Volltexttransformation frühneuzeitlicher Drucke – Ergebnisse und Perspektiven des OCR-D-Projekts

Boenig, Matthias
boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Engl, Elisabeth
engl@hab.de
Herzog-August-Bibliothek Wolfenbüttel, Deutschland

Baierer, Konstantin
konstantin.baierer@sbb.spk-berlin.de
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Deutschland

Hartmann, Volker
volker.hartmann@kit.edu
Karlsruher Institut für Technologie

Neudecker, Clemens
clemens.neudecker@europeana-newspapers.eu
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Deutschland

Einleitung

Das schriftliche Kulturgut des deutschsprachigen Raums aus dem 16.–18. Jahrhundert wird schon seit Jahrzehnten in den Verzeichnissen der im deutschen Sprachraum erschienenen Drucke (VD) zusammengetragen. Ein signifikanter Anteil der verzeichneten Titel wurde der Forschung bereits durch die Bereitstellung von Volldigitalisaten oder einzelnen Schlüsselseiten leichter zugänglich gemacht. Die Verfügbarmachung von Volltexten ist dagegen noch ein Desiderat der Forschung. Das DFG-Projekt OCR-D nimmt sich seit Oktober 2015 im Rahmen der Koordinierten Förderinitiative zur Weiterentwicklung von Verfahren für die Optical Character Recognition (OCR) dieser Aufgabe an, indem es eine modular aufgebaute Open Source-Software entwickelt, deren Werkzeuge alle für die Texterkennung nötigen Schritte abdecken sollen. Der modulare Ansatz ermöglicht es, die technischen Abläufe und Parameter der Texterkennung stets nachzuvollziehen und maßgeschneiderte Workflows zu definieren, die jeweils optimale Ergebnisse für spezifische Titel aus dem Zeitraum des 16. bis 19. Jahrhunderts liefern. Zudem werden Antworten auf die damit verbundenen konzeptionellen, informationswissenschaftlichen und organisatorischen Fragen gefunden.

Künftig sollen mithilfe der OCR-D-Software Volltexte generiert werden, die zum einen von Forschenden zur Recherche verwendet werden können. Zum anderen könnten diese zum Ausgangspunkt für Studien im Bereich der Digital Humanities (DH) werden, wobei auch auf diese Texte die textkritische Methode anzuwenden ist. Gerade bei einer automatisierten Weiterverarbeitung der erzeugten Volltexte ist es für Forschende unerlässlich, die Genese der von ihnen verwendeten Daten kritisch zu hinterfragen. Nur so können Eigenheiten der Daten, die Resultat von zuvor genutzten “Spielräumen” sind, von DH-Forschenden erkannt und in ihrem Umgang mit der Datengrundlage berücksichtigt werden. Nicht nur diese interpretatorischen Spielräume sind zu betrachten, sondern auch, welche konkreten Implementierungen den DH die gewünschten “Spielräume“ für die Erkenntnisgenerierung geben. Im Folgenden wird in vier Thesen eine notwendige Begrenzung der Spielräume vorgenommen. Diese Begrenzung ergibt sich aus dem Vergleich mit anderen Projekten und der heute gängigen Praxis. Ziel ist es, den Forderungen der DH nach qualitativ hochwertigen Volltexten gerecht zu werden.

Im Rückblick

Das Projekt hat sich in den vergangenen vier Jahren mit verschiedenen Themen auf der DHd zur Diskussion gestellt (Boenig et al 2016; Boenig et al 2018; Baierer et al 2019). Zu Beginn standen methodische Fragen, wie die Textqualität erhöht werden kann. Dabei wurden statistische Methoden vorgestellt, die auf Basis eines Vergleichs von mindestens zwei erstellten Textfassungen entwickelt wurden. Im Rahmen des Themas “Kritik der digitalen Vernunft” wurden die DH befragt, wie in den Geisteswissenschaften Ergebnisse ohne Ground Truth und Referenzdaten gewonnen bzw. verifiziert werden. Diesem Desiderat begegnete das Projekt OCR-D mit dem Vorschlag von Transkriptionsrichtlinien für die Erfassung von Ground Truth-Daten[1] und in der Folge mit der Definition von spezifischen Metadaten. Bei dem 2019 veranstalteten Workshop konnten Wissenschaftler und Wissenschaftlerinnen sowie Interessierte Einblicke in den OCR-D-Workflow erhalten. An Beispielen konnten die Möglichkeiten der Software demonstriert und getestet werden. Die Diskussion, Hinweise und Fragen wurden soweit wie möglich in OCR-D umgesetzt.

Thesen

Das Ziel der prototypischen Implementierung des OCR-D-Workflows und damit der Generierung von Forschungsdaten, die sich durch eine erkennbare XML-

Strukturierung sowie eine hohe Zeichen- und Textqualität auszeichnen, wird im ersten Quartal 2020 erreicht werden. Dies stellt jedoch nicht das Ende des Weges dar, sondern eher den Beginn der nun folgenden Volltexttransformation. Letztlich besteht die Aufgabe darin, ca. 1 Mio. frühneuzeitliche Titel mit ca. 250 Mio. Seiten, die zum Teil bereits als Bilddigitalisate vorliegen, zu Volltextdigitalisaten zu transformieren.

1. Die Volltexttransformation der Bestände stellt eine Herausforderung für Bibliotheken und Archive dar. Die vorhandenen institutionellen und interinstitutionellen Vorgehensweisen und Konventionen sind möglichst zentral aufeinander abzustimmen, damit die Aufgabe in absehbarer Zeit gelöst wird.[2]

Es gibt bereits einige Projekte, in denen (Teil-)Bestände und Sammlungen volltextdigitalisiert wurden.[3] Deren Nutzen für die DH wird jedoch v.a. durch zwei Faktoren begrenzt: Zum einen weisen die erstellten Volltexte aufgrund fehlender Standards bzw. Konventionen im Bereich von Text- und Strukturerkennung eine große Bandbreite in der Transkription der Texte und der Benennung von Textstrukturen auf, die deren automatisierte Auswertung und Bearbeitung durch die DH erschweren. Zum anderen gibt es bislang keine zentrale Anlaufstelle, die die Bereitstellung und auch die Erstellung von Volltexten steuert. Dadurch sind die existierenden Volltexte sowohl für die Forschung, als auch für die volltextdigitalisierenden Einrichtungen weniger sichtbar, was die Gefahr aufwändiger und teurer Doppelarbeiten erhöht.

2. Die Volltexttransformation auf Basis von Erkennungssoftware, die neuronale Netze nutzt, setzt Trainingsdaten voraus. Diese fundamental wichtigen Daten sind systematisch aus vorhandenen Ressourcen zu gewinnen und aktiv zu erweitern.

Mit ihren Förderinitiativen von 2010 und 2013 hat die DFG die Bedeutung der Forschungsdaten und des zugehörigen Managements erkannt.[4] Heute sollten Projekte von Beginn an mit entsprechenden Forschungsdatenmangementplänen aufgesetzt und die entstehenden Daten in den zuvor bereitgestellten Repositorien verwahrt werden.[5] Gerade bei der automatisierten Texterfassung im Rahmen von Editionsprojekten werden in der Regel aber nur die abschließend bearbeiteten und korrigierten Daten veröffentlicht. Eine Nachnutzung dieser Daten ist in vielfacher Hinsicht nur begrenzt möglich. Dabei spielt nicht nur das Format der Daten, sondern auch die Methodik der Datenerfassung eine entscheidende Rolle. Für die Nachnutzung ist eine Transformation dieser Daten nötig, die entweder von den Nutzenden zu leisten ist, oder von den bestandsverwaltenden Einrichtungen angeboten werden könnte. Um eine solche Transformation zu gewährleisten, sind sowohl Richtlinien als auch entsprechende Metadaten zu etablieren, damit vergleichbare und konsistente Daten bereitgestellt werden können.[6]

3. Die Volltexttransformation wird für einen Teil der Dokumente ein Prozess sein, der sich über einen größeren Zeitraum wiederholt.

Digitale Daten müssen beständig gepflegt und aktualisiert werden. Dies haben auch Bibliotheken als Herausforderung der digitalen Transformation ihrer Bestände erkannt (vgl. Kempf 2015: 277–278). Werden lernende Systeme für die Text- und Strukturerkennung genutzt, können diese in absehbaren Intervallen verbessert werden.[7] Denn die Verbesserung bestehender Algorithmen sowie die Nutzung zusätzlicher oder verbesserter Trainingsdaten führt auch zu besseren Ergebnissen in der Text- und Strukturerkennung, wie sich beispielsweise im GoogleBooks-Projekt[8] zeigt. Diese wiederkehrende Prozessierung muss konzeptionell berücksichtigt werden.

4. Die Volltexttransformation muss in ihrer Qualität von den Nutzenden beurteilbar sein.

Bibliotheken geben den Nutzenden mit ihrem Bestand und dessen Erschließung ein Qualitätsversprechen. Die Nutzenden können sich auf die vorhandenen Daten verlassen und sie z.B. in Bibliographien verwenden. Das Volltextangebot aus der automatischen Texterkennung kann dagegen häufig nur unpräzise als “schmutzige OCR”[9] bezeichnet werden. Diese pauschale Angabe ermöglicht den DH keine verlässliche Qualitätseinschätzung und führt dazu, dass Volltextbestände oft a priori als minderwertig eingeschätzt werden. Daher besteht die Gefahr, dass projektintern eine erneute Volltextdigitalisierung durchgeführt wird, die nicht immer sinnvoll ist, da die Erkennung teilweise nur durch eine Korrektur verbessert werden könnte. Oder es könnten im umgekehrten Fall auf Grund einer ungenauen bzw. zu groben Einschätzung aufwendige Korrekturen vorgenommen werden. In beiden Fällen werden finanzielle und personelle Ressourcen verschwendet.

Lösungen und Desiderate des OCR-D-Projekts

Zu 1: Die bisherigen umfassenden Bilddigitalisierungsarbeiten im VD17 wurden über einen Masterplan gesteuert, um die große Anzahl an Titeln effizient, in nachnutzbarer Form verarbeiten zu können und Doppelarbeiten zu vermeiden. Ein ähnliches Vorgehen, bei dem die zu prozessierenden Titel an interessierte Einrichtungen verteilt werden, dürfte auch für die Volltexttransformation der VD zielführend sein. Die Voraussetzungen und Rahmenbedingungen für diese Arbeiten wurden von dem OCR-D-Koordinierungsprojekt um die Jahreswende 2019/2020 durch eine Umfrage mit den VD-Bibliotheken zusammengetragen. OCR-D wird die mehrjährige Projekterfahrung im Austausch mit den verschiedenen Stakeholdern nutzen, um die Nachnutzbarkeit von Daten und Abläufen zu verbessern, sowohl mit technischer Dokumentation und Best Practices,

als auch als Katalysator für einen ergebnisorientierten, inklusiven Diskurs zur Etablierung von Standards.

Zu 2: Für die Transkription von Texten gibt es unzählige Richtlinien, die von verschiedenen Fächern, Arbeitskreisen und Forschungsprojekten entsprechend ihrer jeweiligen Anforderungen aufgestellt und wiederum an die spezifischen Erfordernisse bestimmter Transkriptionsprojekte angepasst wurden. Bei diesen Gruppen ist zum einen ein Bewusstsein dafür zu schaffen, ihre Transkriptionen auch mit Blick auf deren Nachnutzbarkeit durch andere Projekte anzufertigen. Zum anderen sind interdisziplinär erarbeitete und gültige Transkriptionsrichtlinien ein großes Desiderat der Forschung. Erste Impulse hierfür könnten große Fördergeber wie bspw. die DFG geben, indem Praxisrichtlinien geschaffen werden, die von Antragstellern zu beachten sind. Das OCR-D-Projekt ist zudem darum bemüht, seine auf Grundlage des DTA-Basisformats erstellten Transkriptionsrichtlinien interdisziplinär zur Nutzung durch weitere Projekte zu kommunizieren.

Zu 3: Modelltraining mit tesstrain und okralact

Das Projekt ocropy, die Python-Implementierung von Tom Breuels OCRopus-Projekt, brachte neben Werkzeugen für die Text- und Strukturerkennung auch Werkzeuge für das Erstellen von Ground Truth und das Trainieren neuer Modelle mit sich. Mit diesen Werkzeugen und einigen Anpassungen lassen sich auch die auf ocropy basierenden Weiterentwicklungen Calamari und Kraken trainieren. Insbesondere für tesseract, die mit Abstand am meisten genutzte Open Source OCR, gab es bis 2018 kaum Dokumentation oder Tooling für das Training. Daher wurde im Rahmen von OCR-D *ocrd-train* entwickelt, eine Makefile-basierte Lösung zum Trainieren von Tesseracts LSTM-Engine, das inzwischen unter dem Namen *tesstrain* vom Tesseract-Entwicklerteam gepflegt und weiterentwickelt wird.[10] Die Aufrufe zum Training von Texterkennungsmodellen und insbesondere das Inventar an freien Parametern sind allerdings in hohem Maße engine-spezifisch, keineswegs trivial und erfordern zur optimalen Feinadjustierung manuelle Intervention. Daher entwickelt OCR-D seit 2019 das Werkzeug okralact,[11] das über ein komfortables Webinterface und ein skalierbares Backend ein Training aller relevanter Open Source OCR-Engines mit einem einheitlichen Interface ermöglichen wird.

Zu 4: Nachkorrektur und Qualitätsanalyse

Innerhalb des OCR-D-Projektes beschäftigen sich zwei Projekte mit der automatischen, bzw. semi-automatischen Nachkorrektur von OCR-Texten. Das Hauptproblem dabei ist es, historische Schreibweisen und Druckfehler von OCR-Fehlern zu unterscheiden. Für moderne Texte würde eine reine Rechtschreiberkennung genügen, wie sie in jedem Textverarbeitungsprogramm verfügbar ist. Die Projekte kooperieren und haben verschiedene Verfahren entwickelt, basierend auf einem Fehler-Profiler, neuronalen Netzen oder endlichen Automaten. Als trainierbare Algorithmen werden sie, analog zur Struktur- und Texterkennung, mit mehr und besseren Trainingsdaten bessere Ergebnisse liefern. Was "besser" bedeutet ist noch Gegenstand der Forschung. OCR-D bringt sich in die Entwicklung ein und unterstützt tatkräftig Projekte wie dinglehopper[12] (ein Werkzeug zur Fehlervisualisierung). Gerade im Bereich der Ground-Truth-freien Evaluation von Text und der Qualitätsanalyse von Strukturdaten gibt es noch große Lücken im Software-Portfolio, die zu schließen sich OCR-D auch weiterhin befleißigen wird.

Ausblick

Ab der ersten Jahreshälfte 2020 werden die entwickelten Software-Komponenten im OCR-D-Workflow verankert sein. Damit tritt diese Software immer mehr aus dem Projektstadium heraus und wird in den produktiven Einsatz überführt. Um kontinuierlich gute Erkennungsergebnisse mit dem aus fast vier Jahrhunderten stammenden Material zu erhalten, sind Optimierungen notwendig. Dabei wird stets darauf abgezielt, Forschungsdaten aus den digitalen Beständen der Bibliotheken zu erzeugen und nicht unstrukturierte Textdaten. So wird die Volltexttransformation in einem umfassenden Maße Grundlagen für datenzentrierte Digital Humanities schaffen.

Fußnoten

1. Im Kontext von OCR bezeichnet *Ground Truth* manuell korrigierte, fehlerfreie Transkriptionen. Diese werden zum einen für das Training von OCR-Engines, zum anderen für die Evaluation der OCR-Ergebnisse benötigt.
2. Der Gedanke folgt der neunten Empfehlung (“Establish an ‘OCR Service Bureau’”) aus dem Report von Smith und Cordell (2018).
3. Vgl. bspw. die folgenden Projekte, die sich auf unterschiedlich große (Teil-)Bestände beziehen: Helmstedter Drucke Online: http://www.hab.de/de/home/wissenschaft/forschungsprofil-und-projekte/helmstedter-drucke-online.html ; Über 14.000 preußische Drucke des 17. Jahrhunderts online verfügbar: https://blog.sbb.berlin/ueber-14-000-preussische-drucke-des-17-jahrhunderts-online-verfuegbar/ ; Projekt Digi20 https://digi20.digitale-sammlungen.de/de/fs1/about/static.html
4. Nachdem im Jahr 2010 der Aufbau von Infrastrukturen für Forschungsdaten von der DFG ausgeschrieben worden war, wurde drei Jahre später das Förderprogramm „Informationsinfrastrukturen für Forschungsdaten“ eingerichtet. Vgl. DFG 2019: 7.
5. Zur aktuellen Situation des Datenmanagements und der Rolle, die Bibliotheken in diesem Bereich einnehmen (könnten), vgl. Neuroth et al 2019.
6. Die Notwendigkeit einheitlicher Richtlinien wird besonders an Projekten wie “Venice Time Machine” deutlich, dessen bereits vorhandenen 8 TB an Daten aufgrund fehlender einheitlicher Richtlinien und

Vorgehensweisen bei der Erfassung der Metadaten für die Forschung vermutlich wertlos sind. Vgl. Castelvecchi 2019: 607.
7. Kempf geht davon aus, dass mit OCR-Software nie völlig fehlerfreie Volltexte generiert werden können. Vgl. Kempf 2015: 274.
8. Während die OCR-Ergebnisse im Rahmen des GoogleBooks-Projekts zunächst insgesamt unbefriedigend, für gebrochene Schriften vollkommen unbrauchbar waren, konnten ab dem Jahr 2008 einzelne Frakturtexte in ausreichender Qualität prozessiert werden. In den letzten Jahren konnte die Erkennungsrate noch deutlich gesteigert werden. Vgl. Wikisource: Google Book Search.
9. Bspw. gibt Google die Fehlerquote im Google Books pauschal mit 1,37 % an (vgl. Kempf 2015: 272). Diese für die wissenschaftliche Nutzung hohe Fehlerrate unterscheidet sich, bedingt durch die Vielfalt an Typen und Layouts sowie den großen Publikationszeitraum der digitalisierten Bücher, von Text zu Text deutlich.
10. https://github.com/tesseract-ocr/tesstrain
11. https://github.com/OCR-D/okralact
12. https://github.com/qurator-spk/dinglehopper

Bibliographie

Baierer, Konstantin / Boenig, Matthias / Hartmann, Volker / Hermann, Elisa / Neudecker, Clemens (2019): *„Vom gedruckten Werk zu elektronischem Volltext als Forschungsgrundlage“* (Workshop) (https://zenodo.org/record/2596095/files/2019_DHd_BookOfAbstracts_web.pdf , S. 58).

Boenig, Matthias / Würzner, Kay-Michael / Binder, Arne / Springmann, Uwe (2016): *„Über den Mehrwert der Vernetzung von OCR-Verfahren zur Erfassung von Texten des 17. Jahrhunderts.“* Vortrag auf der DHd 2016, 7.12.03.2016 in Leipzig (http://dhd2016.de/boa.pdf#page=103).

Boenig, Matthias / Federbusch, Maria / Herrmann, Elisa / Neudecker, Clemens / Würzner, Kay-Michael (2018): *„Ground Truth: Grundwahrheit oder Ad-Hoc-Lösung? Wo stehen die Digital Humanities?“*. Vortrag auf der DHd2018, 28.02.2018 in Köln (http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf#page=221).

Castelvecchi, Davide (2019): *“Venice ‘Time Machine’ Project Suspended amid Data Row. Disagreements between International Partners Leave Plans to Digitize the Italian City’s History in Limbo”* in: *Nature* 574: 607.

DFG (2019): *„Weiterentwicklung des Förderprogramms ‚Informationsinfrastrukturen für Forschungsdaten‘"* https://zenodo.org/record/2650866 [6.3.2019 / 26.9.2019].

Kempf, Klaus (2015): *„Data Curation oder (Retro-)Digitalisierung ist mehr als die Produktion von Daten“* in: *o-bib* 4: 268–278.

Neuroth, Heike / Rothfritz, Laura / Petras, Vivien / Kindling, Maxi (2019): “Digitales Datenmanagement als neue Aufgabe für wissenschaftliche Bibliotheken” in: Bibliothek. Forschung und Praxis 43: 421–431.

Smith, David A. / Cordell, Ryan (2018): *“A Research Agenda for Historical and Multilingual Optical Character Recognition”* http://hdl.handle.net/2047/D20297452 [9.12.2019].

Wikipedia, Die freie Enzyklopädie (2019): *„Google Books“*. https://de.wikipedia.org/w/index.php?title=Google_Books&oldid=189583765 [16.6.2019 / 25.9.2019].

Wikisource (2019): *“Google Book Search”* https://de.wikisource.org/wiki/Wikisource:Google_Book_Search [22.8.2019 / 26.9.2019].