Digital Curation and AI

Opportunities and Risks for Cultural Heritage Institutions

Clemens Neudecker

> ChatGPT prompt: What is the potential of combining AI and cultural heritage?
>
> The combination of AI and cultural heritage has the potential to revolutionize the way that cultural heritage is studied, preserved, and experienced. By harnessing the power of AI, cultural heritage organizations can ensure that the legacy of our ancestors is preserved for future generations and that its significance is understood and appreciated by all. AI-powered tools can assist in the analysis and interpretation of large amounts of data and help to reveal new insights and connections that might otherwise have gone unnoticed. This can lead to a deeper understanding of cultural heritage and its significance.
>
> ChatGPT prompt: What can cultural heritage contribute to AI?
>
> Cultural heritage can contribute to AI in several ways. Firstly, cultural heritage provides a rich source of data and information for AI systems to learn from. By training AI models on vast collections of cultural heritage artifacts and information, AI can be made more sophisticated and able to tackle complex tasks, such as image recognition, natural language processing, and data analysis. This can help to advance the field of AI and lead to new and innovative applications. Secondly, cultural heritage can provide a unique perspective on the world and help AI to understand the complexities of human societies, cultures, and beliefs.

The two paragraphs above were generated by ChatGPT[1] from OpenAI and demonstrate what the neural network that lies behind it has already learned about artificial intelligence in the context of cultural heritage. What is interesting here, however, is that the answer to the initial prompt focuses solely on the potential of applying AI tools 'to assist in the analysis and interpretation of large amounts of data', whereas only the second, follow-up prompt, which explicitly shifts the focus towards the possible contribution of cultural heritage to AI, reveals that there are also significant

1 https://chat.openai.com/chat (all URLs here accessed in June 2023).

opportunities for AI to benefit from cultural heritage. This paper will present two examples for the adoption of AI at the Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (Berlin State Library, SBB[2]) and highlight some of the recent criticisms of data practices in the domain of AI, but also propose various ideas and opportunities for digital curators and cultural heritage institutions regarding what they can contribute to AI.

Opportunities

Numerous use cases, from text recognition to image analysis or the classification and contextualization of cultural heritage content demonstrate how digitization and curation in galleries, libraries, archives, and museums (GLAM), can benefit from AI (Markus/Neudecker/Isaac 2021; Cordell 2020). But digitization too often ends with merely capturing images (scanning). In order to support more advanced and diverse use of digitized collections, the digitized objects need to be processed further so as to analyse, expose, and enrich the content contained within the scans. For example, text recognition (in other words, optical character recognition [OCR] or handwritten text recognition) extracts the text from a scan to make it machine-readable, layout analysis can structure the various types of content on a scanned page into different sections like text, images, tables, et cetera, and methods from the domain of natural language processing (NLP) can be utilized to extract information (for instance, named entities) from the text or to enrich it semantically (for example, with links to a knowledge base), to name just a few applications. Altogether, the abovementioned processes can be useful in creating machine-readable corpora or datasets from digitized collections, which can in turn again help improve machine learning methods and models (Lee 2022). AI methods and models have provided significant improvements for all of the above applications, for instance, for recognizing text in historical prints (Wick/Reul/Puppe et al. 2018) or handwritten documents (Muehlberger/Seaward/Terras et al. 2019), for document layout analysis (Shen/Zhang/Dell et al. 2021 and Huang/Lv/Cui et al. 2022), for content-based retrieval (Brantl/Schweter 2022), or in the area of named entity recognition and linking (Ehrmann et al. 2020; 2022). In the following section, two example projects at the SBB in the area of AI for digitized cultural heritage will be discussed along with their main contributions and outputs.

2 https://staatsbibliothek-berlin.de/.

QURATOR: Automated Curation Technologies for Digitized Cultural Heritage

The SBB was first able to explore the capabilities of various AI and/or deep learning methods for digitized collections in the QURATOR project.[3] QURATOR—Curation Technologies (Rehm/Bourgonje/Hegele et al. 2020) was a research project funded by the German Federal Ministry of Research and Education (BMBF) from 2018 to 2021, based in the Berlin/Brandenburg metropolitan region. The consortium of ten project partners from research and industry pooled their expertise in areas such as language technologies, knowledge technologies, artificial intelligence, and machine learning. The project's main goal was developing a sustainable technology platform that supports knowledge workers in various industries. The platform is intended to simplify the curation of digital content and accelerate it dramatically. AI technologies are integrated into curation technologies and curation workflows in the form of industry solutions covering the entire lifecycle of content curation. The solutions developed focus on curation services for the sectors of culture, media, health care, and industry.

Within QURATOR, the SBB was responsible for the subproject Automated Curation Technologies for Digitized Cultural Heritage. The SBB aims to digitize all its copyright-free historical collections and to make them available on the web for use by researchers. To achieve this goal, the SBB has developed a variety of applications that leverage deep learning techniques.

First, to perform various statistical analyses and clusters of the collections of the SBB, a tool[4] was developed that converts bibliographical data from the library catalogue and digitized collection metadata into a format that is suitable for machine learning research and development, in this case, a Python Pandas DataFrame. Most of the applications created as part of QURATOR target the digitized collections of the SBB. While roughly 200,000 historical documents have been digitized[5] thus far, only a small fraction (14 per cent) already have a searchable digital full text in addition to the scanned image of a page. Accordingly, several tools were developed for a typical text digitization workflow. Any image of a document page should be optimized for OCR processing, which can include various image processing tasks like page cropping (removing borders), deskewing (rotating the page to be horizontally aligned), dewarping (straightening text lines), and scaling (up- or downscaling of the image). The next step is very crucial—especially in the case of documents with complex layouts. In document layout analysis, the aim is to detect any different content regions in the document (for instance, text blocks, tables, illustrations, et cetera)

3 https://qurator.ai/.
4 https://github.com/qurator-spk/mods4pandas.
5 https://digital.staatsbibliothek-berlin.de/.

and to segment (that is, determine their boundaries as pixel coordinates) and classify them (that is, assign the correct content type). Another extremely important task is detecting the correct sequence of the various content regions (reading order detection).

For document layout analysis, the SBB developed the Eynollah tool.[6] It is based on a ResNet-U-Net encoder-decoder architecture and utilizes multiple trained models in combination with heuristics. It can detect up to eight distinct content classes (background, separator, text, table, image, header/heading, marginalia, drop capital/initial). The overall goal is to feed the output into an OCR model as text line images for text recognition purposes. Since current state-of-the-art OCR models operate only on binarized images of text lines, another image processing model was needed to provide the functionality for binarization. This was done by training a hybrid CNN-transformer model[7] on data from past DIBCO competitions[8] and other freely available binarization datasets such as the Palm Leaf Heritage Dataset (Burie et al. 2016) and documents from the Persian Heritage Image Binarization Competition (Ayatollahi/Nafchi 2013). For the text recognition, the SBB opted for the Calamari (Wick/Reul/Puppe et al. 2018) system, for which a model was trained using the GT4HistOCR (Springmann/Reul/Dipper et al.2018) dataset. Experiments conducted at the SBB show that combining Eynollah with Calamari can reduce the character error rate in historical printed documents to roughly 1 to 2 per cent. Additionally, the Calamari OCR engine and model were integrated[9] with the OCR-D framework (Neudecker/Baierer/Federbusch 2019).

Since a few OCR errors may remain in a recognized text, the extent to which deep learning can be used for automated post-correction of OCR results was also explored. For this, the decision was made to follow the example of machine translation, which meant that a model[10] was trained for the task of translating an OCR result containing errors into a perfectly correct text (Schaefer/Neudecker 2020). Finally, to complete the text recognition pipeline, a tool[11] for Ground Truth-based quality evaluation of OCR results was implemented. Compared with other available tools and metrics for OCR evaluation (Neudecker/Baierer/Gerber et al. 2021), this tool is optimized for large documents with consideration of reading order, and provides a visual alignment of OCR and Ground Truth to easily spot errors.

Furthermore, two additional strands of work addressed the reuse of outputs from the layout analysis and text recognition processes. Since document layout anal-

6 https://github.com/qurator-spk/eynollah.

7 https://github.com/qurator-spk/sbb_binarization.

8 https://dib.cin.ufpe.br/#!/datasets.

9 https://github.com/qurator-spk/ocrd_calamari.

10 https://github.com/qurator-spk/sbb_ocr_postcorrection.

11 https://github.com/qurator-spk/dinglehopper.

ysis produces information on image regions within pages, these images were extracted based on their coordinates and analysed for similarities using an image similarity model[12] based on ImageNet (Deng/Dong/Socher et al. 2009). This facilitates content-based image retrieval, in other words, by uploading an image, it is possible to automatically retrieve the most similar images from within the digitized collections of the SBB.

Most importantly, at the end of the text recognition workflow, there should be a digital and structured text with a high accuracy, which can then serve as a suitable input for various natural language processing (NLP) tasks. Within NLP, the SBB focussed on a few common tasks that are supposed to be particularly in demand by users and can also benefit information retrieval in the digitized collections, namely, named entity recognition (NER) and named entity disambiguation and linking (EL). The NER system[13] developed by the SBB is based on Bidirectional Encoder Representations from Transformers, or BERT (Devlin/Chang/Lee et al. 2019). To adapt the original BERT model for historical texts containing OCR errors, an unsupervised pre-training was done using a selection of 2,333,647 German-language pages from the SBB's digitized collections, followed by additional supervised training on openly available gold-standard data for NER (Labusch/Neudecker/Zellhöfer 2019). Furthermore, to disambiguate the entities recognized and link them to authority data (in this case, Wikidata QIDs), knowledge bases were constructed using Wikipedia and Wikidata for German, French, and English, and a purpose-trained BERT context disambiguation model was developed (Labusch/Neudecker 2020) that decides for a given entity whether and which QID should be linked, based on the local context and a comparison with the knowledge bases.

Since there can be no machine learning project without data annotation, a very simple annotation tool[14] was also produced. The tool is a simple browser-based HTML and Javascript application which can use an IIIF-compliant Image API. It operates on a simple, tab separated values file and can be used either to annotate a given text with named entities or to transcribe or correct the text to create Ground Truth for OCR purposes.

Mensch.Maschine.Kultur: AI for Digital Cultural Heritage

A second research project in the area of AI for digital cultural heritage at the SBB received funding from the German Federal Government Commissioner for Culture and the Media (BKM) and is currently being carried out by the SBB. The project

12 https://github.com/qurator-spk/sbb_images.

13 https://github.com/qurator-spk/sbb_ner.

14 https://github.com/qurator-spk/neat.

Mensch.Maschine.Kultur – Künstliche Intelligenz für das digitale kulturelle Erbe[15] (Human.Machine.Culture—Artificial Intelligence for Digital Cultural Heritage) consists of four subprojects that pursue different objectives in coordination with each other and combines them with suitable AI procedures. The project will take place over three years (2022–25) and feature a mix of 50 per cent research activities and 50 per cent adaptation and implementation of deep learning prototypes for productive use in the library.

Subproject 1, 'Intelligent methods for generic document analysis', provides AI methods for document analysis, with the aim of obtaining high-quality full texts and structural data extracted from the variety of information contained in the digitized collections (text, image, layout). This work package therefore goes beyond the recognition of texts, and also separates image elements and analyses the layout to facilitate the structured representation of texts such as those in newspapers and magazines.

Subproject 2, 'Image analysis tools for digital cultural heritage', extends the work begun in the predecessor project QURATOR on image similarity searches through recognizing, extracting, and classifying digital image content.

Subproject 3, 'AI-supported content analysis and subject indexing', assists the experts in the specialist departments of the SBB with semi-automated procedures for subject indexing (based on Suominen 2019; Kasprzik 2020), and systematically incorporates their expertise. Furthermore, fully automated procedures for the recognition of entities such as persons, places, and organizations will support searches within material from the digitized collections in the library's discovery system. In order to predict any ethical issues arising from the use of AI for this, an Ethical Foresight Analysis (Bubinger/Dinneen 2021) has been conducted in collaboration with a researcher from the Humboldt University Berlin.

Subproject 4, 'Data provision and curation for AI', bundles and documents data that have been specifically prepared for research and use in AI contexts, and makes these datasets publicly available for subsequent use. In addition, guidelines on how to identify and deal with qualitatively or ethically problematic holdings and content are being developed in collaboration with the broader community.

Risks

Against the backdrop of these promising developments, it must also be made clear that there are certain risks in applying AI technologies and models developed predominantly by the private sector to cultural data without fully understanding the implications.

15 https://mmk.sbb.berlin/.

Most current AI technologies are based on neural networks that are being trained to derive stochastic models from large amounts of data. Therefore, the capabilities of the AI models are highly dependent on the type and quality of the data they were trained on. But, too often, it is not clear what the source of the data that AI models were trained on is, and what perhaps undesirable implications and biases the model draws from the data. Many questions arise when investigating how data for training AI models is assembled: What are the sources from which the data have been obtained? Have selection criteria been used, and if so, what are they? Are questions regarding data quality addressed? If the data contain problems, are they known and documented? Is there someone to reach out to in the case of issues?

Jo and Gebru (2020) describe a laissez-faire attitude towards dataset development in AI: rather than collecting and curating datasets with care and intentionality, practitioners in the AI field frequently have adopted an 'anything-goes' approach, rather as if, as one data scientist put it, 'if it is available to us, we ingest it' (Holstein/Wortman Vaughan/Daumé III et al. 2019). A comprehensive survey of dataset development and use in AI was produced by Paullada et al. (2020). It found numerous issues with data collection and advocated for a more cautious and thorough understanding of the data. Furthermore, when the source data is annotated, tagged, or classified, this task is often outsourced to unskilled cheap labour, where no attribution is given for this—in fact interpretative—work, and annotation standards and guidelines are not in place or at least are not made publicly available.

Artists were among the first to raise public attention by exposing flaws in AI datasets. In their ImageNet Roulette, Paglen and Crawford (2019) allowed people to upload a selfie and then classified it based on ImageNet (Deng/Dong/Socher et al. 2009), one of the largest image databases with image classifications. The resulting classifications revealed offensive or derogatory statements, such as classifying an image of a pregnant person as 'lazy'. A large part of the ImageNet dataset was subsequently removed from public access. Prabu and Birhane (2020) investigated ImageNet further and still found many harmful representations of women.

Another example is the work of Harvey and LaPlace (2021), who examined publicly available datasets for face recognition. They created the website exposing.ai, which offers a Flickr reverse image that enables users to check if any of their images have been used in face recognition datasets. It was found that many datasets used images of people collected from Flickr without their providing consent for this use. In the case of VGGFace2, a widely used dataset for face detection, 3.3 million images of 9,000 individuals were used to train the model, but none of the individuals provided consent and many of the images were copyrighted. A large portion of the VGGFace2 datasets was subsequently pulled from the web—without any explanation or justification being provided.

But not only image-based AI can suffer from bad quality or biased training data. Large language models like BERT, GPT, et cetera are also trained on massive quantities of text that are scraped from publicly available online sources like search engines, social media platforms, et cetera with no transparent policies regarding selection. At Shutterstock, a 'List of Dirty Naughty Obscene and Otherwise Bad Words (LDNOOBW)' with more than 400 entries for the English language and numerous other languages has been created and released on GitHub.[16] When creating the Colossal Clean Crawled Corpus, AI researchers at Google reported using the list to filter the web pages included in the corpus (Raffel/Shazeer/Roberts, Adam 2020). Not only is it highly problematic that such filtering plainly also removes valid educational or medical content that includes, for instance, sexual terms from the LDNOOBW, which the AI model will fail to learn, but it also begs the question of what else might get filtered out or excluded from the training data without our knowledge. And, more importantly, who determines what gets filtered out or removed?

Cultural Heritage Institutions and Digital Curation for AI

Against this backdrop, what and how can cultural heritage institutions contribute to AI? The following section will look at two areas where cultural heritage might have something to offer for the research and development of AI, data, and curation.

An obvious way for cultural heritage to contribute to AI is by providing data. Thanks to ongoing mass digitization, large quantities of scanned documents, images, and other cultural objects are becoming available digitally. But it must not be overlooked that the collections held in cultural heritage institutions also contain content that is problematic from a social, legal, and/or ethical perspective. Due to copyright, a vast majority of the material that is digitized is of a historical nature, and can therefore contain content that expresses views that are obsolete or discriminating, for instance, since they represent a racist, colonial, or Western perspective, or lack representation of women and other marginalized groups (Manžuch 2017).

What is therefore needed to turn digitized cultural heritage data into collections as data for AI are quality checks based on transparent standards, contextualization, and active curation by domain experts. AI itself can also be used for this, for example, to identify and flag potentially problematic content in cultural heritage collections for review by experts, as done by the Contentious Contexts Corpus (Brate/Nestorov/Vogelmann 2021), in which a classifier was developed for automated detection of contentiousness in historical Dutch newspapers. In addition, cultural heritage institutions need to create fundamentally better and more suitable ways to distribute and curate digitized cultural heritage as data in AI contexts, for

16 https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words.

instance, by providing their data in the formats and on the platforms that are dominant in the AI community (Darby/Coleman/Engel et al. 2022). A promising example for shared benefits for cultural heritage and AI is the HuggingFace BigScience workshop, which brought together experts to curate a dataset for training a multilingual language model (BLOOM) and included a hackathon BigLAM[17] in order to incorporate datasets and AI models from the cultural heritage sector into the HuggingFace platform.

Another not as immediately obvious but potentially even more meaningful way to contribute to AI lies in the curation practices that are at the core of cultural heritage institutions. Cultural heritage institutions have a long tradition and great expertise in curation based on established quality standards. Curators are subject experts trained in various domains to provide contextualization and also pay a great deal of attention to detail. In addition, as public and non-profit organizations, cultural heritage institutions provide a higher level of transparency, trust, neutrality, and reliability. Values like these, as already embedded in many cultural heritage institution policies and applied by cultural heritage practitioners, are paramount for a responsible curation practice. Furthermore, cultural heritage institutions tend to show greater sensitivity and are more rigorous with regard to sensitivity to data sovereignty and to adhering to laws regulating privacy and personal data protection. Last but not least, research data repositories in cultural heritage institutions have typically been established based on long-term preservation policies that also provide ways to track changes and updates to datasets or even the de-publication of datasets transparently and more reliably than in the case of data residing somewhere on the web or even on some torrent site.

On the other hand, the curatorial practices in cultural heritage institutions also need to be questioned and updated on an ongoing basis in order to develop awareness and reflect on and resolve biases and issues. The collections and classification systems used for them are themselves products of particular biases that require identification, documentation, and communication (Coleman 2020). The international GLAM community has begun tackling this through collaborative communities such as the EuropeanaTech Task Force on AI in relation to GLAMs, or the international AI4LAM community. Work is underway in these groups to develop standards and recommendations for the curation of cultural heritage as datasets for AI while considering legal, social, and ethical aspects, such as by adapting the 'Datasheets for Digital Cultural Heritage' based on Gebru/Morgenstern/Veccione et al. 2021 and 'Model Cards' based on Mitchell/Wu/Zaldivar et al. 2019. In summary, cultural heritage institutions can make important contributions to AI, either through the provision of data or by means of their curation practices. But in order to do so, they must invest further in responsible curation and data stewardship that

17 https://huggingface.co/biglam.

cares about quality, transparency, and awareness of biases and ethical issues in data and create good examples of open, high-quality, and actively curated datasets that are useful in further advancing AI (Neudecker 2022).

Conclusion

To conclude, the advances in AI are really in the position to unlock the content in digitized cultural heritage, but practices of responsible data collection and curation need to be adopted in the domain of AI so as to reduce the risks and implications of issues in AI training data. While cultural heritage institutions can build on their long tradition and experience with curation, these also need updating in order to address the concerns and issues that arise when their data is used within AI. Digital curation that cares about quality, transparency, and awareness of biases, from which good datasets can ultimately emerge, can benefit the development and use of AI in science and industry as well as society as a whole. Or as Paullada et al. (2021, p. 10) put it:

> In closing, we advocate for a turn in the culture towards carefully collected datasets, rooted in their original contexts, distributed only in ways that respect the intellectual property and privacy rights of data creators and data subjects, and constructed in conversation with the relevant scientific and scholarly fields required to create datasets that faithfully model tasks and tasks which target relevant and realistic capabilities. Such datasets will undoubtedly be more expensive to create, in time, money and effort, and therefore smaller than today's most celebrated benchmarks. This, in turn, will encourage work on approaches to machine learning (and to artificial intelligence beyond machine learning) that go beyond the current paradigm of techniques idolizing scale. Should this come to pass, we predict that machine learning as a field will be better positioned to understand how its technology impacts people and to design solutions that work with fidelity and equity in their deployment contexts.

References

Ayatollahi, Seyed Morteza/Nafchi, Hossein Ziaei (2013). Persian Heritage Image Binarization Competition (PHIBC 2012). Proceedings of the 1st Iranian Conference on Pattern Recognition and Image Analysis (PRIA), 1–4. https://doi.org/10.1109/PRIA.2013.6528442 (all URLs here accessed in August 2023).

Brantl, Markus/Schweter, Stefan (2022). Neue Wege der Bildsuche an der Bayerischen Staatsbibliothek. Zeitschrift für Bibliothekswesen und Bibliographie 69 (6), 328–37. https://doi.org/10.3196/186429502069646.

Brate, Ryan/Nestorov, Andrei/Vogelmann, Valentin et al. (2021). Capturing Contentiousness: Constructing the Contentious Terms in Context Corpus. Proceedings of the 11th Knowledge Capture Conference, 17–24. https://doi.org/10.1145/3460210.3493553.

Bubinger, Helen/Dinneen, Jesse David (2021). Actionable Approaches to Promote Ethical AI in Libraries. Proceedings of the Association for Information Science and Technology 58 (1), 682–84. https://doi.org/10.1002/pra2.528.

Burie, Jean-Christophe/Coustaty, Mickael/Hadi, Setiawan et al. (2016). ICFHR2016 Competition on the Analysis of Handwritten Text in Images of Balinese Palm Leaf Manuscripts. Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 596–601. https://doi.org/10.1109/ICFHR.2016.0114.

Coleman, Catherine Nicole (2020). Managing Bias When Library Collections Become Data. International Journal of Librarianship 5 (1), 8–19. https://doi.org/10.23974/ijol.2020.vol5.1.162.

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

Crawford, Kate/Paglen, Trevor (2021). Excavating AI: The Politics of Images in Machine Learning Training Sets. AI & Society 36 (4), 1105–16. https://doi.org/10.1007/s00146-021-01162-8.

Darby, Andrew/Coleman, Catherine Nicole/Engel, Claudia et al. (2022). AI Training Resources for GLAM: A Snapshot. arXiv:2205.04738. https://doi.org/10.48550/arXiv.2205.04738.

Deng, Jia/Dong, Wei/Socher, Richard et al. (2009). ImageNet: A Large-Scale Hierarchical Image Database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–55. https://doi.org/10.1109/CVPR.2009.5206848.

Devlin, Jacob/Chang, Ming-Wei/Lee, Kenton et al. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019, 4171–86. Available online at https://aclanthology.org/N19-1423.pdf.

Ehrmann, Maud/Romanello, Matteo/Flückiger, Alex et al. (2020). Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. CEUR Workshop Proceedings, 2696. Available online at https://ceur-ws.org/Vol-2696/paper_255.pdf.

Ehrmann, Maud/Romanello, Matteo/Najem-Meyer, Sven et al. (2022). Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. CEUR Workshop Proceedings, 3180. Available online at https://ceur-ws.org/Vol-3180/paper-83.pdf.

Gebru, Timnit/Morgenstern, Jamie/Veccione, Briana et al. (2021). Datasheets for Datasets. Communications of the ACM 64 (12), 86–92. https://doi.org/10.1145/3458723.

Harvey, Adam/LaPlace, Jules (2021). Exposing.ai. Available online at https://exposing.ai/.

Holstein, Kenneth/Wortman Vaughan, Jennifer/Daumé III, Hal et al. (2019). Improving Fairness in Machine Learning Systems: What do Industry Practitioners Need? Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–16. https://doi.org/10.1145/3290605.3300830.

Huang, Yupan/Lv, Tengchao/Cui, Lei et al. (2022). LayoutLMv3: Pre-Training for Document AI with Unified Text and Image Masking. Proceedings of the 30th ACM International Conference on Multimedia, 4083–91. https://doi.org/10.1145/3503161.3548112.

Jo, Eun So/Gebru, Timnit (2020). Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 306–16. https://doi.org/10.1145/3351095.3372829.

Kasprzik, Anna (2020). Putting Research-Based Machine Learning Solutions for Subject Indexing into Practice. CEUR Workshop Proceedings, 2535. Available online at https://ceur-ws.org/Vol-2535/paper_1.pdf.

Labusch, Kai/Neudecker, Clemens/Zellhöfer, David (2019). BERT for Named Entity Recognition in Contemporary and Historical German. Proceedings of the 15th Conference on Natural Language Processing, 8–11. Available online at https://konvens.org/proceedings/2019/papers/KONVENS2019_paper_4.pdf.

Labusch, Kai/Neudecker, Clemens (2020). Named Entity Disambiguation and Linking on Historic Newspaper OCR with BERT. CEUR Workshop Proceedings, 2696. Available online at https://ceur-ws.org/Vol-2696/paper_163.pdf.

Lee, Benjamin Charles Germain (2022). The 'Collections as ML Data' Checklist for Machine Learning & Cultural Heritage. arXiv:2207.02960. https://doi.org/10.48550/arXiv.2207.02960.

Manžuch, Zinaida (2017). Ethical Issues in Digitization of Cultural Heritage. Journal of Contemporary Archival Studies 4 (2). Available online at https://elischolar.library.yale.edu/jcas/vol4/iss2/4/.

Markus, Gregory/Neudecker, Clemens/Isaac, Antoine et al. (2021). AI in Relation to GLAMs Task Force Report. Available online at https://pro.europeana.eu/project/ai-in-relation-to-glams.

Mitchell, Margaret/Wu, Simone/Zaldivar, Andrew et al. (2019). Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), 220–29. https://doi.org/10.1145/3287560.3287596.

Muehlberger, Guenter/Seaward, Louise/Terras, Melissa et al. (2019). Transforming Scholarship in the Archives through Handwritten Text Recognition. Journal of Documentation 75 (5), 954–76. https://doi.org/10.1108/JD-07-2018-0114.

Neudecker, Clemens/Baierer, Konstantin/Federbusch, Maria et al. (2019). OCR-D: An End-to-End Open Source OCR Framework for Historical Printed Documents. Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, 53–58. https://doi.org/10.1145/3322905.3322917.

Neudecker, Clemens/Baierer, Konstantin/Gerber, Mike et al. (2021). A Survey of OCR Evaluation Tools and Metrics. Proceedings of the 6th International Workshop on Historical Document Imaging and Processing, 13–18. https://doi.org/10.1145/3476887.3476888.

Neudecker, Clemens (2022). Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries. CEUR Workshop Proceedings, 3234. Available online at https://ceur-ws.org/Vol-3234/paper2.pdf.

Paullada, Amandalynne/Raji, Inioluwa Deborah/Bender, Emily M. et al. (2021). Data and its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research. Patterns 2 (11). https://doi.org/10.1016/j.patter.2021.100336.

Prabhu, Vinay Uday/Birhane, Abeba (2020). Large Image Datasets: A Pyrrhic Win for Computer Vision? 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 1536–46. https://doi.org/10.1109/WACV48630.2021.00158.

Raffel, Colin/Shazeer, Noam/Roberts, Adam et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. The Journal of Machine Learning Research 21 (1), 5485–51. Available online at https://jmlr2020.csail.mit.edu/papers/volume21/20-074/20-074.pdf.

Rehm, Georg/Bourgonje, Peter/Hegele, Stefanie et al. (2020). Qurator: Innovative Technologies for Content and Data Curation. CEUR Workshop Proceedings, 2535. Available online at https://ceur-ws.org/Vol-2535/paper_17.pdf.

Schaefer, Robin / Neudecker, Clemens (2020). A Two-Step Approach for Automatic OCR Post-Correction. Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 52–57. Available online at https://aclanthology.org/2020.latechclfl-1.6/.

Shen, Zejiang/Zhang, Ruochen/Dell, Melissa et al. (2021). LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. arXiv:2103.15348. https://doi.org/10.48550/arXiv.2103.15348.

Springmann, Uwe/Reul, Christian/Dipper, Stefanie et al. (2018). Ground Truth for Training OCR Engines on Historical Documents in German Fraktur and Early Modern Latin. Journal for Language Technology and Computational Linguistics 33 (1), 97–114. https://doi.org/10.21248/jlcl.33.2018.220.

Suominen, Osma (2019). Annif: DIY Automated Subject Indexing Using Multiple Algorithms. LIBER Quarterly: The Journal of the Association of European Research Libraries, 29 (1), 1–25. https://doi.org/10.18352/lq.10285.

Wick, Christoph/Reul, Christian/Puppe, Frank (2018). Calamari—A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. Digital Humanities Quarterly 14 (2). arXiv:1807.02004. https://doi.org/10.48550/arXiv.1807.02004.