Kuratierung in Kulturerbe-Einrichtungen

Beitrag für eine qualitäts- und wertebasierte Kultur künstlicher Intelligenz

Clemens Neudecker & Jörg Lehmann

Die Geschwindigkeit der Entwicklung von künstlicher Intelligenz (KI) ist immens. In immer kürzeren Abständen preisen die großen IT-Unternehmen Google, Meta, Microsoft und OpenAI die vielfältigen Verheißungen des neuen technologischen Wunderwerks an. Sogar das baldige Auftreten allgemeiner künstlicher Intelligenz (AGI), also zu selbstständigem Denken fähiger und die intellektuellen Fähigkeiten von Menschen übersteigender KI, wird bereits von manchen herbei phantasiert. Auch in weiten Teilen des GLAM Sektors (Galleries, Libraries, Archives, Museums) sind die Beschäftigten zunehmend alarmiert, denn KI-Systeme versprechen vieles, was zum Kernbereich der Tätigkeiten in Bibliotheken gehört, zukünftig automatisiert oder sogar besser zu erledigen – wodurch perspektivisch auch Arbeitsplätze bedroht sein können. Weithin strahlende Beispiele für durch KI erzielte Durchbrüche im Bereich des kulturellen Erbes stehen hingegen noch aus. Zumindest in Teilbereichen konnten jedoch bereits bedeutende Erfolge erzielt werden, so etwa bei Texterkennung von historischen Drucken (Reul et al., 2018) und Handschriften (Colutto et al., 2019; Kiessling et al., 2019), bei Bildklassifikation und Bildähnlichkeitssuche (Lee et al., 2020; Brantl and Schweter, 2022) oder auch im Bereich der Verarbeitung natürlicher Sprache[1], wie etwa Named Entity Recognition und Entity Linking (Ehrmann et al., 2022). Auch für spezifisch bibliothekarische Tätigkeiten wie die Sacherschließung wird bereits seit einigen Jahren an maschineller Unterstützung durch KI gearbeitet (Suominen, 2019), wenngleich die Ergebnisse aktuell meist noch nicht die Qualität menschlicher Erschließung erreichen (Mödden et al., 2018; Kasprzik, 2020; Kasprzik, 2023). In der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (SBB) wird das Thema KI vor allem in Drittmittelprojekten vorangetrieben. So beteiligt sich die SBB an der Entwicklung der OCR-D Software mit KI-basierten OCR-Verfahren (Neudecker et al., 2019), entwickelte und erprobte im Projekt QURATOR (Rehm et al., 2020) verschiedene Verfahren für die KI-gestützte Erschließung und Kuratierung und setzt aktuell die vielversprechendsten Erkenntnisse daraus in einem weiteren Projekt mit dem Titel »Mensch.Ma-

1 Natural Language Processing (NLP).

schine.Kultur«[2] um. Vernetzung und Austausch zu KI finden zudem bereits auf nationaler Ebene[3], innerhalb Europas[4] sowie international[5] statt. Einen Überblick zu Themen und Herausforderungen für maschinelles Lernen in Bibliotheken bietet der Bericht »Machine Learning + Libraries: A Report on the State of the Field« von Ryan Cordell (2020).

Tatsächlich sind die neuesten Algorithmen und Modelle jedoch häufig nicht unmittelbar für die Anwendung auf kulturelles Erbe geeignet. Dies liegt darin begründet, dass die Leistungsfähigkeit von KI-Modellen immer auf den Daten beruht, mit denen sie trainiert wurden. Und während nur wenig Informationen öffentlich dazu vorliegen, welche Trainingsdaten OpenAI, Microsoft, Google und Meta für ihre Modelle heranziehen, so wird spätestens beim Testen der KI deutlich, dass diese bislang nur in geringem Maße auf die Besonderheiten kultureller Artefakte aus mehreren Jahrhunderten eingestellt sind. Hierin verbirgt sich zugleich eine große Chance für Kulturerbe-Einrichtungen: Gute Trainingsdaten für KI zeichnen sich neben ihrem Umfang auch durch Vielfalt und Relevanz aus. Man könnte sogar so weit gehen, zu behaupten, dass gute Trainingsdaten für KI immer auch eine Kuratierung erfordern. Kuratierung wird hier in einem weiten Sinne verstanden, der sich auf mehrere Unteraufgaben erstreckt. Zunächst muss durch Selektion sichergestellt werden, dass die Trainingsdaten auch wirklich zu der Aufgabe passen, die eine KI daraus lernen soll. Die Trainingsdaten sollten zudem auf ihre Qualität geprüft und ggf. enthaltene Fehler oder Falschinformationen gekennzeichnet oder korrigiert werden. Eine Strukturierung der Trainingsdaten oder deren Aufteilung in Segmente nach bestimmten inhaltlichen oder technischen Kriterien kann je nach Anwendungsfall auch notwendig sein. Die Verknüpfung mit weiteren Ressourcen und Wissensquellen stellt eine Form der Kontextualisierung dar, die ebenfalls zu einer besseren Qualität der Daten beiträgt. Nicht zuletzt müssen die Trainingsdaten dauerhaft und eindeutig zugänglich gemacht sowie im Falle von Korrekturen und Aktualisierungen diese Änderungen durch Versionierung transparent gemacht werden. Vor diesem Hintergrund wird unmittelbar deutlich, dass Bibliotheken hierfür hervorragende Voraussetzungen mitbringen. Schließlich gehören alle der genannten Tätigkeiten zu den Kernaufgaben des Informations- und Datenmanagements in Bibliotheken und Forschungsdatenmanagements in wissenschaftlichen Kontexten, auch wenn dies für den Bereich des maschinellen Lernens nur bedingt zutrifft. Umso mehr könnten Bibliotheken hier ihre Expertise nutzbar machen. So können sie sowohl durch bessere Trainingsdaten für KI als auch durch deren Kuratierung zu einer fortlaufenden Verbesserung und Ausgewogenheit von KI-Modellen beitragen.

Auch die Sammlungen selbst, soweit bereits urheberrechtsfrei und digitalisiert, bieten sich für das Trainieren von KI an. Sie sind über lange Zeiträume von Expert*innen aufgrund ihrer Relevanz für die Sammlung ausgewählt worden. Zudem bilden sie neben der historischen Dimension auch die Vielfalt der Perspektiven in den jeweiligen Wissenschaftsdisziplinen möglichst breit ab. Man kann vermuten, dass beispielsweise OpenAI

2 Mehr dazu siehe https://mmk.sbb.berlin/ [26.02.2024].

3 Mehr dazu siehe https://wiki.dnb.de/display/FNMVE/Netzwerk+maschinelle+Verfahren+in+der+Erschliessung [26.02.2024].

4 Mehr dazu siehe https://pro.europeana.eu/project/ai-in-relation-to-glams [26.02.2024]

5 Mehr dazu siehe http://ai4lam.org/ [26.02.2024]

für das Trainieren einer neuen Version von ChatGPT inzwischen auf Datenpartnerschaften[6] setzt, weil die frei im Internet verfügbaren Datenquellen bereits ausgeschöpft sind. Würden hingegen die in den Kulturerbe-Einrichtungen gesammelten und kuratierten Inhalte als Trainingsdaten für KI zugänglich und nutzbar gemacht, so könnte dies noch erhebliche Fortschritte ermöglichen.

Kulturerbe-Datensätze bringen die Herausforderungen, Schattenseiten und Risiken an die Oberfläche, die sich aus der Verwendung großer Datenmengen für das Training von KI-Anwendungen ergeben. Historisches Text- und Bildmaterial beinhaltet die Sicht der jeweiligen Zeit, die wiederum aus heutiger Bewertung ethisch oder sozial problematisch sein kann. So können beispielsweise Texte des ausgehenden 19. und frühen 20. Jahrhunderts das biologistisch-rassistische Denken der Zeit reproduzieren, im Datenmaterial können diskriminierende oder kolonialistische Positionen und Stereotype enthalten sein. Ein weiteres Beispiel für diese sogenannten »biases«, die sich in den Daten als statistische Verzerrungen niederschlagen, ist die Tatsache, dass bis ins späte 19. Jahrhundert hinein die weit überwiegende Mehrheit der Urheber*innen von Texten Männer waren, zählen sie doch zu denjenigen, die in der Geschichte alphabetisiert und sozial privilegiert waren.

Biases aber finden sich nicht nur in den Daten, sondern werden auch regelmäßig entlang des gesamten *Machine-Learning-Workflows* erzeugt. Ein gutes Beispiel dafür sind die Annotationen, die für maschinelles Lernen aufbereitete Datensätze häufig enthalten. Dies können Klassifikationen sein, etwa die Vergabe von Library of Congress Subject Headings (vgl. Berman, 1971). Es können aber auch Bewertungen durch die Annotator*innen sein, wie ›gut‹, ›schlecht‹, ›okay‹, ›toxisch‹ oder numerische Werte einer Likert-Skala. Solche Annotationen fügen die Zeitgenoss*innen manuell hinzu; daher sind ihnen die Positionalität dieser Menschen oder Klassifikationsvorgaben eingeschrieben. Entsprechend der Sozialisation und des kulturellen Hintergrunds divergieren diese Positionsnahmen und können zu heterogenen Einschätzungen des historischen Materials führen (Santy et al., 2023). Kulturerbe-Einrichtungen zeichnen sich durch ein hohes Bewusstsein für Positionalität aus, denn der im Zeitverlauf erfolgende Wechsel in der Bewertung historischen Materials entspricht der Berufserfahrung der Mitarbeiter*innen. Hiervon könnte die *Machine Learning*-Gemeinschaft lernen: Konsequenterweise sollten auch die Annotationsrichtlinien die Möglichkeit in Betracht ziehen, dass Menschen mit unterschiedlichem sozialem und kulturellem Hintergrund auch zu unterschiedlichen Beurteilungen kommen (Denton et al., 2021) und damit einen Bias aufweisen, der sich wiederum auf die KI-Anwendungen auswirkt.

In der Fachliteratur wird häufig diskutiert, wie sich Biases in Datensätzen vermeiden oder ausgleichen lassen (Blodgett et al., 2020; Birhane, Prabhu und Kahembwe, 2021; Field et al., 2022). Seltener wird hingegen in Betracht gezogen und erörtert, dass auch in den Phasen des Designs und des Trainings des durch das maschinelle Lernen erzeugten Modells Biases eingeführt oder verstärkt werden können. Hierfür kann es verschiedene Ursachen geben. Beispielsweise wird der annotierte Datensatz in der Regel prozessiert, indem die von den Annotator*innen vergebenen Labels validiert und die Urteilerübereinstimmung (*inter-rater reliability* oder *inter-annotator agreement*) berechnet werden. Die

6 Mehr dazu siehe https://openai.com/blog/data-partnerships [26.02.2024].

Wahl des Maßes für die Urteilerübereinstimmung oder des Schwellenwerts, oberhalb dessen die Annotationen verworfen werden, liegt im Ermessen der Forscher*innen. Unter den Expert*innen gibt es keinen Konsens im Hinblick auf den Umgang damit (Cambo und Gergle, 2022). Meist müssen die von den Annotator*innen vergebenen Labels jedoch vereinheitlicht werden, um eine eindeutige Zuordnung der Labels herzustellen, von der der Algorithmus maschinellen Lernens das Ziel ableitet, demgemäß das Modell erstellt wird. Bei dieser Aggregation der Labels wird häufig dasjenige Label benutzt, das in der absoluten oder relativen Mehrheit der Annotator*innen verwendet wurde. In der Konsequenz wird dadurch die Mehrheitsmeinung der Annotator*innen privilegiert, Minderheitenmeinungen werden unterdrückt und ein Bias eingeführt oder verstärkt.

Diese Beispiele zeigen einerseits das Risiko auf, dass ethisch und sozial problematische Inhalte oder auf andere Weise eingeführte Biases durch maschinelles Lernen noch verstärkt werden. Andererseits weisen Kulturerbe-Einrichtungen eine Reihe von Ressourcen und Potenzialen auf, die solchen Risiken vorbeugen oder mögliche statistische Verzerrungen abmildern können. So verfügen Mitarbeiter*innen in Kulturerbe-Einrichtungen über eine ausgezeichnete Kenntnis der Sammlungen in ihren Fachbereichen sowie des analogen und digitalisierten Materials und setzen bei der Auswahl des Materials das vorhandene Expert*innenwissen und institutionell etablierte Verfahren ein (Jo und Gebru, 2020). Bei der Zusammenstellung von Datensätzen wird von den Kurator*innen der Urheberrechtsschutz beachtet oder nur gemeinfreies Material verwendet, eine Vorgehensweise, die gerade bei der Erstellung großer Sprachmodelle nicht immer befolgt wird (Chang et al., 2023). Bibliotheken digitalisieren große Mengen an Dokumenten. Meist erschließen sie sie als Volltexte und stellen sie mitsamt qualitativ hochwertigen Metadaten bereit. Die Bibliotheken machen diese Datenbestände als offene Daten frei und dauerhaft verfügbar, denn dies ist eine Auflage der öffentlichen Hand. Daher sind diese Daten äußerst attraktiv für maschinelles Lernen, beispielsweise für das Training von stochastischen Sprachmodellen. Eine der Schwächen dieser Modelle – das »Halluzinieren« – resultiert nämlich aus ihrem Unvermögen, zwischen Fakten und Fiktion zu unterscheiden (Mahowald et al., 2023), aber auch aus einem Mangel an Daten, deren Vielfalt oder Relevanz. Bibliographische Daten aus Bibliotheken können hier eine verlässliche Grundlage bilden, um zu einer Reduktion in der Häufigkeit der Halluzinationen zu führen. Sie können als externe Daten bereitgestellt werden, um von Sprachmodellen generierte Antworten von höherer Qualität zu gewinnen. Bei der *Retrieval-Augmented Generation*[7] werden dem *Prompt*[8] akkurate Kontexte mitgegeben, etwa über einen *Knowledge Graph*[9], eine (Vektor-)Datenbank oder ein Embedding-Modell. Die Antwort des Sprachmodells wird durch den erweiterten Kontext spezifischer, besser kontrollierbar und aktueller (Gao et al., 2024). Darüber hinaus sind die Bestände an Volltexten auch deshalb qualitativ hochwertig, weil die Verlage, die diese Werke gedruckt haben, für die orthographische, syntaktische und lexikalische Qualität der Werke sorgten. Im Gegensatz da-

7 Retrieval-Augmented Generation bezeichnet eine Technik, bei der durch eine KI generierte Antworten zusätzlich durch Information Retrieval qualifiziert werden.

8 Die Eingabe für ein generatives Sprachmodell.

9 Ein Knowledge Graph ist eine als Graph strukturierte Form der Wissensrepräsentation.

zu stehen Inhalte von sprachlich minderer Qualität, wie sie für Internetnutzer*innen typisch sind, etwa auf Foren oder in sozialen Medien.

Einer Studie zufolge werden hochqualitative Textdaten im Englischen noch vor dem Jahr 2026 nahezu vollständig für Trainingszwecke vorliegen und danach in der Menge nur noch langsam anwachsen (Villalobos et al., 2022) – während für viele andere Sprachen nur wenig oder gar keine KI-Modelle existieren. Dementsprechend werden die von Bibliotheken digitalisierten gemeinfreien Bestände, aber auch Texte, die in *Open Access* verfügbar sind, zukünftig noch im Wert steigen. Schließlich sind Mitarbeiter*innen von Kulturerbe-Einrichtungen mit der Herausforderung vertraut, Beschreibungen von Beständen in standardisierter Form auszuarbeiten und vorzulegen. Dieses Desiderat wurde erst in jüngster Vergangenheit im Hinblick auf Datensätze (Gebru et al., 2021) und *Machine Learning*-Modelle (Mitchell et al., 2019) benannt. Die Aufgabenstellung besteht hier darin, jenen Kontext bereitzustellen, den die möglichen Nutzer*innen dieser Datensätze und Modelle benötigen, um sie aufgabenadäquat einsetzen zu können. Derartige Dokumentationen sollen nicht nur die Biases beschreiben, die den Datensatz charakterisieren, sondern auch die Besonderheiten von Kulturerbe-Datensätzen herausarbeiten: Dazu gehören etwa die vielfältigen Auswahlprozesse, die zur Strukturierung des Datensatzes geführt haben, seine Heterogenität, also die Streuung in Bezug auf Zeit, Ort, Sprachen, soziale Schichten und kulturelle Kontexte, sowie den Anwendungsfall, für den der Datensatz zusammengestellt wurde (Alkemade et al., 2023; Lee, 2023). Derartige die Datenpublikation begleitende Dokumentationen unterstützen nicht nur die *Machine Learning*-Gemeinschaft bei ihrer Arbeit, sondern reduzieren auch das Potenzial, dass Menschengruppen Schaden zugefügt werden könnte, wenn ein bestimmter Datensatz – etwa aus kolonialen Kontexten – dafür benutzt wird, ein *Machine Learning*-Modell zu trainieren.

Die Kulturen im Umgang mit künstlicher Intelligenz unterscheiden sich zwischen Kulturerbe-Einrichtungen und kommerziellen Herstellern deutlich im Hinblick auf Offenheit: Die von Kulturerbe-Einrichtungen entwickelten *Machine Learning*-Modelle werden in der Regel als Open-Source-Dateien veröffentlicht, da sie mit Unterstützung von Fördermitteln erstellt wurden. Ebenso ist die Datengrundlage in aller Regel frei zugänglich verfügbar. Darüber hinaus dürfen Kulturerbe-Einrichtungen urheberrechtlich geschützte Werke, die sich dauerhaft in ihren Beständen befinden, digitalisieren und für *Text und Data Mining* verwenden, oder sie verfügen über eigene Lizenzvereinbarungen mit den Rechteinhaber*innen; dies sind in der Regel Verlage. Metadaten sind auf nationaler (Deutsche Digitale Bibliothek, Archivportal D) sowie auf europäischer Ebene (Europeana) unter einer CC0 Lizenz frei verfügbar. Einige Kulturerbe-Einrichtungen verfügen darüber hinaus über eine geeignete Infrastruktur und die Rechenleistung, die für das aufwändige Training der Modelle notwendig sind. Bei Bündelung der Ressourcen in Kulturerbe-Einrichtungen und weiteren öffentlichen Einrichtungen wie Universitäten und Forschungseinrichtungen wären sie daher wohl auch nicht von Cloud-Diensten abhängig.

Zusammenfassend lässt sich festhalten, dass sich die Kultur künstlicher Intelligenz in Kulturerbe-Einrichtungen nicht nur im Hinblick auf ein hohes Problembewusstsein in Bezug auf Biases in den Daten und im *Machine Learning Workflow* auszeichnet, sondern auch durch die Rollen und Funktionen, die Kulturerbe-Einrichtungen im Hinblick

auf Datenangebote und die Bereitstellung von Modellen übernehmen. Hier setzen Einrichtungen wie die Staatsbibliothek zu Berlin auf die *Open Access*-Publikation von Datensätzen, die sie für spezifische Trainingsaufgaben erstellt haben, und auf die *Open Source*-Bereitstellung von kleinen Modellen, die für klar definierte Anwendungsfälle und den Betrieb in geeigneten Kontexten entwickelt wurden. Dies steht dem aktuellen Trend zur Entwicklung einer allgemeinen künstlichen Intelligenz im Sinne von Allzweck-Anwendungen entgegen, deren Größe und Betrieb erhebliche komputationelle Ressourcen erfordert und die Zentralisierungsbewegung im Bereich KI noch verstärkt (Gray Widder, West and Whittaker, 2023). Kulturerbe-Einrichtungen setzen auf Mensch-Maschine-Kooperationen statt auf das Ersetzen von Menschen durch Maschinen. Und im Gegensatz zu den Werten, die beispielsweise im Bereich *Computer Vision* dominieren (Scheuerman, Hanna und Denton, 2021, S. 373:3), räumen GLAM-Institutionen der Sorgfalt Vorrang vor der Effizienz ein, der Kontextualität Vorrang vor der Universalität, der Positionalität Vorrang vor der Unparteilichkeit, und der Datenarbeit Vorrang vor der Arbeit am Modell.

Literatur

Alkemade, Henk/Claeyssens, Steven/Colavizza, Giovanni/Freire, Nuno/Lehmann, Jörg/Neudecker, Clemens/Osti, Giulia/van Strien, Daniel (2023): Datasheets for Digital Cultural Heritage Datasets, in: *Journal of Open Humanities Data*, Volume 9, S. 1–17, https://doi.org/10.5334/johd.124.

Berman, Sanford (1971): *Prejudices and antipathies. A tract on the LC subject heads concerning people*, Metuchen, N.J.: Scarecrow Pr.

Birhane, Abebe/Prabhu, Vinay Uday/Kahembwe, Emmanuel (2021): *Multimodal datasets: misogyny, pornography, and malignant stereotypes*, https://doi.org/10.48550/ARXIV.2110.01963.

Blodgett, Su Lin/Barocas, Solon/Daumé III, Hal/Wallach, Hanna (2020): Language (Technology) is Power: A Critical Survey of »Bias« in NLP‹, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, S. 5454–5476, https://doi.org/10.18653/v1/2020.acl-main.485.

Brantl, Markus/Schweter, Stefan (2022): Neue Wege der Bildsuche an der Bayerischen Staatsbibliothek, in: *Zeitschrift für Bibliothekswesen und Bibliographie*, Bd. 69, Heft 6, S. 328–337, https://doi.org/10.3196/186429502069646.

Cambo, Scott Allen/Gergle, Darren (2022): Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science, in: *CHI Conference on Human Factors in Computing Systems. CHI ›22: CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, S. 1–19, https://doi.org/10.1145/3491102.3501998.

Chang, Kent K./Cramer, Mackenzie/Soni, Sandeep/Bamman, David (2023): *Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4*, https://doi.org/10.48550/ARXIV.2305.00118.

Colutto, Sebastian/Kahle, Philip/Guenter, Hackl/Muehlberger/Guenter (2019): Transkribus. A Platform for Automated Text Recognition and Searching of Historical Doc-

uments, in: *15th International Conference on eScience (eScience)*, IEEE, S. 463–466, https://doi.org/10.1109/eScience.2019.00060.

Cordell, Ryan (2020): *Machine Learning + Libraries: A Report on the State of the Field*, in: LC Labs, Library of Congress, https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf.

Denton, Remi/Díaz, Mark/Kivlichan, Ian/Prabhakaran, Vinodkumar/Rosen, Rachel (2021): *Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation*, https://doi.org/10.48550/ARXIV.2112.04554.

Ehrmann, Maud/Romanello, Matteo/Najem-Meyer, Sven/Doucet, Antoine/Clematide, Simon (2022): Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents, in: *CEUR-WS*, Vol-3180, https://ceur-ws.org/Vol-3180/paper-83.pdf.

Field, Anjalie/Park, Chan Young/Lin, Kevin Z./Tsvetkov, Yulia (2022): Controlled Analyses of Social Biases in Wikipedia Bios, in: *Proceedings of the ACM Web Conference 2022*, S. 2624–2635, https://doi.org/10.1145/3485447.3512134.

Gao, Yunfan/Xiong, Yun/Gao, Xinyu/Jia, Kangxiang/Pan, Jinliu/Bi, Yuxi/Dai, Yi/Sun, Jiawei/Wang, Meng/Wang, Haofen (2024): *Retrieval-Augmented Generation for Large Language Models: A Survey*, https://doi.org/10.48550/arXiv.2312.10997.

Gebru, Timnit/Morgenstern, Jamie/Vecchione, Briana/Wortman Vaughan, Jennifer/Wallach, Hanna/Daumé III, Hal/Crawford, Kate (2021): Datasheets for Datasets, in: *Communications of the ACM*, Volume 64(12), S. 86–92, https://doi.org/10.1145/3458723.

Gray Widder, David/West, Sarah/Whittaker, Meredith (2023): Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI, in: *SSRN Electronic Journal [Preprint]*, https://doi.org/10.2139/ssrn.4543807.

Jo, Eun Seo/Gebru, Timnit (2020): Lessons from archives: strategies for collecting sociocultural data in machine learning, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* ›20*, Barcelona, Spain: ACM, S. 306–316, https://doi.org/10.1145/3351095.3372829.

Kasprzik, Anna (2020): Putting Research-based Machine Learning Solutions for Subject Indexing into Practice, in: *Proceedings of the Conference on Digital Curation Technologies (Qurator 2020)*, CEUR-WS Vol. 2535, https://ceur-ws.org/Vol-2535/.

Kasprzik, Anna (2023): Aufbau eines produktiven Dienstes für die automatisierte Inhaltserschließung an der ZBW: Ein Status- und Erfahrungsbericht, in*: o-bib. Das offene Bibliotheksjournal*, Bd. 10, Heft 1, S. 1–13, https://doi.org/10.5282/o-bib/5903.

Kiessling, Benjamin/Tissot, Robin/Stokes, Peter/Stökl Ben Ezra, Daniel (2019): eScriptorium: An Open Source Platform for Historical Document Analysis, in: *International Conference on Document Analysis and Recognition Workshops (ICDARW)*, IEEE, https://doi.org/10.1109/ICDARW.2019.10032.

Lee, Benjamin C.G./Weld, Daniel S. (2020): Newspaper Navigator: Open Faceted Search for 1.5 Million Images, in: *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, S. 120–122, https://doi.org/10.1145/3379350.3416143.

Lee, Benjamin Charles Germain (2023): The »Collections as ML Data« Checklist for Machine Learning & Cultural Heritage, in: *Journal of the Association for Information Science and Technology*, S. 1–22, https://doi.org/10.1002/asi.24765.

Mahowald, Kyle/Ivanova, Anna A./Blank, Idan A./Kanwisher, Nancy/Tenenbaum, Joshua B./Fedorenko, Evelina (2023): *Dissociating language and thought in large language models*, https://doi.org/10.48550/ARXIV.2301.06627.

Mitchell, Margaret/Wu, Simone/Zaldivar, Andrew/Barnes, Parker/Vasserman, Lucy/Hutchinson, Ben/Spitzer, Elena/Raji, Inioluwa Deborah/Gebru, Timnit (2019): Model Cards for Model Reporting, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* ›19*, Atlanta GA USA: ACM, S. 220–229, https://doi.org/10.1145/3287560.3287596.

Mödden, Elisabeth/Schöning-Walter, Christa/Uhlmann, Sandro (2018): Maschinelle Inhaltserschließung in der Deutschen Nationalbibliothek. Breiter Sammelauftrag stellt hohe Anforderungen an die Algorithmen zur statistischen und linguistischen Analyse, in: *BuB-Forum Bibliothek und Information*, S. 30–35.

Neudecker, Clemens/Baierer, Konstantin/Federbusch, Maria/Boenig, Matthias/Würzner, Kay-Michael/Hartmann, Volker/Herrmann, Elisa (2019): OCR-D: An end-to-end open source OCR framework for historical printed documents, in: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019)*, ACM, NY, USA, S. 53–58, https://doi.org/10.1145/3322905.3322917.

Rehm, G. et al. (2020): QURATOR: innovative technologies for content and data curation, in: *Proceedings of the Conference on Digital Curation Technologies (Qurator 2020)*, CEUR-WS Vol. 2535, https://ceur-ws.org/Vol-2535/paper_17.pdf.

Reul, Christian/Springmann, Uwe/Wick, Christoph/Puppe, Frank (2018): State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines, *CoRR*. https://doi.org/10.48550/arXiv.1810.03436.

Santy, Sebastin/Liang, Jenny T./Le Bras, Ronan/Reinecke, Katharina/Sap, Maarten (2023): *NLPositionality: Characterizing Design Biases of Datasets and Models*, https://doi.org/10.48550/arXiv.2306.01943.

Scheuerman, Morgan Klaus/Hanna, Alex/Denton, Emily (2021): Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development, in: *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), S. 1–37, https://doi.org/10.1145/3476058.

Suominen, Osma (2019): Annif: DIY automated subject indexing using multiple algorithms, in: *LIBER Quarterly: The Journal of the Association of European Research Libraries*, Bd. 29, Heft 1, S. 1–25, https://doi.org/10.18352/lq.10285.

Villalobos, Pablo/Ho, Anson/Sevilla, Jaime/Besiroglu, Tamay/Heim, Lennart/Hobbhahn, Marius (2022): *Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning*, https://doi.org/10.48550/ARXIV.2211.04325.