Staatlichen Archive Bayerns 1975–1977](#)). Each of the historical court records was thus converted into a regest with 12 metadata entries, recording characteristics such as national and religious backgrounds, party affiliation, and profession (see Figure 6). This inventory constitutes what [Scheltjens \(2023\)](#) terms a ‘legacy collection’: a historical data compilation that does not comply with contemporary FAIR principles but preserves unique knowledge accumulated by a previous generation of researchers: knowledge that would be prohibitively expensive to reproduce from scratch. Our workflow can be productively described through Scheltjens’ concept of ‘upcycling,’ which theorizes a digital historical research practice aimed at increasing the scientific value of such legacy collections. Central to this framework is the analysis of paradata—the reconstruction of the information creation processes that preceded and accompanied the original compilation. Applied to our case, the 1970s inventory is not merely a ‘source to be digitized’ but a structured interpretive artifact whose epistemological choices—what to record, what to omit, and how to categorize—must be critically reconstructed before computational analysis can proceed.

```
(8649) 12938 Prozeß gegen den Invaliden Johann FÜRFANGER
          (geb. 19. Spt. 1896) aus München wegen
          Schwarzhandels.

          Urteil: 4 Jahre Zuchthaus, 1000.-RM Geld-
          strafe oder 20 Tage Zuchthaus, Ab-
          führung von 800.-RM Mehrerlös, Si-
          cherungsverwahrung
          (§ 1 KwVO; § 1 Preis-StrafVO; § 20a
          StGB)

          Das Reichsjustizministerium regte Nichtig-
          keitsbeschwerde mit dem Ziel der Todesstrafe
          an; gest. 28. Mrz. 1945.
          19. Mai 1943 - 16. Apr. 1945
          (2 SKLs 97/44)
```

Figure 6. Sample file index from the archive inventory of the original court files of the Munich Special Court (1933–1945), digitized by the Passau Chair of Computational Humanities.

6.3. The Limitations of Data-Driven Macroscopic Historical Research

Since 2021, this ‘secondary source’ from the 1970s was digitized as part of a collaborative project between the Passau Department of Computational Humanities and the Bavarian State Archives, and all of the regesta it contains were automatically transferred to a database using OCR and parsing. Although research on resistance against National Socialism has traditionally focused on significant individual resistance fighters, computer-assisted analysis of this data set can provide new macroscopic insights into repression and resistance as phenomena affecting society as a whole ([Ernst et al. 2023](#); [Gerstmeier et al. 2022](#)).

Even though the compilation work carried out by professional historians in the 1970s was meticulous and thus offers the possibility of accessing the content on a systematic level, there are nevertheless some inconsistencies and historical desiderata. Sometimes, poor print quality of the archive inventory has a limiting effect on OCR processing. In addition,

the respective legal bases were only recorded for proceedings since 1939. Metadata on the prosecutors and judges involved were not even part of the data model underlying the archive inventory, which is symptomatic of the nature of the state-funded reappraisal of Nazi injustice in the 1970s. Quite a few of the lawyers who became perpetrators during the ‘Third Reich’ were still alive thirty years after the end of the war or were even still in office. These inconsistencies pose a challenge for digital ‘upcycling’ in the sense of automated processes.

6.4. Analytical Yield and Epistemological Implications

The computational analysis of the Munich Special Court archive inventory fundamentally expands the scope of historical inquiry on repression and resistance during the Nazi era. By transforming a semi-structured corpus of around 10,000 case summaries into machine-readable data, it enables a shift from selective interpretation to systematic analysis without abandoning qualitative insight. The inventory’s standardized structure condenses heterogeneous case files into comparable units, making the corpus computationally accessible and enabling, for the first time, a comprehensive perspective on the full body of cases. The key analytical yield lies in pattern recognition: computational methods allow historians to detect correlations, for instance, between charges and sentencing, to reconstruct typical judicial practices, and to trace developments over time. Where qualitative approaches ask what happened in a particular case, computational analysis can ask how the system functioned overall, making visible the structural logic of Nazi justice rather than only its individual manifestations. Crucially, computational methods do not replace close reading but reframe it. They provide a macroscopic baseline against which individual cases can be interpreted, creating an iterative research logic in which large-scale analysis identifies regularities and anomalies while qualitative analysis explains and contextualizes them. The result is a more robust form of historical knowledge that combines empirical breadth with interpretive depth. A central epistemic consequence of this approach lies in the formalization of interpretation. Categories—types of charges, sentencing patterns, defendant profiles—must be explicitly defined, operationalized, and critically reflected upon. This differs from traditional qualitative work, where such assumptions often remain implicit. Computational analysis thus increases methodological transparency, but it also introduces new challenges: abstraction, potential loss of nuance, and the need to mediate between historical language as used in the original court records, the editorial choices of the 1970s compilers, and the analytical frameworks of the present.

6.5. Theoretical Analysis: A Multi-Generational Transformation Chain

Returning to the framework developed in Section 3, the Munich Special Court case extends the transformation chain model in a direction the first two case studies do not: across historiographical generations. Unlike the *Kataster* and the *Schematismus*, where the chain begins with a primary source and proceeds through digital interventions applied directly to that source, the starting point here is itself a secondary source—an interpretive artifact produced by professional historians under specific institutional and political conditions.

The transformation chain for this case proceeds as follows:

- **Original court records (1933–1945):** The primary sources—heterogeneous case files comprising indictments, verdicts, correspondence, and personal data.
- **Regestierung by historians (1975–1977):** Systematic extraction of 12 metadata fields per case. Loss: narrative complexity, procedural detail, the full texture of each file. Gain: structured comparability across 10,000+ cases. Critical omission: metadata on prosecutors and judges was deliberately excluded—a decision shaped not by archival

negligence but by the political constraints of the state-funded reappraisal of Nazi injustice at a time when many of the jurists involved were still alive or in office.

- **Print publication as archive inventory:** Fixation in a specific typographic format. Loss: flexibility of the underlying data model. Gain: durability, citability, institutional authority.
- **Digitization and OCR (2021 ff.):** Scanning and text recognition. Loss: print quality inconsistencies limit OCR accuracy. Gain: machine-readability and full-text searchability.
- **Parsing into a database structure:** Automated extraction of structured fields. Loss: formatting nuances, implicit hierarchies in the print layout. Gain: systematic queryability and statistical analysis.
- **Computational analysis:** Pattern recognition across the full corpus. Loss: individual case uniqueness. Gain: macroscopic perspective on repression and resistance as societal phenomena.

What distinguishes this chain from the *Kataster* and *Schematismus* pipelines is the presence of what Scheltjens (2023) calls *paradata* at its center: the 1970s *Regestierung* from this perspective becomes a historically situated interpretive act whose premises—what counted as relevant metadata, what fell outside the data model—are inscribed into the dataset as structural biases. The systematic absence of judicial personnel data is the most striking example: it is invisible to the OCR pipeline and undetectable by the parsing algorithm and would remain analytically invisible without domain-specific historical knowledge of the political conditions under which the inventory was compiled.

This case thus serves as a stress test for the transformation chain framework itself. It demonstrates that epistemic reversibility, the ability to move upstream from data to source, encounters a specific obstacle in multi-generational chains: the interpretive decisions of the earlier generation are partially irrecoverable from the digital dataset alone. The chain can be traced, but it cannot be fully reversed without returning to the original court files. This limitation does not invalidate the computational analysis, but it specifies exactly where the analysis is constrained and what questions it cannot answer.

7. Conclusions

This paper began with the assertion that the digitization of historical sources is not a mechanical reproduction but a complex process of translation. By comparing three distinct extraction pipelines—the AI-supported segmentation of the *Franziseischer Kataster* (cadastral maps), the layout analysis of the *Hof- und Staatsschematismus* (Administrative Directories), and the datafication of the Munich Special Court archive inventory—we have demonstrated that the transformation from ‘source’ to ‘data’ requires a continuous negotiation between historical logic and algorithmic constraints. Each case study reveals a different facet of this negotiation: the *Kataster* foregrounds the interplay between image resolution and spatial semantics; the *Schematismus* exposes how visual layout encodes institutional hierarchies that must be formalized before they can be computationally parsed; and the Special Court inventory illustrates how the transformation chain extends across historiographical generations, as a secondary source produced in the 1970s—itsself already an interpretive reduction of the original court records—becomes the object of a further digital ‘upcycling’ whose possibilities and limitations are shaped by the epistemological choices of its earlier creators.

7.1. Parameters as a Transdisciplinary Negotiation

Our findings challenge the notion of the ‘black box’ in Digital Humanities. Across all three case studies, the quality of the output depended not on the raw power of the algorithms but on specific manual adjustments—such as the deliberate downsampling of

maps, the generation of synthetic layout templates, or the parsing strategies adapted to the idiosyncrasies of a 1970s print publication—that functioned as hermeneutic interventions.

This observation bears a crucial consequence for the practice of Digital History: setting these parameters correctly requires a fundamental transdisciplinary negotiation. It is insufficient for historians to merely deliver data and for computer scientists to merely process it. Rather, both parties are forced to acquire a functional literacy in the other's discipline. This resonates with Hitchcock's call for historians to 'confront the digital' by interrogating the underlying code and algorithms with the same rigor usually applied to archival sources (Hitchcock 2013). The historian must grasp the concept of the 'receptive field' to understand why the machine misinterprets a map, while the engineer must understand the legal concept of 'public space' to understand why a street is not just a gap between buildings. The code, therefore, becomes a trading zone, in the sense developed by Kemman (2021) for Digital History collaborations, where the premises of historical critique and engineering logic must be explicitly aligned. In Kemman's terms, the historians involved in our case studies function as 'digital history brokers' who connect and translate between domain expertise and computational practice, maintaining their disciplinary identity while acquiring functional literacy in the other's field. This resonates with broader calls for a 'critical digital humanities' that interrogates rather than naturalizes computational methods (Berry and Fagerjord 2017).

7.2. *The Machine as a Stress Test for Source Quality*

Furthermore, the training process generated a 'hermeneutic feedback loop' that offered deep insights into the sources themselves. The resistance of the machine—its inability to classify certain segments—often pointed not to algorithmic failure, but to inconsistencies within the historical material. In the case of the Cadastre, the model's error maps effectively visualized the 'cracks' in the surveyor's logic; in the *Schematismus*, the layout analysis revealed shifting editorial standards that a human reader might have glossed over; in the Special Court inventory, the automated processing exposed inconsistencies in the 1970s compilation—from varying print quality affecting OCR accuracy to the systematic absence of metadata on prosecutors and judges—a lacuna that reflects not archival negligence but the political constraints governing the state-funded reappraisal of Nazi injustice at the time. In this capacity, the computer vision model acts as a stress test for the archive. It allows us to assess the internal consistency and quality of the original source material with a precision that the naked eye could rarely achieve. The 'artificial' gaze thus sharpens the 'historical' gaze, revealing the fragility of the bureaucratic order the sources claim to represent.

7.3. *Scaling the Hermeneutic Circle*

Ultimately, the value of this 'operationalized hermeneutics' lies in its scalability. Traditional historical scholarship is often trapped in a tension between the deep analysis of individual cases (micro-history) and the generalization of systemic trends (macro-history). While Guldi and Armitage (2014) famously called for a return to the *longue durée* to address this, they offered little guidance on how to technically achieve it with fragmented archives. The methods presented here offer a bridge across this divide. By training models that encode expert domain knowledge—specifically tailored to the logic of the source—we can process entire archives with a consistency that no human reader could maintain. The Special Court case illustrates this most vividly: where research on resistance against National Socialism has traditionally centered on prominent individual figures, the comprehensive datafication of over 10,000 criminal proceedings enables a macroscopic perspective on repression and everyday resistance as societal phenomena—precisely the shift from exemplary to comprehensive analysis that our theoretical framework demands. This does not

replace the historian's judgment but scales it. Ideally, this leads to a 'statistically secured hermeneutics,' where the interpretation of the past is supported by the comprehensive analysis of its material traces.

Author Contributions: Conceptualization and methodology W.T.G. and M.R.; writing—original draft preparation, W.T.G.; writing—review and editing, M.R. and M.G.; project administration, W.T.G. and M.R.; funding acquisition, W.T.G. and M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/PAT1763723 and 10.55776/FG3100.

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: The principal project contributors for Sections 4 and 5 are Jörg Baumann and David Fleischhacker; significant conceptual contributions were further provided by Roman Kern. Figures 1, 2, 4 and 5 were produced by Jörg Baumann and David Fleischhacker. Open Access Funding by the University of Graz.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OCR	Optical Character Recognition
HTR	Handwritten Text Recognition
AI	Artificial Intelligence
CNN	Convolutional Neural Network
GIS	Geographic Information System/Science
DPI	Dots Per Inch
DIKW	Data–Information–Knowledge–Wisdom
IoU	Intersection over Union
R-CNN	Region-based Convolutional Neural Network
PDF	Portable Document Format
LiDAR	Light Detection and Ranging
STS	Science and Technology Studies

Notes

- ¹ The Boa Vista expedition is recounted in Chapter 2 ('Circulating Reference') of Latour (1999b, pp. 24–79). For earlier formulations of these ideas on 'immutable mobiles' and 'centers of calculation,' see Latour (1986).
- ² Note that the workflow in (Fleischhacker et al. 2025) is a developmental milestone and differs from the current project workflow.

References

- Antenhofer, Christina, Christoph Kühberger, and Arno Strohmeier, eds. 2024. *Digital Humanities in den Geschichtswissenschaften*. Wien: UTB/Böhlau, vol. 6116. [CrossRef]
- Armaselu, Florentina, and Andreas Fickers, eds. 2024. *Zoomland: Exploring Scale in Digital History and Humanities*. Studies in Digital History and Hermeneutics. Berlin and Boston: De Gruyter Oldenbourg, vol. 7. [CrossRef]
- Bauer, Volker. 2002. Territoriale Amtskalender und Amtshandbücher im Alten Reich. Bilanz eines Forschungsprojekts. *Rechtsgeschichte (Zeitschrift des Max-Planck-Instituts für europäische Rechtsgeschichte)* 1: 71–89. [CrossRef]
- Baumann, Jörg, David Fleischhacker, Wolfgang Goederle, Nicole Kamp, and Roman Kern. 2025. Automated Segmentation of Historical Cadastral Maps: A Machine Learning Approach to the Franziszeischer Kataster. Paper presented at AGIT Conference, Salzburg, Austria, July 2–3.
- Bernheim, Ernst. 1908. *Lehrbuch der historischen Methode und der Geschichtsphilosophie: Mit Nachweis der wichtigsten Quellen und Hilfsmittel zum Studium der Geschichte*, 6th ed. Leipzig: Duncker & Humblot.
- Berry, David M., and Anders Fagerjord. 2017. *Digital Humanities: Knowledge and Critique in a Digital Age*. Cambridge: Polity Press.

- BEV—Bundesamt für Eich- und Vermessungswesen, ed. 2017. *200 Jahre Kataster. Österreichisches Kulturgut 1817–2017*. Wien: BEV—Bundesamt für Eich- und Vermessungswesen.
- Bloch, Marc. 1953. *The Historian's Craft*. New York: Knopf.
- Bloor, David. 1999. Anti-Latour. *Studies in History and Philosophy of Science Part A* 30: 81–112. [CrossRef]
- Boonstra, Onno, Leen Breure, and Peter Doorn. 2004. *Past, Present and Future of Historical Information Science*. Amsterdam: NIWI-KNAW.
- Boshof, Egon, Kurt Düwell, and Hans Kloft. 1973. *Grundlagen des Studiums der Geschichte: Eine Einführung*, 1st ed. Cologne: Böhlau.
- Bowker, Geoffrey C., and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. Cambridge: MIT Press.
- Capurro, Rafael. 2010. Digital hermeneutics: An outline. *AI & Society* 25: 35–42. [CrossRef]
- Clavert, Frédéric, and Caroline Muller. 2017–2026. *Le goût de l'archive à l'ère numérique*. Collaborative Online Book. Ongoing Collaborative Publication. Available online: <https://gout-numerique.net/> (accessed on 10 June 2026).
- Clavert, Frédéric, and Serge Noiret, eds. 2013. *L'histoire contemporaine à l'ère numérique/Contemporary History in the Digital Age*. Brussels: Peter Lang.
- Cohen, Daniel J., and Roy Rosenzweig. 2005. *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*. Philadelphia: University of Pennsylvania Press.
- Collingwood, Robin George. 1946. *The Idea of History*. Oxford: Oxford University Press.
- Crymble, Adam. 2021. *Technology and the Historian: Transformations in the Digital Age*. Topics in the Digital Humanities. Urbana, Chicago and Springfield: University of Illinois Press. [CrossRef]
- Daston, Lorraine, and Peter Galison. 2007. *Objectivity*. New York: Zone Books.
- Deicke, Aline, Jonathan D. Geiger, Marina Lemaire, and Stefan Schmunk. 2024. *Living Handbook "Digitale Quellenkritik"*. First published version, July 2024. Ongoing collaborative project. Hamburg: Digital Humanities im deutschsprachigen Raum.
- Döring, Karoline Dominika, Stefan Haas, Mareike König, and Jörg Wettlaufer, eds. 2022. *Digital History: Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft*. Berlin: De Gruyter Oldenbourg. [CrossRef]
- Droysen, Johann Gustav. 1868. *Historik: Vorlesungen über Enzyklopädie und Methodologie der Geschichte*. Munich: Oldenbourg.
- Drucker, Johanna. 2011. Humanities Approaches to Graphical Display. *Digital Humanities Quarterly* 5. [CrossRef]
- Eberle, Oliver, Jochen Büttner, Hassan El-Hajj, Grégoire Montavon, Klaus-Robert Müller, and Matteo Valleriani. 2024. Historical insights at scale: A corpus-wide machine learning analysis of early modern astronomic tables. *Science Advances* 10: ead1719. [CrossRef] [PubMed]
- Ernst, Marlene, Sebastian Gassner, Markus Gerstmeier, and Malte Rehbein. 2023. Categorising Legal Records—Deductive, Pragmatic, and Computational Strategies. *Digital Humanities Quarterly* 17. [CrossRef]
- Feucht, Rainer. 2008. *Flächenangaben im österreichischen Kataster*. Diploma thesis, Technische Universität Wien, Institut für Geoinformation und Kartografie, Vienna, Austria.
- Fickers, Andreas. 2012. Towards a New Digital Historicism? Doing History in the Age of Abundance. *Journal of European Television History and Culture* 1: 19–26. [CrossRef]
- Fickers, Andreas. 2020. Update für die Hermeneutik. *Geschichtswissenschaft auf dem Weg zur digitalen Forensik? Zeithistorische Forschungen/Studies in Contemporary History* 17: 157–68. [CrossRef]
- Fickers, Andreas. 2021. Authenticity. Historical Data Integrity and the Layered Materiality of Digital Objects. In *Digital Roots. Historicizing Media and Communication Concepts of the Digital Age*. Edited by Gabriele Balbi, Nelson Ribeiro, Valérie Schafer, and Christian Schwarzenegger. Studies in Digital History and Hermeneutics. Berlin and Boston: De Gruyter Oldenbourg, vol. 4, pp. 299–312. [CrossRef]
- Fickers, Andreas, Juliane Tatarinov, and Tim van der Heijden. 2022. Digital History and Hermeneutics—Between Theory and Practice: An Introduction. In *Digital History and Hermeneutics: Between Theory and Practice*. Edited by Andreas Fickers and Juliane Tatarinov. Studies in Digital History and Hermeneutics. Berlin: De Gruyter Oldenbourg, vol. 2, pp. 1–20. [CrossRef]
- Fleischhacker, David, Roman Kern, and Wolfgang Göderle. 2025. Enhancing OCR in historical documents with complex layouts through machine learning. *International Journal on Digital Libraries* 26: 3. [CrossRef]
- Fogel, Robert William, and Stanley L. Engerman. 1974. *Time on the Cross: The Economics of American Negro Slavery*. Boston: Little, Brown.
- Föhr, Pascal. 2019. *Historische Quellenkritik im Digitalen Zeitalter*. Glückstadt: Werner Hülsbusch. Ph.D. dissertation, Universität Basel, Basel, Switzerland. Available online: <https://edoc.unibas.ch/64111/> (accessed on 10 June 2026).
- Frické, Martin. 2009. The knowledge pyramid: A critique of the DIKW hierarchy. *Journal of Information Science* 35: 131–42. [CrossRef]
- Generaldirektion der Staatlichen Archive Bayerns, ed. 1975–1977. *Widerstand und Verfolgung in Bayern 1933–1945. Hilfsmittel. Archivinventare, Bd. 3: Sondergericht München, 7 Teile*. Munich and Regensburg: Generaldirektion der Staatlichen Archive Bayerns.
- Gerstmeier, Markus, Simon Donig, Sebastian Gassner, and Malte Rehbein. 2022. Die Archivinventare zum Sondergericht München (1933–1945) digital. Quellenwert—Verdatung—Erkenntnisperspektiven. *Archivalische Zeitschrift* 99: 215–51. [CrossRef]
- Gitelman, Lisa, ed. 2013. *Raw Data Is an Oxymoron*. Cambridge: MIT Press.
- Goederle, Wolfgang. 2015. Die räumliche Matrix des modernen Staates: Die Volkszählung des Jahres 1869 im Habsburgerreich im Lichte von Latours zirkulierender Referenz. *Schweizerische Zeitschrift für Geschichte—Revue Suisse d'histoire* 65: 414–27.

- Graham, Shawn, Ian Milligan, Scott B. Weingart, and Kim Martin. 2022. *Exploring Big Historical Data: The Historian's Macroscope*, 2nd ed. Hackensack: Imperial College Press.
- Guldi, Jo, and David Armitage. 2014. *The History Manifesto*. Cambridge: Cambridge University Press.
- Hiltmann, Torsten. 2024. (Epistemologische) Grundlagen der Anwendung digitaler Methoden in den Geschichtswissenschaften. In *Digital Humanities in den Geschichtswissenschaften*. Edited by Christina Antenhofer, Christoph Kühberger and Arno Strohmeier. Wien: UTB/Böhlau, vol. 6116, pp. 43–59.
- Hitchcock, Tim. 2013. Confronting the Digital: Or How Academic History Writing Lost the Plot. *Cultural and Social History* 10: 9–23. [\[CrossRef\]](#)
- Howell, Martha, and Walter Prevenier. 2001. *From Reliable Sources: An Introduction to Historical Methods*. Ithaca: Cornell University Press.
- Hudson, Pat. 2000. *History by Numbers: An Introduction to Quantitative Approaches*. London: Arnold.
- Jannidis, Fotis, Hubertus Kohle, and Malte Rehbein, eds. 2017. *Digital Humanities: Eine Einführung*. Stuttgart: J.B. Metzler. [\[CrossRef\]](#)
- Journet, Nicholas, Muriel Visani, Boris Mansencal, Kieu Van-Phan, and Antoine Billy. 2017. DocCreator: A new software for creating synthetic ground-truthed document images. *Journal of Imaging* 3: 62. [\[CrossRef\]](#)
- Kain, Roger J. P., and Elizabeth Baigent. 1992. *The Cadastral Map in the Service of the State: A History of Property Mapping*. Chicago: University of Chicago Press.
- Kemman, Max. 2021. *Trading Zones of Digital History*. Studies in Digital History and Hermeneutics. Berlin and Boston: De Gruyter Oldenbourg, vol. 1. [\[CrossRef\]](#)
- Kiessling, Benjamin. 2019. Kraken—An Universal Text Recognizer for the Humanities. Paper presented at the Digital Humanities Conference 2019, Utrecht, The Netherlands, July 9–12.
- Knowles, Anne Kelly, ed. 2008. *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship*. Redlands: ESRI Press.
- Langlois, Charles-Victor, and Charles Seignobos. 1898. *Introduction aux études historiques*. Paris: Hachette.
- Latour, Bruno. 1986. Visualization and Cognition: Drawing Things Together. In *Knowledge and Society: Studies in the Sociology of Culture Past and Present*. Edited by Henrika Kuklick. Greenwich: JAI Press, vol. 6, pp. 1–40.
- Latour, Bruno. 1999a. Circulating Reference: Sampling the Soil in the Amazon Forest. In *Pandora's Hope: Essays on the Reality of Science Studies*. Cambridge: Harvard University Press, Chapter 2, pp. 24–79.
- Latour, Bruno. 1999b. *Pandora's Hope: Essays on the Reality of Science Studies*. Cambridge: Harvard University Press.
- Manovich, Lev. 2020. *Cultural Analytics*. Cambridge: MIT Press.
- Milligan, Ian. 2019. *History in the Age of Abundance? How the Web Is Transforming Historical Research*. Montreal: McGill-Queen's University Press.
- Moretti, Franco. 2013. *Distant Reading*. London: Verso.
- Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, Hervé Déjean, Markus Diem, Stefan Fiel, and et al. 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation* 75: 954–76. [\[CrossRef\]](#)
- Noflatscher, Heinz. 2004. "Ordonnances de l'hôtel", Hofstaatsverzeichnisse, Hof- und Staatskalender. In *Quellenkunde der Habsburgermonarchie (16.–18. Jahrhundert). Ein exemplarisches Handbuch*. Edited by Josef Pauser, Martin Scheutz and Thomas Winkelbauer. Wien and München: R. Oldenbourg Verlag, pp. 59–75.
- Oberbichler, Sarah, and Cindarella Petz. 2025a. Evaluating Bias within an Epistemological Framework for AI-Based Research in the Humanities. In *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale. Proceedings del XIV Convegno Annuale AIUCD 2025*. Venezia: AIUCD, pp. 52–59.
- Oberbichler, Sarah, and Cindarella Petz. 2025b. *Working Paper: Implementing Generative AI in the Historical Studies*. Geneva: Zenodo. [\[CrossRef\]](#)
- Österreichische Nationalbibliothek. 2011. *ALEX—Historische Rechts- und Gesetzestexte: Staatshandbuch*. Digitale Sammlung. Digitalisierte Ausgaben des Hof- und Staatsschematismus (1702–1918). Vienna: Österreichische Nationalbibliothek.
- Putnam, Lara. 2016. The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast. *American Historical Review* 121: 377–402. [\[CrossRef\]](#)
- Rehbein, Malte. 2021. A Responsible Knowledge Society within a Colourful World. *Philosophisches Jahrbuch* 128: 156–68. [\[CrossRef\]](#)
- Rehbein, Malte. 2024. Künstliche Intelligenz und Datenmobilisierung zwischen Geschichtswissenschaft und Archiv. Über Forschung in der digitalen Transformation und die Notwendigkeit des Umdenkens. *Archivalische Zeitschrift* 101: 11–27. [\[CrossRef\]](#)
- Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39: 1137–49. [\[CrossRef\]](#) [\[PubMed\]](#)
- Rheinberger, Hans-Jörg. 1997. *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford: Stanford University Press.
- Rowley, Jennifer. 2007. The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science* 33: 163–80. [\[CrossRef\]](#)
- Rygiel, Philippe. 2017. *Historien à l'âge numérique*. Villeurbanne: Presses de l'Enssib. [\[CrossRef\]](#)

- Scagliola, Stefania, Daniele Guido, Andreas Fickers, and Gerben Zaagsma. 2018. Ranke.2: A Teaching Platform for Digital Source Criticism. Launched October 2018. Available online: <https://ranke2.uni.lu/> (accessed on 10 June 2026).
- Scharr, Kurt. 2017. Kataster und Grundbuch im Kaisertum Österreich. Ausgangssituation und Entwicklung bis 1866. In *200 Jahre Kataster. Österreichisches Kulturgut 1817–2017*. Edited by BEV—Bundesamt für Eich- und Vermessungswesen. Wien: BEV—Bundesamt für Eich- und Vermessungswesen, pp. 37–51.
- Scheltjens, Werner. 2023. Upcycling Historical Data Collections: A Paradigm for Digital History? *Journal of Documentation* 79: 1325–45. [CrossRef]
- Schlögl, Michaela. 2008. Ein digitales Abbild Kakanien. *Die Presse*, March 18.
- Schöch, Christof. 2013. Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities* 2: 2–13.
- Siebold, Anna, and Matteo Valleriani. 2022. Digital Perspectives in History. *Histories* 2: 170–77. [CrossRef]
- Spiegel, Gabrielle M. 2007. Revising the past/revisiting the present: How change happens in historiography. *History and Theory* 46: 1–19. [CrossRef]
- Strasser, Bruno J., and Paul N. Edwards. 2017. Big data is the answer ... but what is the question? *Osiris* 32: 328–45. [CrossRef] [PubMed]
- Stulpe, Alexander, and Matthias Lemke. 2016. Blended Reading. Theoretische und praktische Dimensionen der Analyse von Text und sozialer Wirklichkeit im Zeitalter der Digitalisierung. In *Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*. Edited by Matthias Lemke and Gregor Wiedemann. Wiesbaden: Springer VS, pp. 17–61. [CrossRef]
- Thaller, Manfred. 2017. Historical Information Science: Is there such a Thing? New Comments on an old Idea. In *From History to Applied Computer Science in the Humanities*. Historical Social Research Supplement. Cologne: GESIS, vol. 29, pp. 260–86. [CrossRef]
- Tosh, John. 2021. *The Pursuit of History: Aims, Methods and New Directions in the Study of History*, 7th ed. London: Routledge.
- von Brandt, Ahasver. 2012. *Werkzeug des Historikers: Eine Einführung in die Historischen Hilfswissenschaften*, 18th ed. Stuttgart: Kohlhammer.
- Zhou, Zongwei, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer, pp. 3–11. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.