Evaluating bias within an epistemological framework for AI-based research in the humanities

Sarah Oberbichler*❖, Cindarella Petz*☼

DH Lab, Leibniz Institute of European History,
Alte Universitätstr. 19, 55116 Mainz, Germany
❖ oberbichler@ieg-mainz.de
☼petz@ieg-mainz.de
*These authors contributed equally to this work.

ABSTRACT (ENGLISH)

In this paper, we propose an epistemological framework for AI-based digital research in the humanities. Based on two of our current research projects using Large Language Models (LLMs) for various analysis tasks, we demonstrate the critical assessment of bias examination as one vital aspect of our proposed epistemological framework. There, we are focusing on biases inherent to the LLMs, and biases, that emerge due to the application of the LLM, which significantly impact research outcomes. This includes how language-specific prompts may favor certain languages, resulting in differing outcomes, as well as how LLMs can amplify biases present in source materials, such as reproducing problematic labels or using discriminatory language. Source critical approaches and bias output evaluation helps identify these patterns and allows researchers to make informed decisions when using LLMs for analysis tasks such as information extraction.

Keywords: LLMs; Output Evaluation; Bias Examination; FAIR AI; Ethical AI

TITLE AND ABSTRACT (ITALIANO)

La valutazione dei bias in un framework epistemologico per la ricerca basata sull'IA nelle scienze umane

In questo lavoro proponiamo un framework epistemologico per la ricerca digitale basata sull'intelligenza artificiale nelle scienze umane. Basandoci su due dei nostri attuali progetti di ricerca che utilizzano Large Language Models (LLM) per vari compiti di analisi, dimostriamo che la valutazione critica dei *bias* è un aspetto vitale del nostro framework epistemologico proposto. In questo caso, ci concentriamo sui *bias* intrinseci agli LLM e su quelli che emergono dalla loro applicazione, che hanno un impatto significativo sui risultati della ricerca. Ciò include il modo in cui i prompt testuali possono favorire alcune lingue, dando luogo a risultati diversi, nonché il modo in cui gli LLM possono amplificare i pregiudizi presenti nei materiali di partenza, come la riproduzione di denominazioni problematiche o l'uso di un linguaggio discriminatorio. Gli approcci critici alle fonti e la valutazione dei *bias* nei risultati aiutano a identificare questi schemi e consentono ai ricercatori di prendere decisioni informate quando utilizzano gli LLM per compiti di analisi come l'estrazione di informazioni.

Parole chiave: LLM; Valutazione degli Output; Analisi dei *bias*; IA FAIR; IA Etica

1. An epistemological framework for AI-based research in the humanities

Digital research is changing rapidly through lowered barriers to digital and computational tools — in particular based on Large Language Models (LLMs) — and intensified efforts to link FAIR cultural heritage data with them.[1] This presents us with both opportunities to democratize knowledge through broader engagement with cultural heritage, but also increases the challenge on how to provide guidance for the responsible use of those resources and the appropriate application of these new tools.

1 An example for such an effort is the European ECHOES project, which is designed to establish the European Collaborative Cloud for Cultural Heritage (ECCCH), a shared platform to link heritage professional and researchers to open data and scientific resources from GLAM institutions fostering collaboration across previously fragmented communities. There, advanced digital tools and specifically AI technology plays a crucial within the supported projects, such as AUTOMATA and TEXTaiLES (https://www.echoes-eccch.eu/projects/).

To address this, we argue that established digital epistemological frameworks within the Digital Humanities disciplines need to be extended and adapted to include the deployment of AI models: to combine traditional practices such as verifying information with new approaches to critically assess complex AI systems, their prompts, and outputs.

We therefore introduce an epistemological framework for AI-based research in the humanities, providing guidance for the critical reflection and systematic examination of how knowledge is created, validated, and transmitted through digital tools and data — specifically when working with generative AI. This includes critical evaluation of sources, methodologies, and tools used in knowledge production, as well as understanding how these elements shape our interpretation and understanding of cultural heritage material.

Drawing from the authors' research projects on "Transnational Flows of News in Historical Newspapers" and on the "Construction of Political Criminality during the Dollfuß-/Schuschnigg-Regime", this paper demonstrates how the application of such a framework can enhance methodological rigor and critical reflection in the AI-assisted analysis of cultural heritage resources based on the example of understanding and evaluating biases within sources and algorithms.

This framework goes beyond interpretability and explainability approaches – such as technical tools like Neuronpedia[2], which visualizes internal model activations, or surface-level transparency like DeepSeek's[3] natural language explanations of its reasoning steps — and black-box vs. white-box models. Both black-box and white-box AI models face fundamental challenges with explicability and explainability. While black-box models inherently struggle with explainability due to their opaque nature, white-box models, despite being interpretable in theory, can become equally challenging to explain when their complexity increases. Both struggle with explicability, the broad challenge of enabling accountable usage and appropriate understanding across different stakeholders. The complexity of AI systems means that even when we can trace how a decision is made (as in white-box models), we may still fail to make this understanding accessible and meaningful to the stakeholders who need to work with these systems (Herzog, 2022).[4]
In order to mitigate this and to understand and use AI responsibly, this requires examining:

- The quality and reliability of knowledge sources
- The appropriateness and trustworthiness of the methods
- The validation, evaluation, and ethical implications of how we use this knowledge

The proposed framework as visualized in Figure 1 can provide such a comprehensive approach to assessing AI models from an epistemological standpoint, connecting source, data, and tool criticism with prompt and output evaluation (compare to Oberbichler & Petz 2025). This framework addresses three interconnected dimensions:

The first dimension focuses on source and data criticism, where traditional research practices must evolve to meet new challenges. Epistemologically, the reliability of knowledge is directly tied to the quality of its sources. While it has a long tradition in humanities research to evaluate representativeness, appropriateness, quality, authenticity, and bias of used sources, these need to be extended also onto LLMs, which are both data and tool / method. Additionally, researchers must carefully consider the legal dimensions of their data: copyright protections, licensing restrictions, and the handling of personal data under EU AI Act and or national laws requirements which need to influence the choice of which models and tools to work with, and how the corpora needs to be created and analyzed. Compliance with these regulations ensures that the knowledge generation process adheres to legal and ethical standards and enhances trustworthiness.

2 Neuronpedia, https://docs.neuronpedia.org/.

3 DeepSeek, https://chat.deepseek.com/ .

4 Similarly, the European Data Protection Supervisor (2023) stressed the need for transparent documentation in order for a model to be understood beyond mere technical transparency.

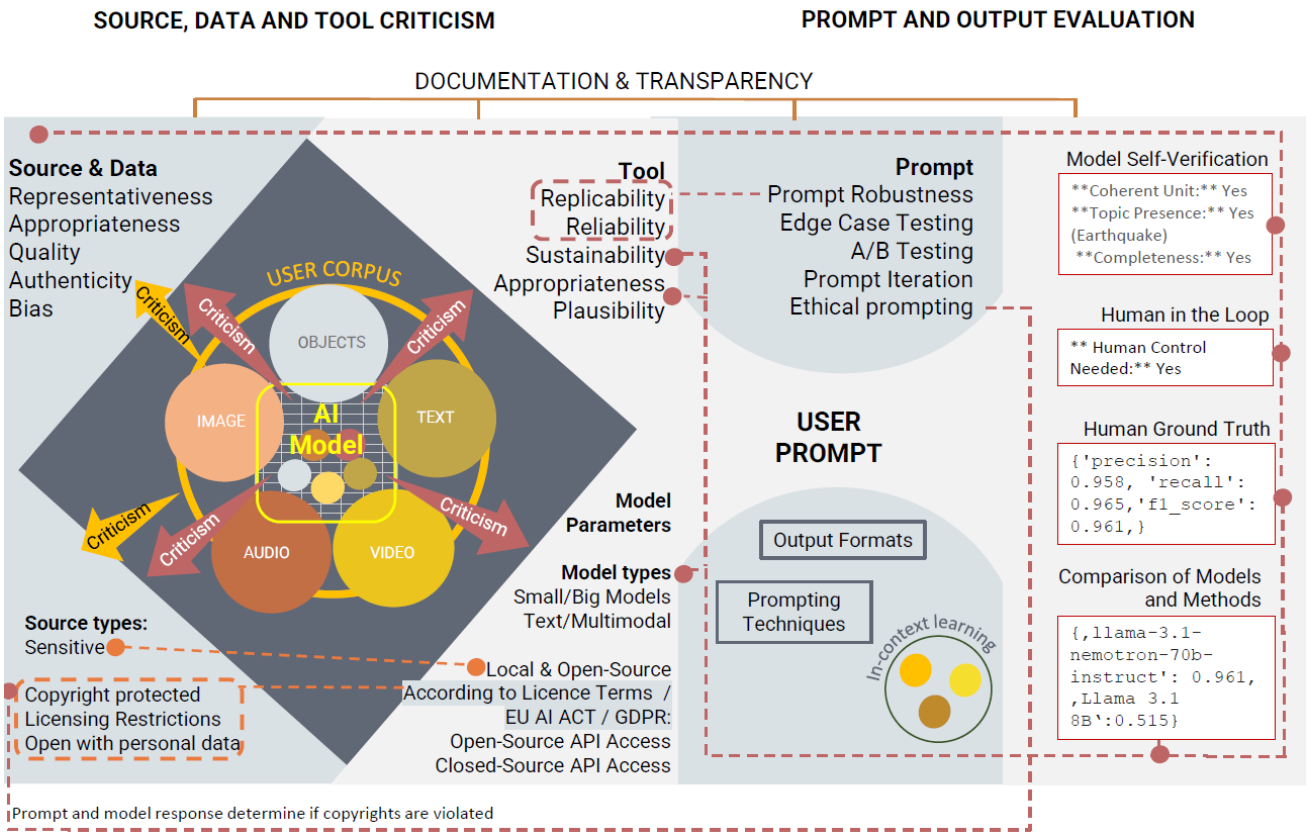

Figure 1. Epistemological framework for AI-based research in the humanities (Image: S. Oberbichler & C. Petz 2024).

The second dimension centers on tool and model criticism, emphasizing the replicability, reliability, sustainability, appropriateness, and plausibility of chosen LLM tools. These factors are crucial for justifying the methods used in knowledge generation. The tools and methods must be reliable and appropriate to produce trustworthy knowledge. This includes careful consideration of model parameters and types, whether working with small or large models, text-only or multi-modal systems. For researcher or any person using LLMs as an analysis tool, it is suggested to evaluate different access options, weighing the implications of open-source / open-weight versus closed-source model options, and assessing the advantages and limitations of local versus cloud-based deployment options. The recent FAIRification of LLMs — i.e., the extension and re-interpretation of the FAIR data principles to include AI models, algorithms, software and tools which produce data — answered to the need for thorough technical documentation of model characteristics[5] and persistent identification (compare e.g., Ravi et al., 2022; Huerta et al, 2023; Hupont et al., 2023; Alkemade et al., 2023; Golpayegani et al., 2024). This transparency furthers the scientific re-usability of trained models and AI tools, and increases both the reliability and interpretability of models, allowing for cross-comparisons and the reproducibility of results. If copyright or licensing restrictions are present in the data base at question, using local models should be considered the most responsible option. While small local models have the advantage of not needing additional computing power, comparative evaluations can determine whether larger LLMs achieve significantly better results that justify their usage.

The final dimension involves prompt and output evaluation, requiring the implementation of robust prompting techniques and evaluation guidelines. This includes systematic edge case testing, A/B testing, prompt iteration, ethical prompting, and the integration of model self-verification protocols. These elements ensure that the interaction with LLMs is both consistent and ethical. Critical to this dimension is

5 This includes model training data, input and output data types, configurations, dependencies, benchmark metrics, and uncertainty quantification for performance evaluation, as well as usage (and storage) guidelines in model cards.

the establishment of the human-in-the-loop validation processes and systematic comparison of different models and methods to ensure optimal results.
Throughout all these dimensions, maintaining comprehensive documentation and transparency of decisions remains crucial. An example template of such a documentation is shown in the *AI model Research Documentation Sheet* that is part of this framework (Oberbichler, 2024).

2. Practical application of the epistemological framework on the example bias examination

Based on two of our current research projects using AI models for various analysis tasks, we will demonstrate the critical assessment of bias examination as one vital aspect of our proposed epistemological framework. Biases manifest in AI models on various levels: in the model design, in the training data and it's selection, which then may translate into biased model outputs, and in its (mis-)application on context the model wasn't trained for (compare e.g., Ferrer et al., 2021). Source and data selection biases can result in misrepresentation of the context at study due to unbalanced domain distributions, cultural, temporal and language constraints, or demographic skews. This can lead to demonstrable social biases in model outputs, including prejudices and even discrimination based on gender, age, ethnicity, religion, disability status, and socioeconomic class (compare e.g., Dai et al., 2024; Gallegos et al., 2024; Navigli et al., 2023; European Union Agency for Fundamental Rights, 2022). Source critical approaches and bias output evaluation helps identify these patterns and allows researchers to make informed decisions when using LLMs for analysis tasks such as information extraction.

Amplification of bias and Ethical AI in historical court records

Based on the research project on "The Construction of Political Criminality during the Dollfuß-/ Schuschnigg-Regime (1933—1938)," we reflect on the amplification of biases through LLMs, both due to the models used and due to inherent biases within the source basis at study. There, we are interested in the appropriateness of the AI-assisted analysis:

1. Do different models have different capabilities for text extraction of key characteristics of the trial based on what they have been trained on? How does prompt engineering facilitates this?
2. How can we combat the amplification of biases inherent to the source basis at study? And does Ethical AI[6] obscure the analysis of discriminatory language present in historical sources?

In this study, we work with historical court records, which offer an ultimately non-objective worldview imposed by the court personnel, the judiciary, and the police onto the defendants and the case at trial. Schwerhoff (2011, p. 40) characterized the nature of court records as 'involuntary ego-documents' of the accused, where the self-declarations of the accused become altered through the court's perspectives and their recording decisions, obscuring factual representation. In our case study on the semantic construction of political criminality in the prosecution of political-religious offenses during the autocratic Dollfuß-/Schuschnigg-regime in Austria (1933—1938), this becomes evident, when specific labels are imposed on the accused within the trials, influencing their case outcomes. In the context of political-religious offenses, these are often moralizing, evoke ethnic differences, and/or dependent on the political orientation of the accused (e.g., framing someone as morally corrupt, Socialist, Jewish, or in health-related terms), revealing hidden patterns of the justice systems' labeling processes of morals, deviant behaviors, as well as political judiciary. When found guilty of charges, these labels and the punishment of conviction too have repercussions for the social standing of the accused, leading to stigmatization and ostracization in Austrian society.
The identification of these inherent perspectives of court records tie into the first dimension of our proposed framework. Another aspect of this are legal considerations: court records as sources of sensitive data require the adherence to their legal protections not only when publishing the study's results and accompanying datasets with person-related details, but already when analyzing these records with generative AI. This makes it necessary to opt for localized data and prompt processing with our chosen

6 Ethical AI in the context of non-maleficent and just AI is trained to avoid unfair biases (compare to overview on Ethical AI frameworks in Prem, 2023, p. 702).

model (or on German Data Privacy Regulations-bound servers when using high performance computing services).
The second and third dimension of our proposed framework are closely connected when assessing model characteristics, optimizing prompts, and evaluating outputs: what data have the chosen models been trained on, and how does this influence the results? Which prompts result in appropriate results? This makes it necessary to opt for open-access models in order to be able assess the appropriateness and plausibility of chosen LLMs when systematically comparing results from multiple models and different prompting strategies with a human-created ground truth.
While the source and data biases described above necessarily reappear in model output when utilizing them as Retrieval Augmented Generators, we do need to make sure not to amplify discriminatory historical terms. This study demonstrates the need for context-awareness of the human-in-the-loop for output evaluation in order to frame and possibly even offer a counter-narrative for the automatic findings of the model used, and suggest further editing to combat problematic or discriminatory language when presenting findings and when publishing the underlying dataset.

Bias evaluation when analyzing multilingual historical newspaper corpora

Using the research project "Transnational Flows of News: Analysis and Visualization of Historical News across Languages and Countries, 1850–1950" as an example, we demonstrate how systematic evaluation of AI model outputs contributes to understanding the model's appropriateness. This is crucial when working with multilingual historical sources where cultural and linguistic contexts vary greatly. We specifically want to determine whether model's training data sufficiently represents historical and linguistic diversity for accurate analysis, and whether its design effectively accommodates human-in-the-loop processes: and

1. Do different models have varying capabilities across languages? Does the language of the prompt matter?
2. How good is the historical context understanding? Are the models able to find articles related to a specific event without providing further historical context?
3. Are LLMs biased towards specific output formats?
4. How well are the models aligned for historical analysis?

The underlining project for this evaluation examines transnational news flows in Germany, France, Great Britain, Italy and Switzerland from 1850-1950, focusing on news about (re)migration and environmental / natural disasters. The goal of this study is to provide new insights into the creation and manipulation of narratives across languages and time.

Crucial for this project is the creation of a topic-specific corpus containing well separated articles. Previous layout based article separation methods currently don’t provide good enough (>70% accuracy) results (Sun et al., 2024) or are not flexible enough to be adapted for this research project. Therefore, we created a novel approach using OCR'ed text and context-understanding for article extraction via LLMs. For this purpose, we created an evaluation framework for three LLMs tasks:

- Classification: Testing models' ability to classify articles containing OCR mistakes as relevant or not relevant to the specific topic of the 1908 Messina earthquake.
- Extraction: Evaluating accuracy in extracting complete relevant articles from newspaper in different languages
- Boundary Detection: Assessing ability to correctly mark beginning and end of articles using different output formats, especially when several articles were published in the same newspaper issue.

The proposed bias evaluation within this case study occurs through multiple steps within the epistemological framework. Analyzing both language and historical domain bias involves addressing representation issues within model training data, which falls under data and source criticism (first and second dimension of our proposed framework. Language evaluation also constitutes part of tool / model criticism, as selecting a monolingual model would be inappropriate for multilingual datasets. Although

critical assessment of model training sources is often limited, both language and domain biases can be quantified by comparing model outputs with human ground truth (third dimension). Such comparisons can reveal whether a model performs significantly worse when extracting information from non-English texts compared to English texts. Historical domain bias can be evaluated by including both topic-relevant and non-relevant texts in a test corpus. Comparing the model's classification results against human-created ground truth will demonstrate whether the model possesses sufficient contextual understanding to complete this task. Additionally, closely examining a model's reasoning and explanations for classifying sources as relevant or non-relevant can provide valuable insights into its historical knowledge and its ability to integrate geographical, cultural, and historical information (Mauermann and Oberbichler,2025). Model design biases, such as distortions introduced through a model's alignment or instruction-following approach, are part of tool/model criticism and fall under the aspect of model appropriateness. Model alignment and instruction-following issues may manifest as failures to follow specific instructions—such as the directive not to truncate when extracting long articles, to preserve OCR mistakes, or to adhere to a specified output format. These issues can be quantified by comparing model outputs to provided ground truth. Another critical consideration is a model's confidence alignment. A well calibrated model should be able to highlight cases of ambiguity or uncertainty to facilitate human verification. Models that exhibit overconfidence in their answers fail to create space for human-in-the-loop approaches. This type of bias can be evaluated by comparing models' self-estimated confidence levels with output evaluation results, evaluating instructions that encourage human interaction, or analyzing the logits of a model (Mauermann and Oberbichler, 2025).

In both case studies, these approaches to identify biases demonstrate the interconnectedness of our proposed framework's dimensions, when critically assessing our sources, models, operationalisations, and outputs. This multilayered approach equips researchers with the necessary critical framework for understanding both the strengths and limitations of a model.

3. Outlook: Bias and Ethical AI

The intersection of AI technology and humanities research presents both opportunities and challenges for knowledge democratization. Our proposed epistemological framework offers a systematic approach to understanding and addressing these challenges, particularly regarding the complex relationship between ethical AI principles and historical research integrity.
Through our case studies - examining political criminality in Austrian court records and analyzing multilingual newspaper coverage - we demonstrate how ethical AI considerations must be carefully balanced with historical authenticity. While ethical AI aims to minimize discriminatory language and biased representations, humanities researchers must sometimes preserve these elements to accurately document historical discrimination and prejudice. While this preservation is integral for historical inquiry, when presenting and publishing these results and underlying datasets, every effort needs to be made to mitigate discriminatory language and labels. Our proposed framework helps researchers navigate this tension by:

1. Recognize and document biases inherent in both historical sources and AI models
2. Develop systematic evaluation processes that account for linguistic and cultural diversity
3. Maintain human oversight while leveraging AI capabilities
4. Balance ethical AI principles with accurate historical representation

This approach to understanding AI tools helps counter biases and distortions, which is central to democratizing access to knowledge and cultural heritage. By implementing robust evaluation frameworks, we can better harness AI's potential while maintaining the rigorous standards of humanities research.
The future of AI technology in the humanities will require continuous refinement of these approaches. Our work provides a foundation for this ongoing development, while acknowledging that ethical AI evaluation is not a fixed destination but rather a continuous process of improvement and adaptation.

ACKNOWLEDGEMENTS
No AI-based tools were used for creating the content of this text.

REFERENCES

Alkemade, H., Claeyssens, S., Colavizza, G., Freire, N., Irollo, A., Lehmann, J., Neudecker, C., Osti, G., & van Strien, D. (2023). Datasheets for Digital Cultural Heritage Datasets. https://doi.org/10.5281/zenodo.8375034.

European Data Protection Supervisor. (2023). TechDispatch: explainable artificial intelligence. #2/2023. Publications Office. https://data.europa.eu/doi/10.2804/802043.

European Union Agency for Fundamental Rights. (2022) Bias in Algorithms – Artificial Intelligence and Discrimination. Vienna, 2022. https://fra.europa.eu/de/publication/2022/bias-algorithm#publication-tab-1.

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in AI: A cross-disciplinary perspective. IEEE Technology and Society Magazine, 40(2), 72-80. https://doi.org/10.1109/MTS.2021.3056293.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey (No. arXiv:2309.00770). arXiv. https://doi.org/10.48550/arXiv.2309.00770.

Golpayegani, D., Hupont, I., Panigutti, C., Pandit, H. J., Schade, S., O'Sullivan, D., & Lewis, D. (2024). AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act. https://doi.org/10.48550/ARXIV.2406.18211.

Herzog, C. (2022). On the risk of confusing interpretability with explicability. AI and Ethics, 2(1), 219–225. https://doi.org/10.1007/s43681-021-00121-9.

Huerta, E. A., Blaiszik, B., Brinson, L. C., Bouchard, K. E., Diaz, D., Doglioni, C., Duarte, J. M., Emani, M., Foster, I., Fox, G., Harris, P., Heinrich, L., Jha, S., Katz, D. S., Kindratenko, V., Kirkpatrick, C. R., Lassila-Perini, K., Madduri, R. K., Neubauer, M. S., Psomopoulos, F. E., Roy, A., Rübel, O., Zhao, Z. & Zhu, R. (2023). FAIR for AI: An interdisciplinary and international community building perspective. Scientific Data, 10(1), 487. https://doi.org/10.1038/s41597-023-02298-6.

Hupont, I., Fernández-Llorca, D., Baldassarri, S., & Gómez, E. (2023). Use case cards: a use case reporting framework inspired by the European AI Act. https://doi.org/10.48550/ARXIV.2306.13701
LIAS: Layout Information-Based Article Separation in Historical Newspapers | SpringerLink. (n.d.). Retrieved January 2, 2025, from https://link.springer.com/chapter/10.1007/978-3-031-72437-4_15.

Mauermann, J., & Oberbichler, S (2025). LLM Biases: Expected and Unexpected Model Design Effects in Historical Newspaper Article Extraction on the Messina Earthquake. DH Lab. https://doi.org/10.58079/137qr.

Navigli, R., Conia, S., & Ross, B. (2023). Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information Quality, 15(2), 10:1–10:21. https://doi.org/10.1145/3597307

Oberbichler, S. (2024). AI Model Research Documentation Sheet (AIRDocS). https://doi.org/10.5281/zenodo.14550113.

Oberbichler, S., & Petz, C. (2025). Working Paper: Implementing Generative AI in the Historical Studies (1.0). Zenodo. https://doi.org/10.5281/zenodo.14924737.

Prem, E. From ethical AI frameworks to tools: a review of approaches. AI Ethics 3, 699–716 (2023). https://doi.org/10.1007/s43681-023-00258-9.

Ravi, N., Chaturvedi, P., Huerta, E. A., Liu, Z., Chard, R., Scourtas, A., Schmidt, K. J., Chard, K., Blaiszik, B., & Foster, I. (2022). FAIR principles for AI models with a practical application for accelerated high energy diffraction microscopy. Scientific Data, 9(1), 657. https://doi.org/10.1038/s41597-022-01712-9.

Schwerhoff, Gerd. 2011. Historische Kriminalitätsforschung. Frankfurt/M.: Campus Verlag.

Sun, W., Tran, H. T. H., González-Gallardo, C.-E., Coustaty, M., & Doucet, A. (2024). LIAS: Layout Information-Based Article Separation in Historical Newspapers. In A. Antonacopoulos, A. Hinze, B. Piwowarski, M. Coustaty, G. M. Di Nunzio, F. Gelati, & N. Vanderschantz (Eds.), Linking Theory and Practice of Digital Libraries (pp. 256–272). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72437-4_15.

Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., & Xu, J. 2024. Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24). Association for Computing Machinery, New York, NY, USA, 6437–6447. https://doi.org/10.1145/3637528.3671458.