Absichtserklärung Implementierungsprojekt OCR-D

MARBACH WEIMAR WOLFENBÜTTEL FORSCHUNGS VERBUND

Szenario: *Entwicklung eines generischen Implementierungspaketes von OCR-D als Webdienst*

Antragsteller:

- Prof. Dr. Peter Burschel, Herzog August Bibliothek Wolfenbüttel
- Dr. Jan Horstmann, Klassik Stiftung Weimar
- Prof. Dr. Roland S. Kamzelak, Deutsches Literaturarchiv Marbach

Kooperationspartner: Forschungsverbund MWW; Klassik Stiftung Weimar (Herzogin Anna Amalia Bibliothek, Goethe- und Schiller-Archiv); Herzog August Bibliothek Wolfenbüttel; Deutsches Literaturarchiv Marbach

Kurzdarstellung Projektziele und Anwendungsszenario

Im Zuge der Entwicklungen des Forschungsverbundes Marbach Weimar Wolfenbüttel (MWW) und des dort angesiedelten Digitalen Labors soll der entwickelte OCR-D-Prototyp als Webdienst implementiert werden und die Angebote des Virtuellen Forschungsraums (VFR) ergänzen.

Das Digitale Labor des Forschungsverbunds MWW generiert konkrete Bedarfe aus sechs an den drei Standorten des Verbunds angesiedelten wissenschaftlichen Forschungsprojekten, deren Untersuchungsgegenstände vom 17. bis in das 20. Jahrhundert reichen. Die Forschungsprojekte werden durch Methoden der Digital Humanities unterstützt und angereichert; eine Voraussetzung hierfür sind in vielen Fällen Volltextdigitalisate. Die Projekte sollen die Ebene der Metadatenforschung überschreiten und auch Methoden des Distant Reading einschließen. Die Arbeit des Digitalen Labors findet in verschiedenen DH-Teilprojekten statt (jeweils zwei pro Standort), die in ihrer Gesamtheit den virtuellen Kern von MWW bilden. Digitales Herz des Verbunds ist der Virtuelle Forschungsraum (VFR; https://vfr.mww-forschung.de), in dem nicht nur sämtliche Fallstudien und Forschungsgruppen von MWW, sondern auch Ausstellungen und Veranstaltungen inhaltlich bearbeitet und öffentlich präsentiert werden. Außerdem bietet dieser virtuelle Raum eine sammlungsübergreifende Suche und wird zukünftig ein System zur sammlungserschließenden Forschung sowie eine digitale Publikationsumgebung beinhalten.

Der VFR ist auf Öffnung und Anschlussfähigkeit angelegt und bietet eine Plattform für überregional ausgeführte kollaborative Projekte der sammlungsbezogenen Forschung. Um wissenschaftlich nutzbare Volltextdigitalisate aus den eigenen Sammlungen der drei Standorte

anbieten zu können, soll im VFR ein wissenschaftlich verlässlicher und niedrigschwelliger wie nutzerfreundlicher OCR-Service angeboten werden. Die Idee ist, dass sich der Service primär zur Massenbearbeitung der bereits eingescannten Sammlungen der drei Verbundinstitutionen eignet, jedoch so generisch gestaltet ist, dass er auch (a) von weiteren sammlungsverwaltenden Einrichtungen und (b) von sammlungserschließenden Forschungsprojekten, die anders als Bibliotheken und Archive in der Regel keine METS-Dateien vorliegen haben, genutzt werden kann. Der Service sollte daher sowohl von Experten als auch niedrigschwelliger (etwa über grafische Nutzerinterfaces) genutzt werden können. In diesem Zusammenhang wären Ziele etwa die automatische Verknüpfung von maschinell erkannten Schriftarten mit dem entsprechenden OCR-Modell und die Angabe von Konfidenzwerten (die in der Massendigitalisierung niedriger sein können als in Forschungsprojekten, die eine Nachkorrektur zum Training eines verbesserten Modells leisten können). In den beteiligten Institutionen liegen derzeit in großer Anzahl digitalisierte (aber noch nicht maschinenlesbare) Drucke vor, die in einem Implementierungsprojekt im Sinne des Verbundcharakters von MWW virtuell zusammengeführt werden sollen und jeweils unterschiedliche Herausforderungen an den generischen Webservice stellen.

Kurzdarstellung Vorarbeiten

Alle drei Standorte des Forschungsverbunds haben in den vergangenen Jahren bereits den DFG-Richtlinien entsprechende Druckdigitalisate erstellt, die sich für eine OCR anbieten (etwa die Erstellung von TIFs mit mindestens 300 dpi, häufig auch erheblich höhere Auflösung). Die Quellen des Verbundes decken über 500 Jahre deutscher Literatur- und Geistesgeschichte ab. Diese Diversität der jeweils lokalen Bestände sehen wir als Herausforderung aber zugleich als große Chance, einen tatsächlich *generischen* Webdienst entwickeln zu können.

Die Herzog August Bibliothek Wolfenbüttel ist eine der drei Trägerbibliotheken des VD17-Projekts und hat im Bereich Buchdigitalisierung den längsten Vorlauf. Hier ist daher eine Vielzahl an digitalisierten Drucken vorrätig: Insgesamt handelt es sich um 36.029 Titel mit variierendem Umfang (Stand 31.12.2019), von denen bislang lediglich ein Bruchteil mit OCR-Werkzeugen bearbeitet wurde.

Die Klassik Stiftung Weimar bietet sowohl im bibliothekarischen wie im archivalischen Bestand digitalisierte Drucke. Die Herzogin Anna Amalia Bibliothek war am VD17-Projekt beteiligt und engagiert sich unter anderem im VD18-Projekt (1. Phase mit 3.000 Werken/390.000 Seiten, potentielle Zuteilung von insgesamt 7.800 Werken; aktueller Stand: 300 Werke/20.000 Seiten digitalisiert). Außerdem liegen gefragte Sammlungen wie Goethes Privatbibliothek teilweise digitalisiert vor: Von den insgesamt rund 9.000 Werken sind 2.140 Werke (ca. 300.000 Seiten) für die Digitalisierung vorgesehen (aktueller Stand: 1.100 Werke/125.000 Seiten). In beiden Projekten werden die digitalisierten Bände bereits gemäß der vorkommenden Schrifttypen in Antiqua, Fraktur und Grafik codiert. Das Goethe- und Schiller-Archiv Weimar verfügt gegenwärtig über insg. 15.341 Blatt digitalisierte Drucke, die sich ergänzend und explorativ für Texterkennungsverfahren eignen: darunter Erstdrucke kanonischer Texte, Korrekturbögen, Kalender, Schriftzeichentabellen etc.

Das Deutsche Literaturarchiv Marbach hält Digitalisate im fünfstelligen Bereich vor im Zusammenhang der digitalisierten Zeitschrift *Simplicissimus*, 70 digitalisierte Flugblätter aus dem Zweiten Weltkrieg, oder auch Kracauers sog. "Klebemappen" (ca. 1500 Seiten) als Teil des DFG-Projektes zur Bibliothek von Siegfried Kracauer. Dabei handelt es sich um Kracauers eigene Sammlung seiner journalistischen Beiträge in der Frankfurter Zeitung von 1921 bis 1933. Zudem wurde in der ersten Förderphase des Forschungsverbunds MWW in einer Kooperation zwischen DLA Marbach und GSA Weimar die vollständige Rilke-Korrespondenz im Insel Verlagsarchiv digitalisiert.

Für die Umsetzung der Pilotierungsphase stehen wir für konzeptionelle Absprachen in engem Austausch mit Elisabeth Engl aus dem derzeitigen OCR-D-Koordinierungsprojekt. Außerdem haben wir eine Arbeitsgruppe gegründet, die sich um Christiane Müller, der Verantwortlichen für den Virtuellen Forschungsraum und Mitarbeiterin im Forschungsverbund MWW, gruppiert. Die drei Einrichtungen bringen in vergleichbarem Umfang Personalmittel als Eigenanteil in diese Arbeitsgruppe ein. Die Arbeitsgruppe wird im Zuge der Pilotierungsphase die vorhandenen OCR-D-Pakete hinsichtlich ihrer Implementierbarkeit in einen Webservice evaluieren, mit unterschiedlichen Materialien der drei beteiligten Verbundeinrichtungen die einzelnen OCR-D-Werkzeuge testen und ein Konzept zur Implementierung der geeigneten, einen vollständigen OCR-Workflow abdeckenden OCR-Werkzeuge als Webservice erarbeiten.

Konkret sollen verschiedene Ansätze zur Übergabe der digitalisierten Werke an einen OCR-Server getestet werden. Neben den üblichen Schnittstellen, die auf eine direkte Verbindung zwischen dem Imageserver einerseits und dem OCR-Server andererseits aufbauen, sollen im Rahmen des Implementierungsprojekts auch Möglichkeiten getestet werden, einen OCR-Server basierend auf dem IIIF-Manifest eines Werkes oder anderen XML-Formaten anzusteuern. Im Gegensatz zu einer rein serverbasierten Verbindung wäre dieser Ansatz deutlich barriereärmer. IIIF-Manifeste werden mittlerweile von einer Vielzahl an Einrichtungen zur Verfügung gestellt. Dieses Verfahren würde sich gleichermaßen für die Massenbehandlung eines größeren Sets an Werken eignen, wäre aber ebenso für einzelne Werke nutzbar. Gegenüber der etablierten Schnittstelle läge der Vorteil auf Endnutzerseite darin, dass eine zeitaufwändige Kommunikation mit den bestandshaltenden Einrichtungen im Wesentlichen entfällt, da die IIIF-Manifeste in den verschiedenen digitalen Sammlungen in der Regel frei zugänglich sind.

Sicherzustellen ist dabei, dass Standardprodukte wie z.B. Kitodo bzw. Goobi mit ihren Schnittstellen weiterhin auch direkt an den Webservice andocken können. Geklärt werden sollen außerdem Fragen zur technischen Umsetzung einer zuverlässigen (Langzeit-)Speicherung der OCR-Daten. Ebenso notwendig ist aber auch eine Klärung der datenschutzrechtlichen Fragen.

Jan Horstmann
(auch im Namen von Peter Burschel und Roland Kamzelak)