Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries

Clemens Neudecker[1]

[1]Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Potsdamer Straße 33, 10785 Berlin, Germany

Abstract

Artificial Intelligence and Machine Learning offer enormous potential for applications in the digitization and digital curation of cultural heritage. But cultural heritage institutions have also produced large amounts of digital data that can be suitable to improve AI methods and models. At the same time there are problems and issues with data used in AI industry and research, which frequently lack quality curation and introduce or reinforce biases. What are the main obstacles for reuse of digitized cultural heritage as data for AI, and what can libraries with their quality awareness and long established practices and competencies of curation contribute to the field of AI?

Keywords

cultural heritage, libraries, digitization, curation, artificial intelligence, machine learning, dataset

1. Introduction

Methods and models from the fields of artificial intelligence and machine learning (AI) promise enormous potential for the (semi-)automated curation of digitized cultural heritage in libraries, archives, and museums, as well as for the computational analysis of cultural heritage data such as in the digital humanities. Some examples from digital libraries that illustrate the possibilities include text recognition (OCR) for historical printed documents [1] and even handwriting [2], where AI is now enabling the near-perfect recognition of text from historical documents that were previously highly problematic; new search and browse functionalities resulting from image detection, classification and similarity analysis in digitized cultural heritage sources with the help of AI [3], [4]; the refinement of unstructured text with methods from Natural Language Processing (NLP) such as Named Entity Recognition (NER) and Entity Linking (EL) [5], which e.g. allow for an easier enrichment and contextualization of digitized content through knowledge bases and also opens up new ways for searching and browsing (e.g. by name or place) - but also more traditional library tasks like subject indexing can benefit, such as by gains in efficiency and quality through recommendations and normalizations generated by AI [6], [7]. Increasingly, projects like *Qurator* [8] or *Living with Machines* [9] are demonstrating what is already achievable with AI in the area of cultural heritage digitization and analysis, semantic enrichment, and digital curation, even when working with complex and messy historical sources.

Qurator 2022: 3rd Conference on Digital Curation Technologies, September 19-23, 2022, Berlin, Germany
clemens.neudecker@sbb.spk-berlin.de (C. Neudecker)
https://staatsbibliothek-berlin.de/ (C. Neudecker)
0000-0001-5293-8322 (C. Neudecker)

2. Background

For the training, fine-tuning, and evaluation of AI methods and models, suitably large-scale data are a necessary precondition. However, digital and freely reusable datasets of relevant size, quality and diversity, especially for the historical and culture domain, are still sparse. Here, digitized cultural heritage can potentially help to fill a current gap. But it is important to ensure that the cultural heritage collections (and their metadata) that are being digitized are also made available in the appropriate ways for use in the further development of AI technologies, with appropriate documentation, and on platforms suitable for this purpose. Making cultural heritage collection available in ways suitable for the AI community can in turn trigger new offerings also for those who do not themselves participate directly in AI.

Another important factor in the provision of digitized cultural heritage for AI research (and beyond) is the responsible curation of such data. The ongoing documentation, contextualization and, when necessary, updating and versioning is a desideratum of many datasets currently widely used in AI research; on the other hand, libraries in particular have the competencies and established processes for curation and a high level of quality awareness in this regard. This curation should also extend to include awareness for the identification and appropriate treatment of problematic content in cultural heritage collections with regard not only to quality or copyright, but ethical and social biases and issues in the data. This can lead to enrichments and better descriptions of the holdings which are also useful in other contexts.

To allow the widest possible use in AI development, digitized cultural heritage collections must be openly licensed. In section 5 a use case and example are discussed to illustrate how copyright and legal limitations, but especially ambiguities and uncertainties when dealing with rights and the (re-) distribution of digitized cultural heritage still present considerable barriers and obstacles for the increased uptake of cultural heritage data in AI.

3. Digitized Cultural Heritage for AI: Collections as Data

Most AI models have in common that they are trained using contemporary data sources and thus, for their adaptation and optimization for the cultural heritage domain, where predominantly historical data is being digitized (due to copyright), considerable amounts of suitable data are needed in order to train or fine-tune a given model for the domain. Ground truth data is typically created by manual transcription or annotation, but requires time, care and effort to produce. Crowd-sourcing can sometimes help create substantial amounts of ground truth data [10]. Such user input can then also be used to e.g. adaptively train an OCR algorithm [11]. Besides ground truth, also the noisy OCR of digitized historical collections can be very useful, e.g. for training large language models [12]. Open data from digitized cultural heritage can apparently make a valuable contribution here. What are the reasons why digitized cultural heritage data has so far only been used in isolated cases in the research and development of AI?

Libraries usually publish their digitized collections in online portals for discovery, with the option to search and browse either by metadata or (when available) full-text by keyword search. Thanks to new developments like the International Image Interoperability Framework (IIIF)[1],

[1] https://iiif.io/

standardized ways to distribute and harvest digitized cultural heritage via API are becoming more commonly available.

But APIs alone are not sufficient. Simple download dumps are often more useful to quickly explore what is being offered, without the need to learn or utilize an API. A good example from the cultural heritage sector are the downloadable ”packs” by the National Library of Luxembourg[2]. Packs of different sizes (from a small sample pack of a few GB to very large packs with hundreds of GB), various manifestations of the data (metadata, text only, structured markup) are provided.

Furthermore, while libraries typically publish data in formats that are de-facto standards in the digital library world, such as e.g. XML-based family of formats METS[3]/MODS[4]/ALTO[5], this considerably raises the barrier for reuse of this data in other domains. In contrast, most developers in AI will be more happy to work with formats like CSV or JSON, where numerous software libraries are available to process these further with. When libraries are unable to distribute their data in multiple ways, there should be clear documentation on the formats and how they are used and ideally information or links to technical resources that allow transformation into other formats in a straightforward and reproducible way.

How digitized cultural heritage data can be made more suitable for computational research and use in AI was a central part of the investigations in the *Collections as Data* project [13]. The project produced 10 recommendations in the ”Santa Barbara Statement”[6] for guiding cultural heritage institutions in lowering barriers and encouraging wider computational use of their data, but also asks for commitments to improve ethics and transparency with clear responsibilities for data stewardship. In ”50 Things”, a list with simple and practical measures to create easily consumable and machine-readable data from digitized cultural heritage is provided, alongside with concrete examples to improve documentation and contextualization of cultural heritage data in such a way that they are more attractive and adequate for computational use. The perspectives of collections as data have found wide appreciation in the sector and been adopted by various libraries in the US and also Europe [14].

At the same time, comprehensive reports on AI and machine learning in libraries have been produced in the US [15] and Europe [16], converging on the view that both, great possibilities, but also significant challenges lie in the increased use of AI in cultural heritage, and that there is a need to identify best practices that include responsible curation to fully leverage the possibilities. Cultural heritage organizations more active in AI have started to loosely organize in the international AI4LAM[7] community. A growing number of publications is now focusing on the specific context of AI and its applications in cultural heritage, such as a recent introduction for cultural heritage practitioners for supporting, participating in and undertaking machine learning-based activities [17], an overview and review of AI training resources from cultural heritage and recommendations for future work in this area [18], or a practical checklist for practitioners in libraries that are embarking on machine learning [19].

[2] https://data.bnl.lu/data/historical-newspapers/
[3] https://www.loc.gov/standards/mets/
[4] https://www.loc.gov/standards/mods/
[5] https://www.loc.gov/standards/alto/
[6] https://doi.org/10.5281/zenodo.3066208
[7] https://ai4lam.org

In summary, there is a clear momentum for AI in libraries, but it is also still in its infancy. To unlock the possibilities, libraries can not just rely on the fast progress in AI research, but in order to fully benefit from it, need to invest into more suitable ways to share their data, and into digital curation with a considerably broader scope of use, and responsibilities with regard to managing ethical issues and biases in data.

4. Issues and Biases in AI: models, datasets and collections

Issues and biases in AI models have gained more attention in recent years, with especially artists providing some compelling and revealing illustrations of problems relating to the quality and the lack of proper curation of the source data used to train AI models.

For example, in December 2020 the *VGG-Face2* dataset [20], which is widely used in facial recognition and its applications, was un-published after it became known that images from Flickr were used without observing the legal requirements or the consent of the persons depicted. Artist Adam Harvey has critically examined ethically problematic facial recognition datasets in his work since 2019 and has created a website that allows every Flickr user to check if their images were used in several widely used facial recognition datasets.[8]

Another artist work that critically engages with AI was presented by Kate Crawford and Trevor Paglen [21] who used *ImageNet* [22], a dataset commonly used for image classification, as the basis for "ImageNet Roulette", a website where visitors could have their own images classified on the basis of *ImageNet* - and then often found themselves confronted with statements by the AI with strongly negative connotations.

From the side of industry research on AI ethics, recommendations for ethically sound and transparent standards for publishing datasets [23] and AI models [24] have been proposed, alongside the critique of e.g. large language models [25]. This development also includes calls for more "accountable" curation, as is the practice in cultural institutions [26]. From critical surveys of recent issues pertaining to data in machine learning [27], to critical dataset studies [28], investigations into the power dynamics of image data annotation [29], and shifting the arguments for and against data curation to the question of how much we want to invest into it [30], more perspectives and challenges with data and AI are being exposed.

Within the domains of cultural heritage and digital humanities, this has also led to an increased awareness and several studies such as to determine what ethical issues arise in cultural heritage digitization and how they affect the ways decisions are taken and processes are organized [31], towards the identification of contentious terms and concepts in digitized newspapers [32], on the revelation of politics and biases in the digitization of newspaper collections [33], to examine core concepts in machine learning, generalization and unstructured data, in comparison to library practices for managing bias [34], or describing ethics scenarios of AI which have been developed specifically for information professionals [35]. While the overall landscape for AI in libraries still remains complex and partially unclear [36], finding adequate ways to deal with questions related to ethical AI clearly has to become an integral part in this endeavour.

With larger AI models, access to more computation often also equals better performance. Questions arise such as who has this power and who doesn't. Data Feminism [37] offers

[8] https://exposing.ai

valuable insights about data science and data ethics, and The Trouble With Big Data [38] opens up perspectives on data through the lens of culture rather than social, political or economic trends. Including these critical views on data and AI from the domain of culture and humanities into the development and use of AI is essential to its further benefit, and cultural heritage institutions can make an important contribution.

In order to respond to some of these challenges, for the AI project *Mensch.Maschine.Kultur*[9] at SBB, a full position has been allocated to investigating and documenting the requirements for publication of digitized cultural heritage data for use in AI and research. Guidelines for the responsible curation of digitized cultural heritage data with a particular focus on the identification and treatment of ethical, legal and social aspects will be created. For part of the work on AI for subject classification, a complementary ethical audit will be performed by external experts using Ethical Foresight Analysis [39].

5. Reusing Cultural Heritage Data for AI: a legal gray area?

Given that cultural heritage is made available as data suitable for immediate use in AI, and supported with digital curation that cares about transparency and ethical issues, there nevertheless sometimes remain legal obstacles for the uptake of cultural heritage as data in AI.

Despite comprehensive legal frameworks for digitization of cultural heritage, there are still several legal ambiguities when it comes to the reuse and redistribution of digitized cultural heritage in AI contexts. For example, the release of a dataset is legally considered a publication and thereby can also be considered a from of re-distribution of the source used to produce it. But cultural heritage institutions often shy away from the complexity of assessing if and what legal implication such re-distribution can mean for them. There are e.g. concerns about the (commercial) duplication of data offerings and services, or institutions are restricted by contractual obligations to publishers, service providers or digitization project partners.

But there are also open questions on simple practical issues, such as to what extent the publication of only parts of the source data or of derivatives created from it are affected by licensing conditions. It is unclear for example, how to proceed after the preparation of a new dataset in the form of annotations or transcriptions for segments of the source data with a view on the subsequent distribution of that newly created dataset. In this respect, there is still a lack of clear guidance and recommendations to assist the creators of such datasets in determining suitable licenses and their reuse options. Even considerably open licenses such as the Creative Commons Attribution NonCommercial ShareAlike License, which is widely used by cultural institutions, can restrict reuse and distribution of digitized cultural heritage. For example, if advertising is also placed on the same website where a dataset is made available, this can legally be considered a form of commercial redistribution that would be prohibited by that license.

Accordingly, the legal requirements for reuse of cultural heritage in AI are still largely a gray area and often have limited applicability to the specific case. Consider, for example, the case of the creation of a ground truth dataset for OCR based on digitized cultural heritage from libraries which reuses data produced in the context of Google Books public-private-partnerships [40]. In principle, written permission must be obtained from Google for any use of such digitized

[9] https://blog.sbb.berlin/mensch-maschine-kultur-neues-projekt-zur-kuenstlichen-intelligenz/

material, including non-commercial use, and only the distribution of the scans on the websites of the library partner is directly approved. Thus, whoever wants to reuse these data to create new datasets and accordingly distribute them, must either restrict their own dataset to only include data from institutions that explicitly permit such re-distribution, or which grant bilateral exemptions, or find creative alternatives to indirect distribution - such as e.g. by listing only the URL of the digitized material at the providing institution instead of attaching the actual data. This however creates unnecessary obstacles for further engagement with these resources.

Similarly, the authors of the dataset *GT4HistOCR* [41], that has been derived by transcriptions of lines from historical books taken from the Internet Archive, have chosen to partially randomize the order of the text lines in their dataset in order to prevent complete works from being reconstructed from the individual lines. Such "anticipatory" practice in turn creates hurdles for reuse and makes it more difficult to transparently track and evaluate the provenance of a dataset, which is especially relevant in scientific contexts, e.g. for replicability.

Even established and commonly used licenses for cultural data can only be applied to a limited extent to the use and redistribution in the context of AI, or cover this only incompletely. Cultural heritage institutions and those who want to reuse cultural heritage data need better legal guidelines and support for common reuse scenarios.

6. Summary and Outlook

In summary, it can be said that while AI offers great potential for applications in the cultural sector, at the same time there are still many unsolved challenges relating to the ways in which data is distributed, and also uncertainties and legal gray areas preventing wider distribution and reuse of cultural heritage as data for AI.

On the one hand, cultural heritage institutions need to create fundamentally better and more suitable ways for the publication and redistribution of cultural heritage as data. And they also need to invest into responsible digital curation and data stewardship to provide data that is not only useful for AI, but also as aware of ethical issues and biases as possible.

On the other hand, cultural heritage institutions with their extensive and diverse digital collections and their quality awareness and established standards and processes for curation, can create good examples of open and well-documented datasets, supported by commitments to digital curation that cares about quality, transparency and awareness for biases, from which ultimately the development and use of AI in science and industry as well as society as a whole can benefit.

7. Acknowledgments

This work was partially supported by the Federal German Ministry of Education and Research (BMBF), project grant QURATOR, grant no. 03WKDA1A.

References

[1] E. Engl, M. Boenig, K. Baierer, C. Neudecker, V. Hartmann, Full-text for the early modern age. the contribution of the ocr-d-project to the full-text recognition of early modern prints, Zeitschrift für historische Forschung 47 (2020) 223–250.

[2] P. Kahle, S. Colutto, G. Hackl, G. Mühlberger, Transkribus-a service platform for transcription, recognition and retrieval of historical documents, in: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), volume 4, IEEE, 2017, pp. 19–24.

[3] M. Brantl, K. Ceynowa, T. Meiers, T. Wolf, Visuelle suche in historischen werken, Datenbank-Spektrum 17 (2017) 53–60.

[4] D. Abhishek, B. Giles, Visual analysis of chapbooks printed in scotland, in: The 6th International Workshop on Historical Document Imaging and Processing, 2021, pp. 67–72.

[5] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, Extended overview of clef hipe 2020: named entity processing on historical newspapers, in: CEUR Workshop Proceedings, 2696, CEUR-WS, 2020, p. 38.

[6] O. Suominen, Annif: Diy automated subject indexing using multiple algorithms, 2019.

[7] A. Kasprzik, Putting research-based machine learning solutions for subject indexing into practice, in: CEUR Workshop Proceedings, 2535, CEUR-WS, 2020, p. 8.

[8] G. Rehm, P. Bourgonje, S. Hegele, F. Kintzel, J. M. Schneider, M. Ostendorff, K. Zaczynska, A. Berger, S. Grill, S. Räuchle, et al., Qurator: innovative technologies for content and data curation, arXiv preprint arXiv:2004.12195 (2020).

[9] B. McGillivray, B. Alex, S. Ames, G. Armstrong, D. Beavan, A. Ciula, G. Colavizza, J. Cummings, D. De Roure, A. Farquhar, et al., The challenges and prospects of the intersection of humanities and data science: A white paper from the alan turing institute, 2020.

[10] A. Dumitrache, O. Inel, B. Timmermans, C. Ortiz, R.-J. Sips, L. Aroyo, C. Welty, Empirical methodology for crowdsourcing ground truth, Semantic Web 12 (2021) 403–421.

[11] C. Neudecker, A. Tzadok, User collaboration for improving access to historical texts, Liber Quarterly 20 (2010).

[12] S. Schweter, L. März, K. Schmid, E. Çano, hmbert: Historical multilingual language models for named entity recognition, arXiv preprint arXiv:2205.15575 (2022).

[13] T. Padilla, Responsible Operations: Data Science, Machine Learning, and AI in Libraries. OCLC Research Position Paper., ERIC, 2019.

[14] G. Candela, M. D. Sáez, M. Escobar Esteban, M. Marco-Such, Reusing digital collections from glam institutions, Journal of Information Science 48 (2022) 251–267.

[15] R. Cordell, Machine learning and libraries: a report on the state of the field, 2020.

[16] G. Markus, C. Neudecker, A. Isaac, G. Bergel, W. Bailer, M. Marrero, V. Tzouvaras, J. Oomen, P. van Kemenade, M. Bontje, M. Cuper, S. Bartholmei, J. E. Cejudo, A. Larsson, G. Angelaki, Ai in relation to glams. europeanatech task force report and recommendations, 2021.

[17] D. van Strien, M. Bell, N. R. McGregor, M. Trizna, An introduction to ai for glam, in: Proceedings of the Second Teaching Machine Learning and Artificial Intelligence Workshop, PMLR, 2022, pp. 20–24.

[18] A. Darby, C. N. Coleman, C. Engel, D. van Strien, M. Trizna, Z. W. Painter, Ai training resources for glam: a snapshot, arXiv preprint arXiv:2205.04738 (2022).

[19] B. C. G. Lee, The" collections as ml data" checklist for machine learning & cultural heritage, arXiv preprint arXiv:2207.02960 (2022).
[20] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, in: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), IEEE, 2018, pp. 67–74.
[21] K. Crawford, T. Paglen, Excavating ai: The politics of images in machine learning training sets, AI and Society (2019).
[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.
[23] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, K. Crawford, Datasheets for datasets, Communications of the ACM 64 (2021) 86–92.
[24] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model cards for model reporting, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 220–229.
[25] E. M. Bender, T. Gebru, A. McMillan-Major, M. Mitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623.
[26] E. S. Jo, T. Gebru, Lessons from archives: Strategies for collecting sociocultural data in machine learning, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 306–316.
[27] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, A. Hanna, Data and its (dis) contents: A survey of dataset development and use in machine learning research, Patterns 2 (2021) 100336.
[28] N. B. Thylstrup, The ethics and politics of data sets in the age of machine learning: deleting traces and encountering remains, Media, Culture & Society (2022) 01634437211060226.
[29] M. Miceli, M. Schuessler, T. Yang, Between subjectivity and imposition: Power dynamics in data annotation for computer vision, Proceedings of the ACM on Human-Computer Interaction 4 (2020) 1–25.
[30] A. Rogers, Changing the world by changing the data, arXiv preprint arXiv:2105.13947 (2021).
[31] Z. Manžuch, Ethical issues in digitization of cultural heritage, Journal of Contemporary Archival Studies 4 (2017) 4.
[32] R. Brate, A. Nesterov, V. Vogelmann, J. Van Ossenbruggen, L. Hollink, M. Van Erp, Capturing contentiousness: Constructing the contentious terms in context corpus, in: Proceedings of the 11th on Knowledge Capture Conference, 2021, pp. 17–24.
[33] K. Beelen, J. Lawrence, D. Wilson, D. Beavan, Bias and representativeness in digitized newspaper collections: Introducing the environmental scan, Digital Scholarship in the Humanities (2022).
[34] C. N. Coleman, Managing bias when library collections become data, International Journal of Librarianship 5 (2020) 8–19.
[35] A. Cox, The ethics of ai for information professionals: Eight scenarios, Journal of the Australian Library and Information Association (2022) 1–14.
[36] A. Gasparini, H. Kautonen, Understanding artificial intelligence in research libraries–

extensive literature review, LIBER Quarterly: The Journal of the Association of European Research Libraries 32 (2022).
[37] C. D'ignazio, L. F. Klein, Data feminism, MIT press, 2020.
[38] J. Edmond, N. Horsley, J. Lehmann, M. Priddy, The Trouble With Big Data: How Datafication Displaces Cultural Practices, Bloomsbury Academic, 2021.
[39] H. Bubinger, J. D. Dinneen, Actionable approaches to promote ethical ai in libraries, Proceedings of the Association for Information Science and Technology 58 (2021) 682–684.
[40] D. Lassner, C. Neudecker, J. Coburger, A. Baillot, Publishing an ocr ground truth data set for reuse in an unclear copyright setting., Zeitschrift für digitale Geisteswissenschaften (2021).
[41] U. Springmann, C. Reul, S. Dipper, J. Baiter, Ground truth for training ocr engines on historical documents in german fraktur and early modern latin, arXiv preprint arXiv:1809.05501 (2018).