Das richtige Tool für die Volltextdigitalisierung

Baierer, Konstantin
konstantin.baierer@sbb.spk-berlin.de
Staatsbibliothek zu Berlin, Deutschland
ORCID: 0000-0003-2397-242X

Hinrichsen, Lena
hinrichsen@hab.de
Herzog August Bibliothek, Deutschland
ORCID: 0000-0002-9286-2390

Boenig, Matthias
boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0003-4615-4753

Reul, Christian
christian.reul@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland
ORCID: 0000-0002-1776-1469

Sautter, Lilja
sautter@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland

Mustafa, Mehmed
mehmed.mustafa@gwdg.de
Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen, Deutschland

Will, Larissa
larissa.will@uni-mannheim.de
Universität Mannheim, Deutschland
ORCID: 0009-0004-6220-8939

Sowohl Einrichtungen als auch Forschende, die Volltexte generieren möchten, stehen vor einer Vielzahl an Tools, die entweder Open Source oder kostenpflichtig sind. Jedes der Tools bringt spezifische Vor- und Nachteile mit sich. Eine Auswahl des Tools kann von verschiedenen Faktoren abhängig sein, beispielsweise Art und Menge des Ausgangsmaterials, verfügbare Hardware, verfügbare Softwarekenntnisse sowie die gewünschte Qualität. Auch aufgrund der vielen verschiedenen Anforderungen werden im OCR-D-Projekt[1] unterschiedliche Lösungen entwickelt. „Dazu wurde ein Koordinationsprojekt gebildet, das in der ersten Projektphase Entwicklungsbedarfe identifizierte. Diese wurden in der zweiten Projektphase von insgesamt acht Modulprojekten bearbeitet. In der derzeitigen dritten Projektphase steht die konzeptionelle Vorbereitung für die automatische Generierung von Volltexten für die Verzeichnisse der im deutschen Sprachraum erschienenen Drucke des 16., 17. und 18. Jahrhunderts im Fokus. Außerdem arbeiten vier Implementierungsprojekte daran, OCR-D in bestehende Anwendungen und Infrastrukturen zu integrieren, während drei Modulprojekte OCR-D-Werkzeuge weiter optimieren.“ (https://ocr-d.de/de/about)

OCR4all und eScriptorium

OCR4all[2] ist eine benutzerfreundliche Anwendung für die OCR mit einer grafischen Benutzeroberfläche, die unter anderem ein Training der Layout- und Texterkennung niedrigschwellig ermöglicht. In einem der OCR-D-Implementierungsprojekte wird das „Open-Source-Werkzeug OCR4all so erweitert und angepasst werden, dass Bibliotheken und Archive bei ihrer Massendigitalisierung die im Rahmen des OCR-D-Projekts erarbeiteten Lösungen niederschwellig, flexibel und eigenständig einsetzen können. Eine zusätzliche visuelle Erklärungskomponente soll darüber hinaus Unterstützung bei der Erstellung und Konfiguration optimaler OCR-Workflows bieten.

Als Use Case fungiert die Forschungsbibliothek des GEI Braunschweig mit ihren digitalisierten Schulbüchern des 17. und 18. Jahrhunderts. Um zunehmende Komplexitäten der so entstehenden OCR-Lösung nutzerorientiert aufzufangen, wird die bestehende grafische Benutzerschnittstelle in enger Kooperation und unter Anleitung des HCI Lehrstuhl der Universität Würzburg angepasst und weiterentwickelt.“ (ebenda)

Auch eScriptorium,[3] entwickelt von der Université PSL, ist als Open-Source-Lösung geeignet für DH-Anwender*innen, um manuell und automatisch Transkriptionen zu erstellen. Während der Fokus bei OCR4all auf einer großen Anzahl an verschiedenen OCR-Tools liegt, die hier in einer Anwendung integriert sind, hat eScriptorium Vorteile in den Bereichen des Datenaustauschs und der Ergonomie. eScriptorium und OCR4all haben beide grafische Benutzeroberflächen, die es Forschenden ermöglichen, Volltexte zu erstellen und dabei einen vergleichsweise niedrigschwelligen Einstieg zu haben. Die Tools zu kennen und auszuprobieren, versetzt Forschende in die Lage, für ihre eigenen Projekte geeignete Methoden auszuwählen. eScriptorium und OCR4all sind eine Auswahl unter vielen verschiedenen OCR-Tools, die verfügbar sind. Forschende können nach ihrem Einstieg in die OCR ggf. Weitere, spezialisertere Tools ausprobieren, in vielen Fällen werden sie allerdings mit eScriptorium und/oder OCR4all ihre Bedarfe für die Forschung decken können und haben damit eine gute Grundlage für weitere Arbeiten. Daher können diese beiden Anwendungen im Workshop ausprobiert werden. Dazu können gern eigene Daten mitgebracht werden.

Individuelle OCR-Workflows für Forschende und bestandshaltende Einrichtungen

Je nach Tool ist es ggf. notwendig, geeignete Workflows und Modelle auszuwählen. Schritte einer OCR (Optical Character Recognition) oder einer HTR (Handwritten Text Recognition) können unter anderem umfassen:

- Binarisierung: Das Umwandeln von Bildern in Farbe oder Graustufen in Schwarz-Weiß-Bilder
- Cropping: Ausschneiden des Bildes auf den Bereich, der erkannt werden soll (ohne gegenüberliegende Seiten oder störende Elemente)
- Denoising: Entfernen von störenden Artefakten
- Deskewing: Geraderücken eines schiefen Scans
- Dewarping: Entfernen von Verzerrungen
- Layoutanalyse auf Regionen-, Zeilen-, Wort- und Glyphenebene.
- Texterkennung
- Nachkorrektur

Dabei können diese Schritte auf verschiedene Arten kombiniert werden und einzelne Schritte mehrmals ausgeführt werden. Für jeden Prozessierungsschritt gibt es bei OCR-D bzw. OCR4all in der Regel mehr als einen Prozessor, der zur Verfügung steht. Zudem arbeiten manche Prozessoren modellbasiert, sodass ein solches Modell ausgewählt (und ggf. weiter trainiert) werden muss. Weitere mögliche Parameter bei der Erstellung eines Workflows kommen ergänzend hinzu. Um hier die richtige Wahl für Workflows und Modelle bei OCR4all, eScriptorium oder anderen Tools zu treffen, sind grundlegende Kenntnisse der OCR und Deep Learning notwendig, die im Workshop vermittelt werden.

Nutzungsszenarien und Ressourcen von OCR-D

Innerhalb des Projekts OCR-D mit dem Fokus auf Masssenvolltextdigitalisierung werden Ressourcen bereitgestellt, die auch im Rahmen individueller OCR-Arbeiten von Forschenden nützlich sein können.

Aktuell können in OCR-D die Workflows über die Kommandozeile ausgeführt werden. Mit der neuen Version von OCR4all, das die OCR-D-Prozessoren integriert hat, ist es außerdem möglich, die Konfigurationen über eine grafische Benutzeroberfläche vorzunehmen. Dies macht individuelle Workflows zugänglicher. OCR4all als freie und auf Nutzerfreundlichkeit konzentrierte Software bietet alle notwendigen Workflowschritte an und integriert das Tool LAREX (Layout Analysis and Region Extraction) für Layouterkennung und Korrektur von Zwischenergebnissen der Layout- sowie Texterkennung.

Für bestimmte Materialien kann ein einfacher Workflow mit wenigen Schritten ausreichend sein, um eine gute Qualität zu erzeugen, während Materialien mit komplexen Layouts aufwändigere Workflows notwendig machen. Für eine Vorauswahl stellen wir Standard-Workflows sowie das im OCR-D-Projekt entwickelte Benchmarking-Tool QUIVER bereit.[4] Mit dieser Entwicklung erstellt OCR-D Werte für Durchsatz und Qualität bestimmter Workflows auf verschiedenen Materialien.

Werden Modelle (nach-)trainiert und dafür Ground Truth erstellt, erreicht das Vorhaben eine weitere Komplexitätsstufe. Bei der Transkription von Ground Truth helfen die in OCR-D entwickelten und gepflegten Ground-Truth-Guidelines.[5] „Mit den OCR-D-Ground-Truth-Guidelines wurden Richtlinien geschaffen, die eine Format-Dokumentation des vorhandenen OCR-D-Ground-Truth darstellt und als Handlungsanweisung für die Ground-Truth-Erstellung genutzt werden kann. Mit dieser Normierung kann der Ground-Truth technisch validiert werden. Darüber hinaus können vorhandene Transkriptionen auf Grundlage dieses Regelwerkes überprüft und gegebenenfalls in Ground-Truth-Daten umgewandelt werden. Das Datenformat des OCR-D-Ground-Truth ist PAGE-XML. Dieses Format wurde initial durch das PRImA Research Lab an der Universität Salford Greater Manchester entwickelt und innerhalb des EU-Projektes IMPACT grundlegend erweitert. Zurzeit wird es vom PRImA Research Lab betreut. Um eine Weiterentwicklung und Pflege dieses Formates zu gewährleisten, wurde auf Initiative von OCR-D ein PAGE-XML-Board geschaffen.“ (https://ocr-d.de/de/gt-guidelines/trans/) Zusätzlich finden regelmäßige Onlinemeetings (GT-Call)[6] statt, um Fragen zu erörtern.

Workshop

Im ganztägigen Workshop erlangen die Teilnehmenden erforderliche Kenntnisse, um Tools und Workflows für die Volltexterschließung unter der Vielzahl von Angeboten auszuwählen. Dabei legen wir einen besonderen Fokus auf die genannten Open-Source-Produkte wie die OCR-D, OCR4all und eScriptorium.

Geplante Inhalte des Workshops sind:

- Einführung in Deep Learning, OCR, Layoutanalyse, Evaluation, OCR-D und passende Workflows für bestimmte Vorlagen
- Praktisches Arbeiten mit OCR4all und LAREX
- Praktisches Arbeiten mit eScriptorium
- Möglichkeiten zum Hochskalieren mit OCR-D Processing Server/Workern
- Unterstützung beim individuellen Einrichten der Werkzeuge

Bei der Anwendung von OCR4all mit LAREX und eScriptorium ist der Workshop interaktiv gestaltet und die Teilnehmenden können die Tools selbst ausprobieren. Sie erhalten grundlegende Kenntnisse für die Anwendung. Dabei kann anhand selbst mitgebrachter Vorlagen bereits fest-

gestellt werden, wo Potenziale, aber auch Grenzen der OCR liegen können.

Nach Abschluss des Workshops kennen die Teilnehmenden frei verfügbare OCR-Tools und Anlaufstellen für Ressourcen, die sie bei der Einrichtung eines OCR-Workflows benötigen.

Vortragende

Konstantin Baierer (SBB Berlin), Lena Hinrichsen (HAB Wolfenbüttel) und Matthias Boenig (BBAW) arbeiten im OCR-D-Koordinierungsprojekt.

Christian Reul leitet die Digitalisierungseinheit des Zentrums für Philologie und Digitalität „Kallimachos“ (ZPD) and er Universität Würzburg. In OCR-D ist er unter anderem mitverantwortlich für ein Implementierungspaket für OCR4all.

Lilja Sautter ist an der Niedersächsischen Staats- und Universitätsbibliothek Göttingen in der Einheit Software- und Service-Entwicklung tätig. In OCR-D ist sie beteiligt an den Projekten OPERANDI sowie OLA-HD.

Mehmed Mustafa ist Entwickler in der Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen. In OCR-D arbeitet er im Koordinierungsprojekt sowie den Projekten OPERANDI und OLA-HD.

Larissa Will ist Referentin für Forschungsdatenmanagement und Digitalisierung (Digital Humanities) an der Universität Mannheim.

Zielpublikum, technische Ausstattung

Teilnehmende sollten ein spezifisches Interesse daran haben, selbst Volltexte zu erstellen. Dies können Forschende sein, die Volltexte für ihre Projekte benötigen, sowie Mitarbeitende von Einrichtungen, die Volltexte und/oder OCR-Services in der Zukunft anbieten möchten. Um die Tools auszuprobieren, wird ein Laptop benötigt, möglichst mit Windows oder Linux.

Fußnoten

1. https://ocr-d.de/
2. https://www.ocr4all.org/
3. https://gitlab.com/scripta/escriptorium
4. https://ocr-d.de/quiver-frontend/#/workflows?view=list
5. https://ocr-d.de/en/gt-guidelines/trans/
6. https://ocr-d.de/en/community

Bibliographie

eScriptorium. https://gitlab.com/scripta/escriptorium [19.07.2023].

Website OCR-D. https://ocr-d.de [19.07.23].

Website OCR.4all. https://www.ocr4all.org/ [19.07.2023].