Volltexte – die Zukunft alter Drucke
Bericht zum Abschlussworkshop des OCR-D-Projekts

Das DFG-geförderte Projekt OCR-D (https://ocr-d.de) befasst sich seit 2015 mit der Verbesserung von Verfahren zur automatischen Text- und Strukturerfassung historischer Drucke. Übergeordnetes Ziel des Projekts ist es, die technische und konzeptionelle Volltexttransformation der Drucke vorzubereiten, die in den Verzeichnissen der im deutschen Sprachraum erschienenen Drucke (VD16, VD17 und VD18) erfasst sind.

Die zweite Phase des OCR-D-Projekts unter Koordination der Berlin-Brandenburgischen Akademie der Wissenschaften in Berlin (BBAW), der Herzog August Bibliothek Wolfenbüttel (HAB), des Karlsruher Instituts für Technologie (KIT) und der Staatsbibliothek zu Berlin Preußischer Kulturbesitz (SBB) endet in der ersten Hälfte des Jahres 2020. Aus diesem Anlass fand am 12. Februar im Wissenschaftszentrum in Bonn der Workshop "Volltexte – die Zukunft alter Drucke" statt, in dem Erkenntnisse und Desiderate des OCR-D-Projekts vorgestellt und diskutiert wurden. Die Präsentationen zum Workshop sind auf der Projektwebsite[1] eingestellt.

Nach einem kurzen Überblick der Projektkoordinatorin Elisabeth Engl (HAB Wolfenbüttel) zum derzeitigen Stand des Projekts stellten Konstantin Baierer und Clemens Neudecker (SBB Berlin) die OCR-D-Software mit ihren Funktionen und Möglichkeiten vor.

Die OCR-D-Software setzt auf etablierte Standards wie METS auf. Dadurch wird die maximale Kompatibilität mit digitalisierten Bibliotheksbeständen, die gemäß den DFG-Praxisregeln Digitalisierung[2] erzeugt wurden, gewährleistet. Zudem können die Ergebnisse von OCR-D mit dem DFG-Viewer verwendet werden. Als Ausgabeformat für OCR-Ergebnisse wird das PAGE-XML-Format genutzt, das eine größere Annotationstiefe als bspw. das im Bibliotheksbereich verbreitete ALTO[3] bietet. Grundsätzlich folgt die Entwicklung von Spezifikationen und Software im OCR-D-Projekt dem FLOSS Prinzip (free, libre, open source software). Außerdem werden konsequent offene Plattformen wie GitHub genutzt, wodurch die Entwicklungsarbeiten transparent gemacht sowie Interessierte und Anwender frühzeitig eingebunden werden können. Diesen Ansatz verfolgt OCR-D auch in Kooperationen mit weiteren OCR-Partnern wie der DHd (Digital Humanities im deutschsprachigen Raum; Gründung der DHd AG OCR)[4] oder der weit verbreiteten, freien OCR-Software Tesseract (Übernahme des OCR-D Werkzeugs zum Training[5] neuer Modelle[6] ins Tesseract-Projekt). An der Entwicklung des modular aufgebauten OCR-D-Prototypen waren neben dem OCR-D-Koordinierungsprojekt acht Modulprojekte[7] beteiligt.

1 OCR-D. Publikationen, <https://ocr-d.de/de/publications>, Stand: 23.03.2020.
2 DFG. DFG-Praxisregeln "Digitalisierung", <https://www.dfg.de/formulare/12_151/>, Stand: 23.03.2020.
3 The Library of Congress. ALTO. Technical Metadata for Layout and Text Objects, <https://www.loc.gov/standards/alto/>, Stand: 25.03.2020.
4 DHd AG OCR, <https://dhd-ag-ocr.github.io/>, Stand: 23.03.2020.
5 GitHub. Tesstrain, <https://github.com/tesseract-ocr/tesstrain>, Stand: 23.03.2020.
6 Für die Texterkennung müssen Modelle aus umfangreichen Trainingsdaten erstellt werden. Die Modelle lernen dabei aus den Trainingsdaten (Bilddaten und dazugehörige Transkriptionen) Merkmale und deren Wahrscheinlichkeiten. So lassen sich anschließend weitere Eingabedaten (Bilder) automatisiert auf Grundlage der im Modell gespeicherten Merkmale und Wahrscheinlichkeiten in die gewünschten Ausgabedaten (Volltexte) übersetzen.
7 OCR-D. Modulprojekte, <https://ocr-d.de/de/module-projects>, Stand: 23.03.2020.

Für diese inzwischen weitgehend abgeschlossenen Arbeiten wurden zum einen bestehende Werkzeuge übernommen und angepasst, zum anderen neue Werkzeuge implementiert. Die lauffähigen Werkzeuge sind auf GitHub[8] frei zugänglich und können getestet werden.

Mit Blick auf diese Werkzeuge wurde diskutiert, inwieweit die großen Herausforderungen im Bereich der Volltexterkennung im Rahmen der DFG-geförderten OCR-D-Initiative gelöst werden konnten. Die Entwickler/innen der acht OCR-D-Modulprojekte sehen insbesondere bei der Layouterkennung und der Nachkorrektur noch weiteren Forschungsbedarf. Eine vollautomatische Layouterkennung ist nach dem derzeitigen Forschungsstand noch immer eine große Herausforderung, der gleichzeitig aber als Grundlage für die eigentliche Texterkennung große Bedeutung zukommt. In diesem Bereich könnte ein Austausch zwischen OCR-Entwicklern und Geisteswissenschaften fruchtbar sein. Die spezifischen Kenntnisse der Wissenschaftler*innen über die zu prozessierenden Vorlagen, insbesondere deren Layouts, könnten für die automatische Text- und Strukturerkennung nutzbar gemacht werden. Die für die Nachkorrektur der OCR-Ergebnisse trainierten Modelle sind derzeit nur für Drucke aus einem jeweils recht begrenzten Zeitraum erfolgreich einsetzbar. Hier stellt sich zudem die Grundsatzfrage, an welcher Stelle des OCR-Prozesses Erkenntnisse über (potenzielle) Fehler bzw. Fehlerquellen zur Verbesserung der Texterkennung eingesetzt werden sollten. Dieses Wissen, bspw. über häufig verwechselte Zeichen, ist unabdingbar für eine erfolgreiche Nachkorrektur. Es könnte jedoch bereits für die vorhergehenden Schritte der Texterkennung genutzt werden und die Nachkorrektur vorwegnehmen.

Auf weitere Herausforderungen, die auch andere Projekte im Bereich der OCR betreffen, ging Matthias Boenig (BBAW Berlin) ein. Beispielsweise wurde im Zuge der Erstellung von Ground-Truth[9] deutlich, dass mit dem PAGE-Format zwar ein dafür passendes Format vorhanden ist, entsprechende Richtlinien jedoch fehlen. Auf Grundlage des Basisformats des deutschen Textarchivs (DTABf)[10] hat das OCR-D-Projekt Richtlinien mit drei gestuften Leveln[11] formuliert. Mit deren Hilfe können transkribierte Texte u.a. aus Editions-Projekten nach ihrer Kuratierung sowie Umwandlung in das PAGE-XML-Format als Ground-Truth angeboten werden. Die Erfassung von neuem Ground-Truth ist schwierig, da es nur wenige Dienstleister gibt, die diesen in der geforderten Qualität erstellen können.

Die Nachmittagssektion drehte sich insbesondere um Fragen der Überführung der OCR-D-Software in den praktischen Einsatz in Bibliotheken bzw. anderen bestandshaltenden Einrichtungen. Wie Elisabeth Engl in ihrem Vortrag zur *Bibliothekarischen Digitalisierungspraxis* ausführte, haben die im Bereich der (Bild-)Digitalisierung stark bis mittel engagierten deutschen Bibliotheken entgegen ihrem großen Interesse an OCR zu frühneuzeitlichen Büchern bisher nur wenige diesbezügliche Projekte durchgeführt. Die Anforderungen der Bibliotheken an eine OCR-Software können von OCR-D im

8 GitHub. OCR-D, <https://github.com/OCR-D>, Stand: 23.03.2020.

9 OCR-D versteht unter Ground Truth fehlerfreie Transkriptionen, aus denen die für die Volltexterkennung benötigten Trainings- und Testdaten erstellt werden können.

10 Das Basisformat in TEI-XML wird u.a. von der DFG zur Transkription historischer Texte empfohlen. Auf Grundlage dieses Formats wurden die vom DTA bereitgestellten Volltexte annotiert. Vgl. <http://www.deutschestextarchiv.de/doku/basisformat/einfuehrung>, Stand: 25.03.2020.

11 OCR-D. Die Ground-Truth-Guidelines, <https://ocr-d.de/de/gt-guidelines/trans/>, Stand: 23.03.2020.

derzeitigen Stadium nur teilweise erfüllt werden. Alle von den Bibliotheken formulierten Bedarfe werden von der Förderinitiative jedoch bereits in den Projektplanungen berücksichtigt und sollen perspektivisch umgesetzt werden. Dass die OCR-D-Software schon jetzt als robuster, lauffähiger Prototyp verfügbar ist, hat eine erste Teststellung in neun Pilotbibliotheken bewiesen.

Die Workshop-Teilnehmer sehen das fortgeschrittene Entwicklungsstadium der Software als wichtigen Impuls, OCR-D zeitnah in verschiedenen Einrichtungen zu implementieren. Die im Praxiseinsatz gewonnenen Erkenntnisse könnten für deren gemeinschaftliche, perspektivisch durch eine Community getragene Weiterentwicklung genutzt werden. Durch langes Warten auf die *Fertigstellung* eines OCR-D-Produkts könnte womöglich der Anschluss an die aktuellen Veränderungen im Wissenschafts- und Bibliotheksbereich verloren werden. Diese Vorgehensweise baut auf einen intensiven Erfahrungsaustausch zwischen den beteiligten Institutionen, bspw. bei gemeinsamen Workshops und über standardisierte Protokolle. Gerade vorkonfigurierte Workflows könnten effektiv im Verbund erarbeitet werden. Diese werden benötigt, um die OCR-D-Software ressourcensparend auf einer breiten Auswahl an Digitalisaten einsetzen zu können, und wurden bei ersten Praxistests des OCR-D-Prototypen vermisst. Zudem sind Fragen zur Qualitätsanalyse der erstellten Volltexte sowie der absehbaren kontinuierlichen Neuprozessierung von Digitalisaten mit verbesserten Algorithmen zu klären – Herausforderungen, die alle OCR-Lösungen bzw. -projekte gleichermaßen betreffen und nur gemeinschaftlich gemeistert werden können.

Für einen baldigen Einsatz der OCR-D-Software spricht auch die große Nachfrage nach Volltexten, die von verschiedensten Nutzergruppen an bestandshaltende und -verarbeitende Einrichtungen herangetragen wird. Nicht nur die häufig als Hauptzielgruppe der Volltexte angesehenen Digital Humanities, sondern eine deutlich breitere Auswahl an Forschenden aus traditionellen Geisteswissenschaften, aber auch bspw. Wirtschafts- und Materialwissenschaften, haben großes Interesse an Volltexten. Neben historischen Materialien werden graue Literatur oder freigekaufte Publikationen als Volltexte gewünscht. An die Volltexte bestehen je nach Nutzergruppe unterschiedliche Qualitätsanforderungen. Auch Volltexte mit Erkennungsraten von ca. 90 %, die noch deutlich unter dem gemeinhin als wissenschaftlich brauchbar angesehenen Niveau liegen, könnten durch eine Vielzahl von Forschenden für die Volltextsuche genutzt werden. Dafür ist es jedoch wichtig, dass die OCR-Fehler bei der Weiterverarbeitung der Volltexte mit Verfahren der Indexierung, des Text Mining etc. berücksichtigt werden. Die OCR-Entwicklung dürfte bald ein Stadium erreicht haben, in dem Volltexterkennung für historische Drucke zum Standard in Digitalisierungsprojekten wird.

Insgesamt sehen Bibliotheken die Bereitstellung von Volltexten für ihre Nutzer*innen als strategisch wichtige Aufgabe und streben an, OCR-Expertise in ihren Häusern weiter auszubauen. Dieser Service könnte für Bibliotheken ein weiterer Schritt in Richtung einer gleichberechtigten Partnerschaft mit Wissenschaftler*innen für die Forschung sein. Die fortgeschrittene Entwicklung der OCR-D-Software sowie die neue Ausschreibung der DFG zu deren Implementierung[12] könnten für Bibliotheken

12 DFG. Implementierung der OCR-D-Software zur Volltextdigitalisierung, <https://www.dfg.de/foerderung/info_wissenschaft/info_wissenschaft_20_15/index>, Stand: 23.03.2020.

und weitere Einrichtungen zu einem wichtigen Anstoß werden, diese Visionen für ihre historischen Bestände nun in einer gemeinschaftlichen Anstrengung in die Praxis umzusetzen.

Elisabeth Engl, Herzog August Bibliothek Wolfenbüttel
Konstantin Baierer, Staatsbibliothek zu Berlin
Matthias Boenig, Berlin-Brandenburgische Akademie der Wissenschaften
Volker Hartmann, Karlsruher Institut für Technologie
Clemens Neudecker, Staatsbibliothek zu Berlin

Zitierfähiger Link: https://doi.org/10.5282/o-bib/5600