Volltexte für die Forschung: OCR partizipativ, iterativ und on Demand

Anke Hertling, Leibniz-Institut für Bildungsmedien | Georg-Eckert-Institut
Sebastian Klaes, Leibniz-Institut für Bildungsmedien | Georg-Eckert-Institut

Zusammenfassung
Für die Forschungsarbeit mit digitalisierten Quellen stellt die Leistung der Volltexterkennung, also die Genauigkeit der Optical Character Recognition (OCR), eine wesentliche Grundlage dar. Die Volltexterkennung avanciert damit zu einem Qualitätskriterium von digitalen Sammlungen und Bibliotheken müssen als zentrale Digitalisierungsakteure ihrer Verantwortung im Hinblick auf die Evidenz von auf Volltexten basierenden wissenschaftlichen Ergebnissen gerecht werden. Ausgehend von einer Digitalisierung, die explizit an der Zielgruppe der digitalen Forschung ausgerichtet ist, greift der folgende Beitrag Formate und Workflows zur Organisation der Volltexterkennung als partizipativen und iterativen Prozess in Zusammenarbeit mit der Forschung auf. Vor dem Hintergrund der aktuellen OCR-D-Förderphase wird ein on-Demand-Ansatz, bei dem OCR-Prozesse nach spezifischen Bedarfen durchgeführt werden, vorgestellt.

Abstract
For working with digitized sources in research, the quality of full-text recognition, i.e. the accuracy of Optical Character Recognition (OCR), is essential. Full-text recognition is thus advancing to become a quality criterion of digital collections, and libraries – as central actors in digitization – must live up to their responsibility regarding the evidence of scientific results based on full text. Starting from a digitization process that is explicitly oriented towards digital research, the paper discusses formats and workflows for organizing full-text recognition as an iterative and participatory process in collaboration with researchers. Against the background of the current OCR-D funding phase, the paper also presents an on-demand approach for OCR processes according to specific requirements.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/5832

Schlagwörter: Digitalisierung; Volltext; OCR; Optische Zeichenerkennung

Autorenidentifikation:
Hertling, Anke: GND: 1033153737; ORCID: https://orcid.org/0000-0002-3163-2233;
Klaes, Sebastian: ORCID: https://orcid.org/0000-0003-3597-7017

1. Einleitung

Mit der wachsenden Bedeutung einer Forschung, die wie die Digital Humanities (DH) mit digitalen Methoden und Werkzeugen arbeitet, verändern sich die Anforderungen an von Bibliotheken bereitgestellte digitale Sammlungen und somit auch an die Digitalisierung von Quellen. Die

Auseinandersetzung mit der Volltexterkennung stellt dabei neben der Bereitstellung verschiedener Daten- und Exportformate[1] einen, so formuliert es die vom Verband „Digital Humanities im deutschsprachigen Raum“ (DHd) 2019 gegründete Arbeitsgruppe DHd-AG OCR, „Schlüssel für die Umsetzung der Forderungen der DH“[2] dar. Insbesondere bei historischen Quellen, bei denen Schriftarten, Layout, Sprache und Orthographie vielfach variieren, ist die Volltexterkennung mit ihren Prozessen der Vorverarbeitung in Form von Bildoptimierung und Binarisierung (Preprocessing), der Layoutsegmentierung (Region Segmentation), der Zeichenerkennung (Character Recognition) und der Nachbearbeitung im Sinne einer Fehlerkorrektur nach wie vor eine große Herausforderung. 2021 startete die Deutsche Forschungsgemeinschaft (DFG) deshalb in die dritte Phase ihrer OCR-D Förderinitiative und unterstützt die OCR-Weiterentwicklung, um die Volltexttransformation der im deutschen Sprachraum erschienenen Drucke des 16. bis 18. Jahrhunderts zu ermöglichen und so zu optimieren, dass die Quellen nicht nur durchsuch-, sondern verstärkt digital analysierbar sind.[3] Ziel der aktuellen DFG-Förderphase ist es, den seit 2015 entwickelten OCR-D-Softwareprototyp in Workflow- und Digitalisierungssysteme zu integrieren und die Erzeugung qualitativ hochwertiger Volltexte in den bibliothekarischen Regelbetrieb zu überführen.[4] Koordiniert wird die Förderphase von der Herzog August Bibliothek in Wolfenbüttel, die mit ihren Partnerinstitutionen einen regelmäßigen Austausch zwischen den Projektbeteiligten, darunter die Universitätsbibliotheken Braunschweig und Mannheim, die SLUB Dresden, die SUB Göttingen und die Gesellschaft für wissenschaftliche Datenverarbeitung in Göttingen (GWDG), die ULB Sachsen-Anhalt, die Johannes Gutenberg-Universität Mainz und die FAU Erlangen-Nürnberg, organisiert. Die Forschungsbibliothek des Leibniz-Instituts für Bildungsmedien | Georg-Eckert-Institut (GEI) arbeitet gemeinsam mit dem Institut für Mensch-Computer-Medien (HCI) und dem Zentrum für Philologie und Digitalität „Kallimachos“ (ZPD) der Universität Würzburg ebenfalls im Projektverbund und bringt dabei besonders ihre enge Zusammenarbeit mit der Digital-Humanities-Forschung ein. Bereits seit Beginn der Digitalisierung erfolgte der Auf- und Ausbau der digitalen Sammlung am GEI forschungsbasiert. Die langjährige gemeinsame Arbeit mit der Fachcommunity wird im Folgenden dargelegt und entsprechende Instrumente sowie Erfahrungen aus der engen Kooperation mit der Forschung aufgezeigt. Besonders vor dem Hintergrund der Anforderungen aus den Digital Humanities wird ein Verständnis von Volltexterkennung als partizipativer und iterativer Prozess vorgestellt, der neue Workflows und Organisationsstrukturen bei allen Akteuren erfordert. Damit optimierte und nachhaltige OCR-Ergebnisse für die Forschung generiert werden können, wird ein on-Demand-Ansatz entwickelt, bei dem die Volltexterkennung entsprechend spezifischer Forschungs- und Materialbedarfe auf Korpusebene durchgeführt wird.

1 U.a. Gasser, Sonja: Das Digitalisat als Objekt der Begierde. Anforderungen an digitale Sammlungen für Forschung in der Digitalen Kunstgeschichte, in: Andraschke, Udo; Wagner, Sarah (Hg.): Objekte im Netz. Wissenschaftliche Sammlungen im digitalen Wandel, Bielefeld 2020, S. 261–276, hier S. 267 ff. Online: <https://doi.org/10.14361/9783839455715>.

2 Vgl. die auf ihrer Webseite formulierten Arbeitsschwerpunkte der DHd-AG OCR: <https://dig-hum.de/ag-ocr>, Stand: 26.04.2022.

3 Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition. Das OCR-D-Projekt. Online: <https://ocr-d.de/de/about>, Stand: 26.04.2022; sowie Engl, Elisabeth: OCR-D kompakt. Ergebnisse und Stand der Forschung in der Förderinitiative, in: Bibliothek – Forschung und Praxis 44 (2), 2020, S. 218–230. Online: <https://doi.org/10.1515/bfp-2020-0024>. Stand: 26.04.2022>.

4 DFG: Implementierung der OCR-D-Software zur Volltextdigitalisierung. Information für die Wissenschaft Nr. 15 | 27. Februar 2020. Online: <https://www.dfg.de/foerderung/info_wissenschaft/2020/info_wissenschaft_20_15/index.html>, Stand: 26.04.2022.

2. Sondierung der Interessen beim Aufbau der digitalen Schulbuchbibliothek GEI-Digital

Auch in ihrem aktuellen Förderprogramm „Digitalisierung und Erschließung" erklärt die DFG die Wissenschaft zur primären Zielgruppe und die „nachdrückliche Stimulierung und Stärkung wissenschaftlicher Forschung"[5] als eine Voraussetzung für die Förderung von Digitalisierungsprojekten. Eine wissenschaftliche Beteiligung bei der Digitalisierung kann zum Beispiel in Form eines Digitalisierungsbeirates gewährleistet werden. So haben am GEI Vertreter*innen u.a. aus der Geschichts- und Erziehungswissenschaft gemeinsam mit der Forschungsbibliothek einen Digitalisierungsplan mit der Zielsetzung entwickelt, die gesamte deutsche historische Schulbuchsammlung des Instituts, die die Fächer Geschichte, Geographie, Politik, Realien, Werteerziehung/Religion sowie den (Erst-) Leseunterricht umfasst, zu digitalisieren und zugänglich zu machen. Im Fokus der Zusammenarbeit mit dem Digitalisierungsbeirat zum Aufbau der digitalen Schulbuchbibliothek GEI-Digital standen zunächst fachwissenschaftliche Kriterien zur Erstellung von Korpora, die als Grundlage für einen vom GEI 2007 in der Förderlinie „Digitalisierung der DFG-Sondersammelgebiete" eingereichten DFG-Antrag fungierten. Gemeinsame Abstimmungsprozesse waren auch deshalb notwendig, weil Schulbücher als Gebrauchsmedien kaum und selbst nach Gründung der Deutschen Bücherei zumeist nicht systematisch gesammelt wurden und bis heute keine Bibliografie vorliegt, die alle deutschsprachigen Schulbücher verzeichnet. Spätestens ab 1871 gab es zahlreiche Auflagen sowie eine Vielzahl an regional ausdifferenzierten Ausgaben, deren vollständige Digitalisierung Finanz- und Zeitrahmen, wie sie in DFG-Projekten möglich wären, erheblich überschritten hätten. So wurde u.a. mit dem Digitalisierungsbeirat festgelegt, dass mehrere Auflagen nur dann digitalisiert werden, wenn zwischen der frühestmöglichen und der spätesten Auflage signifikante Veränderungen zum Beispiel in Form von Umfangssteigerungen festzustellen sind.[6]

Obwohl die Anfänge der Schulbuchproduktion im 17. Jahrhundert liegen, waren die Entwicklungen des Schulbuchs im 19. Jahrhundert für die Entscheidung des Digitalisierungsbeirats maßgeblich, Bestände aus dieser Zeit prioritär zu digitalisieren. Erst im 19. Jahrhundert setzten sich Schulbücher, wie sie heute gebräuchlich sind, durch. Schulbücher zählen dabei zu den ersten modernen Massenmedien. Mit der Einführung der allgemeinen Schulpflicht Anfang des 19. Jahrhunderts avancierten sie zu einem staatlichen Steuerungsinstrument und wurden zum festen Bestandteil des Unterrichts. Große Verlage wie Ferdinand Hirt in Breslau oder Velhagen & Klasing in Bielefeld richteten ihr Kerngeschäft auf die Konzeption und Produktion von Schulbüchern verschiedener Schulfächer aus.[7] Insbesondere vor dem Hintergrund der Ausdifferenzierung im höheren Schulwesen stieg die Zahl der im Deutschen Kaiserreich publizierten Schulbücher in Bezug auf Auflage und Exemplare stark an und erreichte die

5 DFG: Merkblatt und ergänzender Leitfaden – Digitalisierung und Erschließung, DFG Vordruck 12.15 – 09/21, S. 5. <https://www.dfg.de/foerderung/programme/infrastruktur/lis/lis_foerderangebote/digitalisierung_erschliessung/formulare_merkblaetter/index.jsp>, Stand: 26.04.2022.

6 Hertling, Anke; Klaes, Sebastian: Historische Schulbücher als digitales Korpus für die Forschung. Auswahl und Aufbau einer digitalen Schulbuchbibliothek, in: Nieländer, Maret; De Luca, Ernesto William (Hg.): Digital Humanities in der internationalen Schulbuchforschung. (Eckert. Expertise 9), Göttingen 2018, S. 22–44. Online: <https://repository.gei.de/handle/11428/296> (DOI 10.14220/9783737009539), Stand: 26.04.2022.

7 Jäger, Georg: Der Schulbuchverlag, in Ders. et al. (Hg.): Geschichte des deutschen Buchhandels im 19. und 20. Jahrhundert. Bd. 1: Das Kaiserreich 1870-1918, Teil 2, Frankfurt am Main 2003, S. 62–102.

Schulbuchproduktion zwischen 1871 und 1918 einen quantitativen Höhepunkt. So wurden rund 40% aller Geschichtsschulbücher, die im Zeitraum zwischen 1700 und 1945 in den zum deutschen Staatsgebiet zählenden Territorien erschienen sind, in dieser Zeit veröffentlicht.[8]

Angesichts der forcierten Schulbuchproduktion im Deutschen Kaiserreich sowie der Übernahme des Bildungsmonopols durch den Staat plädierte der am GEI eingesetzte Digitalisierungsbeirat, dass Schulbücher aus dieser Epoche von besonderer wissenschaftlicher Relevanz sind und diese zuerst digitalisiert und möglichst als Volltexte zugänglich gemacht werden sollten. Problematisiert wurde dabei, dass den älteren und zumeist unikalen Quellen aus Sicht der Bestandserhaltung eine höhere Digitalisierungspriorität hätte eingeräumt werden müssen. So wäre die sehr wertvolle historische Fibel-Sammlung aus Sicht der GEI Forschungsbibliothek bevorzugt zu digitalisieren. Die Community der Fibel-Forschung ist jedoch sehr überschaubar und für die Fibeln mit ihren vielen Abbildungen und unterschiedlichen Schriftarten war zu dieser Zeit eine Volltexterkennung nicht zielführend. Notwendig war demnach ein genaues Abwägen und ein gemeinsamer Verständigungsprozess, bei dem das GEI vor allem davon profitierte, dass die Volltexttransformation der digitalen Quellen gleich zu Beginn seiner Digitalisierungsplanungen mit einkalkuliert worden war. Der Digitalisierungsbeirat konnte gleichfalls für die Digitalisierung als Maßnahme der Bestandserhaltung und somit als Grundlage für zukünftige Forschungen sensibilisiert werden. Entsprechende finanzielle und personelle Mittel wurden im Institut eingeplant, um besonders gefährdete Quellen rechtzeitig zu digitalisieren, auch wenn kein unmittelbarer Forschungsbedarf vorliegt. Inzwischen stehen in der digitalen Schulbuchbibliothek GEI-Digital (https://gei-digital.gei.de/viewer/index/) 5.000 historische Schulbücher im Umfang von rund 1,6 Millionen Seiten überwiegend im Volltext frei zugänglich zur Verfügung.[9] Die Digitalisierung von Geographie-, Realien- und Geschichtsschulbüchern aus der Zeit des 17. Jahrhunderts bis 1918 sowie die Digitalisierung von Lesebüchern aus der Zeit des Deutschen Kaiserreichs sind nunmehr abgeschlossen. Neben einer Metadaten- und Volltextsuche im Gesamtbestand bietet GEI-Digital eine differenzierte Recherche in den einzelnen Korpora und verschiedene Ausgabeformate für die Volltexte (u.a. ALTO) und Metadaten (u.a. Dublin Core, METS/MODS, MARC XML).

3. Texterkennung partizipativ und iterativ

In der Zusammenarbeit mit dem Digitalisierungsbeirat wurde zu einem für das Bibliothekswesen recht frühen Zeitpunkt deutlich, dass Volltexte eine unverzichtbare Grundlage und Nukleus für die wissenschaftliche Analyse und Weiterverarbeitung von digitalen Quellen sind. Aus Sicht der Forschungsbibliothek des GEI war es nach zehn Jahren Digitalisierung eine programmatische Notwendigkeit, die OCR-Qualität ihrer Digitalisate zu evaluieren. Eine schon 2014 am GEI durchgeführte Online-Befragung machte die Bedeutung qualitativ hochwertiger Volltexte deutlich. Gegenstand der Umfrage waren die mit Abbyy FineReader Engine 10 und Engine 11 behandelten und auf GEI-Digital bereitgestellten Volltexte. An der Befragung haben 106 GEI-Digital-Nutzer*innen teilgenommen, wobei 70% der Befragten die Qualität der Volltexte generell als „sehr wichtig“ beurteilten.

8 Jacobmeyer, Wolfgang: Das deutsche Schulgeschichtsbuch 1700-1945. Die erste Epoche seiner Gattungsgeschichte im Spiegel der Vorworte, Bd. 1, Berlin 2011, S. 19.

9 Rund 113.000 Seiten wurden nicht mit OCR behandelt, darunter Bestände mit überwiegend graphischen Darstellungen wie zum Beispiel Atlanten.

29% der Befragten waren mit der auf GEI-Digital zugänglichen OCR-Qualität *sehr zufrieden*, jeweils weitere 25% waren *weitestgehend zufrieden* bzw. *zufrieden*. Die Antworten legen nahe, dass ca. ¾ der Nutzer*innen von GEI-Digital mit der Texterkennung zufrieden waren und diese als Mehrwert erkannten. Im Rahmen der Befragung wurden darüber hinaus Perspektiven zur Nachnutzung der Texterkennung erfragt. *Exportmöglichkeiten von Volltexten* wünschten sich 76% der Befragten, wobei 54% der Umfrageteilnehmer*innen *ein direktes Exportieren in Analysewerkzeuge wie beispielsweise TextGrid* als *sehr wichtig* anerkannten. Auf die Frage, ob GEI-Digital-Nutzer*innen fehlerhafte Volltexte selbst korrigieren würden, gaben 61% der Befragten eine positive Rückmeldung. Aufschlussreich waren auch Angaben zur Relevanz bestimmter Funktionen und Suchmöglichkeiten. Eine Funktion oder Möglichkeit zur Analyse von Volltexten-Fehlerraten wurde von 42% als *sehr wichtig* und von weiteren 42% als *wichtig* angesehen. Bei den Suchmöglichkeiten wurde der fehlertoleranten Suche nach Volltexten, bei der nicht nur eine exakte, sondern zudem eine ähnliche Zeichenfolge als Suchoption fungiert, ein hoher Stellenwert eingeräumt. Die Bedeutung der Volltexterkennung für die Forschung zeigte auch die 2016 von der Bayerischen Staatsbibliothek (BSB) im Rahmen des OCR-D-Kooperationsprojekts durchgeführte Befragung. Knapp über die Hälfte der 139 Teilnehmer*innen würde fehlerhafte OCR-Texte zu Forschungszwecken verwenden, denn auch fehlerhafte Volltexte seien durchaus hilfreich, u.a. für eine Volltextsuche. 40% der Befragten hielten fehlerhafte OCR-Texte indessen für nutzlose Daten.[10]

Bekräftigt wurde der Befund, dass die Volltexte von GEI-Digital als verbesserungswürdig anzusehen sind, durch die am Institut zunehmend durchgeführten Digital-Humanities-Projekte, wie zuletzt das Projekt „DiaCollo für GEI-Digital“[11], in dem sprachtechnologische Verfahren auf den GEI-Digital Volltextkorpus eingesetzt wurden. Mit dem im Rahmen der CLARIN-D-Initiative entwickelten Werkzeug DiaCollo[12] wurden typische Wortverbindungen auf Grundlage verschiedener digitaler Schulbuchkorpora ermittelt und die Ergebnisse visuell aufbereitet. Im Zuge der Anpassung von DiaCollo für seine Anwendung in der historischen Schulbuchforschung wurden Unschärfen bei den GEI-Digital-Volltexten nochmals sehr sichtbar. Fehlerhafte OCR beeinträchtigen die wissenschaftliche Analyse in erheblicher Weise und verfälschen Ergebnisse u.a. bei der Kollokationsanalyse. So wird zum Beispiel in dem *Berlinischen neu eingerichteten Schulbuch* aus dem Jahr 1761 das Wort „und“ häufig als „uiw“ erkannt, das im Korpus selten, dann aber wiederum sehr oft mit dem Suchbegriff „Schule“ vorkommt und deshalb als relevantes Kollokat interpretiert wird. Die Dominanz von Fraktur sowie die für Schulbücher typischen komplexen Layoutstrukturen einschließlich Anschauungstafeln, Inhaltsverzeichnissen, Fußnoten oder Abbildungen sind hier materialspezifische Ursachen für einen Qualitätsverlust bei der Volltexterkennung. Ein Schulbuch wird darüber hinaus vielfach von mehreren Generationen genutzt. In Schulbüchern vorkommende Randnotizen und Anstreichungen sind ebenfalls entscheidende OCR-Fehlerquellen.

10 Vgl. die Umfrage zur Verwendung von OCR-Texten aus dem Jahr 2016: <https://ocr-d.de/de/umfrage>, Stand: 26.04.2022.
11 Leibniz-Institut für Bildungsmedien | Georg-Eckert Institut: Diacollo für GEI-digital.<https://diacollo.gei.de/>, Stand: 26.04.2022.
12 CLARIN-D: DiaCollo. Kollokationsanalyse in diachroner Perspektive.<https://www.clarin-d.net/de/kollokationsanalyse-in-diachroner-perspektive>, (Stand: 26.04.2022.

Auch wenn die DFG-Standards bei der Volltexterkennung eingehalten wurden, wuchs die Unzufriedenheit der Forschung mit der OCR-Qualität von GEI-Digital. Selbst bei 98% Erkennungsrate ist bei einer Seite mit rund 2.000 Zeichen mit etwa 40 Fehlern zu rechnen und werden letztlich bis zu 40 Wörter nicht gefunden, so dass Datenaufbereitungsverfahren notwendig sind, um zum Beispiel CLARIN-D-Werkzeuge verlässlich anwenden zu können.[13] Angesichts der hohen Nachfrage nach qualitativ hochwertigen Volltexten hat die Forschungsbibliothek des GEI 2019 die Qualität von verschiedenen OCR-Softwaresystemen evaluiert. Bei der Qualitätsmessung wurde auf das von der DFG in ihren Praxisregeln empfohlene „Bernoulli-Experiment" zurückgegriffen.[14] In einer Stichprobe mit zehn Lesebüchern mit Fraktur-Schriftbild aus der Zeit des Deutschen Kaiserreichs mit einem Umfang von 4.000 Seiten wurde geprüft, ob Zeichen richtig oder falsch erkannt wurden, was erste Rückschlüsse auf die Genauigkeit der OCR-Erkennung und damit die Validität der Daten ermöglicht. Verglichen wurde die Erkennungsquote zwischen dem bislang für den GEI-Digital-Korpus genutzten kommerziellen Software-Anbieter Abbyy FineReader SDK und der Open-Source-Software Tesseract als ein Texterkennungssystem, das Verfahren der Künstlichen Intelligenz in Form sog. rekurrenter neuronaler Netze nutzt und auf trainierbaren Datenmodellen basiert. Tesseract erlaubt demnach das Training spezifischer Modelle u.a. auf Frakturschriften.[15] Bei seinen Tests auf Tesseract 4.0-Basis hat das GEI bei seinen zehn Lesebüchern zwei Modelle eingesetzt: (1) ein Standard-Modell von einem Dienstleister (Dienstleister-Tesseract) sowie (2) ein durch einen Dienstleister spezifisch auf GEI-Daten trainiertes Tesseract-Modell (GEI-Tesseract), wobei durch eine Zufallsfunktion 10.000 Zeichen aus dem Korpus der Lesebücher ausgewählt und durch deren Transkription das Modell trainiert wurde.

Auch wenn das Bernoulli-Experiment nur einen groben Einblick ermöglicht und die zehn Lesebücher bei einem Korpus von insgesamt über 1.351 digitalisierten Lesebüchern nur einen kleinen Ausschnitt darstellen, konnte festgehalten werden, dass die Tesseract-Modelle gegenüber dem kommerziellen Software-Anbieter große Potenziale aufwiesen, was die UB Heidelberg bei ihren 2019 durchgeführten Tests nochmals bekräftigen konnte.[16] Bei den zehn Lesebüchern wurde in wenigen Fällen der Wert von mehr als 496 korrekt erkannten Zeichen überschritten, was einer Tesseract-Erkennungsrate von rund 98% und damit den Vorgaben der DFG entspricht. Mit Blick auf digitale Methoden wie Topic Modeling oder Data Mining wären für die Belastbarkeit der Daten erneute Qualitätsprüfungen und ggf. OCR-Korrekturverfahren notwendig.

13 Nieländer, Maret; Weiß, Andreas: „Schönere Daten" – Nachnutzung und Aufbereitung für die Verwendung in Digital-Humanities-Projekten", in: Nieländer, Maret; De Luca, Ernesto William (Hg.): Digital Humanities, 2018, S. 91–116. <https://repository.gei.de/handle/11428/296> (DOI 10.14220/9783737009539), Stand: 26.04.2022.

14 DFG: Praxisregeln „Digitalisierung". DFG Vordruck 12.151 – 12/16, S. 35, <https://www.dfg.de/formulare/12_151/12_151_de.pdf> Stand: 26.04.2022.

15 Weil, Stefan: tesseract-ocr / tesstrain, <https://github.com/tesseract-ocr/tesstrain/wiki>, Stand: 26.04.2022.

16 Weil, Stefan: Neue Frakturmodelle für Tesseract. Präsentation auf dem Kitodo Anwendertreffen 18.-19. November 2019, S. 3, <https://madoc.bib.uni-mannheim.de/53748/1/2019-11-18.pdf>, Stand: 26.04.2022.

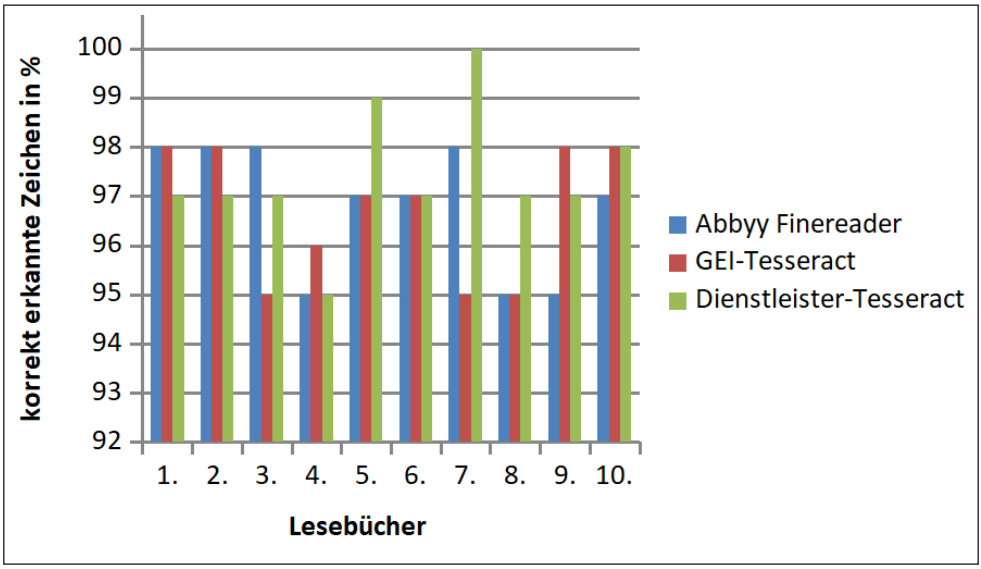

Abbildung: Vergleich OCR-Ergebnisse bei zehn Lesebüchern aus dem Deutschen Kaiserreich durch Bernoulli-Experiment

Neben Tesseract weisen auch andere Open-Source-Engines vielversprechende Ergebnisse auf, die sogar Resultate mit einer Erkennungsrate von über 99% erzielen.[17] Wegweisend ist dabei die Open-Source-Engine Calamari, die die Erkennungsgeschwindigkeit erheblich steigert, die Generierung von Trainingsmaterial erleichtert und auch bei den Erkennungsgenauigkeiten sehr gute Leistungen liefert.[18]

Die Erstellung von Volltexten und ihre Qualität sind in einem zunehmenden Maße abhängig vom Stand der jeweiligen Technologien und der Qualität der Trainingsdaten. Vor dem Hintergrund der fortschreitenden OCR-Weiterentwicklungen, wie sie auch im OCR-D Kooperationsprojekt forciert werden, ist Volltexterkennung demnach als iterativ zu verstehen, und das heißt, dass OCR-Prozesse je nach aktuellem Stand der Technologien mehrfach wiederholt durchgeführt werden sollten, zumal sich die Bedarfe aus der Wissenschaft ebenfalls fortwährend verändern. Mit ihren inzwischen vorliegenden großen Datenmengen kommen Bibliotheken allerdings bei umfangreichen OCR-Prüfungen und -Verbesserungen an die Grenzen ihrer Ressourcen. Maßnahmen zur Prüfung und Optimierung auf Korpusebene erscheinen eher realisierbar. So werden am GEI bei der Planung von Digital-Humanities-Projekten, für die Volltexte zur Verfügung gestellt werden sollen, die digitalen Methoden und Werkzeuge für die Forschungsarbeit besprochen und die Voraussetzungen, die die Volltexte erfüllen müssen, eruiert. Liegen digitale Volltexte bereits vor, testet die Bibliothek an Volltext-Stichproben

17 Reul, Christian et al.: State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines, in: DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts (1.0, p. 366), S. 212–216. Online: <https://doi.org/10.5281/zenodo.2596095>, Stand: 26.04.2022.

18 Wick, Christoph; Reul, Christian; Puppe, Frank: Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition, in: Digital Humanities Quarterly 14 (2), 2020, <https://doi.org/10.48550/arXiv.1807.02004>, Stand: 26.04.2022.

gemeinsam mit den Wissenschaftler*innen sowie mit Kolleg*innen, die an der Weiterentwicklung von digitalen Forschungstools arbeiten, die geplanten digitalen Methoden und Werkzeuge. Falls noch keine Digitalisierung und Volltexterkennung erfolgt ist, werden Probedigitalisierungen und entsprechende OCR-Tests durchgeführt. Auf diese Weise können für ein digitales Forschungsprojekt notwendige Zeit- und finanzielle Ressourcen optimiert ermittelt werden. Gegebenenfalls sind Ressourcen zu berücksichtigen, um Volltext-Daten durch Training zu verbessern oder zu korrigieren. Nach Abschluss der Forschungsarbeit stellt die Bibliothek Informationen zur eingesetzten OCR-Technologie sowie OCR-Erkennungsraten für die Forschungsdokumentation zur Verfügung.

Die Sensibilisierung für OCR-Prozesse und insbesondere die Ermittlung und Transparenz von OCR-Fehlerraten waren am GEI Ausgangspunkt für einen über Korpusprioritäten hinausgehenden Verständigungsprozess mit der Forschung. Texterkennung wurde dabei grundlegend für die Evidenz der digitalen wissenschaftlichen Arbeit sowie als partizipative und iterative Daueraufgabe definiert. Diesem Verständnis folgend, erfordert es eine fortwährende enge Zusammenarbeit zwischen Bibliothek und Forschung, wodurch sich die Rolle von Bibliotheken als Dienstleister, die durch Digitalisierung ihre Quellen für die Forschung zugänglich machen, hin zur Rolle, wie sie die Konzepte des „Embedded" oder des „Liaison Librarian" vorsehen, erweitert. Im Hinblick auf die Aufgabenbereiche für „Liaison Librarians"[19] liegen die Potentiale, um sich im Bereich der Digitalisierung neu zu positionieren, insbesondere in der Vermittlung von Kenntnissen von OCR-Prozessen und damit verbundenen Resultaten („Teaching and Learning"), im Austausch über Anforderungen aber auch Grenzen der Volltexterkennung („Scholary Communication"), im Erkennen und Bereitstellen von neuen OCR-Technologien („Digitale Tools") sowie in der Unterstützung der Forschung bei der Beantragung von Fördermitteln („Fund Raising"). Um die Volltexterkennung entlang digitaler Forschungsfragen und -methodiken zu planen, zu konfigurieren, zu realisieren und zu dokumentieren, sollten Bibliotheken Digital-Humanities-Kompetenzen aufweisen und mit der Forschung auf Augenhöhe kommunizieren und agieren. Die fachwissenschaftliche Forschung wiederum sollte die Volltexterkennung verstärkt als Teil des digitalen Forschungsprozesses anerkennen und dessen kritische Reflexion als signifikanten Bestandteil einer digitalen Quellenkritik verstehen.

4. Texterkennung on Demand

Am GEI bringt die Bibliothek ihre OCR-Expertise nicht nur projektspezifisch ein, sie ist darüber hinaus strukturell in das Forschungsteam „Digital Humanities" am Institut eingebunden und kann somit Bedarfe unmittelbar mit der Forschung diskutieren und Impulse setzen. Basierend auf den Digital-Humanities-Forschungen, die besonders auf hochwertige Volltexte angewiesen sind, wurde gemeinsam die Perspektive einer korpusspezifischen Volltexterkennung im Sinne einer *Volltexterkennung on Demand* entwickelt. Ideal wäre eine gezielte Auswahl und Nutzung von OCR-Anwendungen, die sich an Materialbesonderheiten von Korpora und korpusbasierten Forschungsfragen orientieren. Die

19 Fühles-Ubach, Simone: Vom „embedded" zum „liaison librarian" – Was versprechen die neuen Konzepte?, in: Mittermaier, Bernhard (Hg.): Vernetztes Wissen – Daten, Menschen, Systeme. 6. Konferenz der Zentralbibliothek Forschungszentrum Jülich. 5.–7. November 2012, Proceedingsband (Schriften des Forschungszentrums Jülich Reihe Bibliothek / Library, Band / Volume 21), Jülich 2012, S. 337–350, hier S. 346. Online: <https://juser.fz-juelich.de/record/126960/files/FZJ-2012-00028.pdf>, Stand: 26.04.2022.

Organisation und Durchführung der Volltexterkennung würde dabei nicht mehr ausschließlich in der Hand der Bibliothek liegen, Wissenschaftler*innen initiieren vielmehr OCR-Prozesse selbst. Dieser on-Demand-Ansatz hätte den Vorteil, dass die Volltexterkennung gezielt auf Korpus- und Forschungsbedarfe ausgerichtet wäre. Open-Source-Tools speziell für historische Drucke, wie sie im Rahmen des Kooperationsprojekts OCR-D entwickelt werden, stellen für diesen Ansatz eine gute Grundlage dar. Die meisten OCR-Tools sind allerdings nur mit Programmierkenntnissen zu beherrschen, was beispielsweise Forscher*innen am GEI von ihrer Nutzung abhält.[20] An einen dezidiert nicht technisch versierten Nutzerkreis richtet sich die vom Zentrum *für Philologie und Digitalität „Kallimachos" (ZPD)* an der Universität Würzburg entwickelte Software OCR4all (http://www.ocr4all.org/). OCR4all, in dem Calamari vollständig integriert ist, ermöglicht das Preprocessing, also u.a. die Erkennung der Schriftbereiche in binären und graustufigen Bildern und das Umrechnen schief gescannter Textbereiche in gerade Textblöcke, die Region und Line Segmentation, d.h. die Extrahierung der klassifizierten Layoutregionen und Textzeilen, die Textzeichenerkennung (Character Recognition) sowie die Korrektur der erkannten Texte und die Erstellung werkspezifischer OCR-Modelle in einem Trainingsmodul. OCR4all deckt somit den gesamten OCR-Workflow ab.

Nach Tests mit Wissenschaftler*innen aus dem GEI, die die Anwenderfreundlichkeit und Ergebnisqualität von OCR4all an historischen Schulbuchbeständen eruierten, intensivierte die Forschungsbibliothek des GEI die Zusammenarbeit mit den OCR4all-Entwickler*innen. Gemeinsam mit dem ZPD sowie dem Würzburger Lehrstuhl für Mensch-Computer-Interaktion (HCI) wurde das Projekt „OCR4all libraries" initiiert, das von der DFG im Rahmen der OCR-D Förderlinie gefördert wird. Ziel des Projekts ist es, die OCR4all-Software so zu erweitern und anzupassen, dass die OCR-D-Module niederschwellig und eigenständig sowohl auf Korpus- und Werkebene als auch bei größeren Mengen im bibliothekarischen Kontext eingesetzt werden können. Ein hierfür im Projekt geplantes Graphical User Interface (GUI) soll nicht nur Wissenschaftler*innen befähigen, die um die OCR-D-Module angepasste OCR4all-Software intuitiv zu nutzen. Die GUI und eine visuelle Erklärungskomponente zur Erstellung und Konfiguration optimaler OCR-Workflows soll auch Bibliotheken unterstützen, ihre Volltexterkennung flexibler durchzuführen.

Zur Verbesserung der OCR-Qualität bei Drucken in Fraktur und unnormierten Schriftbildern wird im Projekt ein granularer Ansatz erprobt. Es wird dabei ein Verfahren entwickelt, das eine nach Korpora mit jeweils ähnlicher Materialgrundlage organisierte Volltexterkennung erlaubt. Bei Lesebüchern aus dem Deutschen Kaiserreich sind beispielsweise Inhaltsverzeichnisse eine signifikante OCR-Fehlerquelle. Wurden solche Materialspezifika eruiert, können entsprechende OCR-Modelle für eine Datenoptimierung trainiert werden. Von der Generierung und dem Training entlang von Materialspezifika erhoffen sich die Projektpartner*innen von OCR4all libraries eine vielversprechende OCR-Optimierung auch für Massenverfahren. Die Volltexterkennung auf Forschungs- und Materialbedarfe auszurichten, bedeutet, sie nicht nur wie beschrieben partizipativ und iterativ, sondern verstärkt agil zu gestalten. Die Auszeichnung und Erzeugung fehlerfreier Zeichen und Wörter (Ground-Truth-Daten) als Goldstandard sowie eine breite Bereitstellung trainierter OCR-Modelle ist

20 Nieländer; Weiß: Schönere Daten, 2018, S. 91–116, hier S. 97, <https://repository.gei.de/handle/11428/296> (DOI 10.14220/9783737009539), Stand: 26.04.2022.

dabei essenziell, denn dadurch lassen sich Texterkennungsprozesse umfassender bewerten und verbessern. Inwieweit OCR4all-libraries an von Bibliotheken genutzte Digitalisierungsworkflowsysteme angebunden werden kann, wird in einer im Projekt ebenfalls geplanten Machbarkeitsstudie geprüft. In der Entwicklung und Bereitstellung entsprechender Schnittstellen, etwa zu Goobi, Kitodo und DWork, sehen die Projektpartner*innen einen zentralen Meilenstein zur optimalen Unterstützung der Digitalisierung in Bibliotheken.

OCR-Workflows durch eine intuitive Usability eigenständig zu initiieren und spezifisch zu parametrisieren, hätte für Bibliotheken und Wissenschaftler*innen den Vorteil, flexibel auf Forschungsbedarfe und Materialbesonderheiten reagieren zu können. Auch am GEI noch zu klären ist die Nachnutzung bzw. Integration von optimierten Volltext-Daten in Digitalisierungsworkflow- und Präsentationssysteme und die Sicherung und Dokumentation von Volltextversionen. Verschiedene OCR-Textversionen zu dokumentieren, wäre für die Forschung vor allem im Hinblick auf die Zitierbarkeit von Volltexten und damit die Nachvollziehbarkeit der Forschungsergebnisse wichtig.[21] Die Versionen der Volltexte fungieren in diesem Sinne als Forschungsdaten und sind demnach für ein Forschungsdatenmanagement relevant.

Literaturverzeichnis

- DFG: Praxisregeln „Digitalisierung“. DFG Vordruck 12.151 – 12/16, <https://www.dfg.de/formulare/12_151/12_151_de.pdf>, Stand: 26.04.2022.

- DFG: Merkblatt und ergänzender Leitfaden – Digitalisierung und Erschließung, DFG Vordruck 12.15 – 09/21, <https://www.dfg.de/foerderung/programme/infrastruktur/lis/lis_foerderangebote/digitalisierung_erschliessung/formulare_merkblaetter/index.jsp>, Stand: 26.04.2022.

- DFG: Implementierung der OCR-D-Software zur Volltextdigitalisierung. Information für die Wissenschaft Nr. 15 | 27. Februar 2020, <https://www.dfg.de/foerderung/info_wissenschaft/2020/info_wissenschaft_20_15/index.html>, Stand: 26.04.2022.

- DHd (Digital Humanities im deutschsprachigen Raum): AG OCR – Punkt 2: Arbeitsschwerpunkte, <https://dig-hum.de/ag-ocr>, Stand: 26.04.2022.

- Engl, Elisabeth: OCR-D kompakt. Ergebnisse und Stand der Forschung in der Förderinitiative, in: Bibliothek – Forschung und Praxis 44 (2), 2020, S. 218–230. Online: <https://doi.org/10.1515/bfp-2020-0024>.

- Fühles-Ubach, Simone: Vom „embedded“ zum „liaison librarian“ – Was versprechen die neuen Konzepte?, in: Mittermaier, Bernhard (Hg.): Vernetztes Wissen – Daten, Menschen, Systeme. 6. Konferenz der Zentralbibliothek Forschungszentrum Jülich. 5.-7. November

21 Siehe die Umfrage zur Verwendung von OCR-Texten: <https://ocr-d.de/de/umfrage>, Stand: 26.04.2022.

2012, Proceedingsband. (Schriften des Forschungszentrums Jülich Reihe Bibliothek / Library Band / Volume 21). Jülich 2012, S. 337-350. Online: <https://juser.fz-juelich.de/record/126960/files/FZJ-2012-00028.pdf>, Stand: 26.04.2022.

- Gasser, Sonja: Das Digitalisat als Objekt der Begierde. Anforderungen an digitale Sammlungen für Forschung in der Digitalen Kunstgeschichte, in: Andraschke, Udo; Wagner, Sarah (Hg.): Objekte im Netz. Wissenschaftliche Sammlungen im digitalen Wandel, Bielefeld 2020, S. 261–276. Online: <https://doi.org/10.14361/9783839455715>.

- Hertling, Anke; Klaes, Sebastian: Historische Schulbücher als digitales Korpus für die Forschung. Auswahl und Aufbau einer digitalen Schulbuchbibliothek, in: Nieländer, Maret; De Luca, Ernesto William (Hg.): Digital Humanities in der internationalen Schulbuchforschung. (Eckert. Expertise 9). Göttingen 2018, S. 22-44. Online: <https://repository.gei.de/handle/11428/296> (DOI 10.14220/9783737009539), Stand: 26.04.2022.

- Jacobmeyer, Wolfgang: Das deutsche Schulgeschichtsbuch 1700-1945. Die erste Epoche seiner Gattungsgeschichte im Spiegel der Vorworte, Bd. 1, Berlin 2011.

- Jäger, Georg: Der Schulbuchverlag, in: Ders. et al. (Hg.): Geschichte des deutschen Buchhandels im 19. und 20. Jahrhundert, Bd. 1: Das Kaiserreich 1870-1918, Teil 2, Frankfurt am Main 2003.

- Nieländer, Maret; Weiß, Andreas: »Schönere Daten« – Nachnutzung und Aufbereitung für die Verwendung in Digital-Humanities-Projekten, in: Nieländer, Maret; De Luca, Ernesto William (Hg.): Digital Humanities in der internationalen Schulbuchforschung. (Eckert. Expertise 9), Göttingen 2018, S. 91–116. Online: <https://repository.gei.de/handle/11428/296> (DOI 10.14220/9783737009539), Stand: 26.04.2022.

- Reul, Christian; Springmann, Uwe; Wick, Christoph; Puppe, Frank: State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines, in: DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts (1.0, p. 366), S. 212–216, <https://doi.org/10.5281/zenodo.2596095>.

- Weil, Stefan: Neue Frakturmodelle für Tesseract. Präsentation auf dem Kitodo Anwendertreffen 18.–19. November 2019, S. 3. Online: <https://madoc.bib.uni-mannheim.de/53748/1/2019-11-18.pdf>, Stand: 26.04.2022.

- Weil, Stefan: tesseract-ocr / tesstrain, <https://github.com/tesseract-ocr/tesstrain/wiki>, Stand: 26.04.2022.

- Wick, Christoph; Reul, Christian; Puppe, Frank: Calamari – A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition, in: Digital Humanities Quarterly 14 (2), 2020. Online: <https://doi.org/10.48550/arXiv.1807.02004>.