Absichtserklärung der Universität Hamburg

DFG –Projektantrag im Bereich

„Wissenschaftliche Literaturversorgungs- und Informationssysteme“ (LIS)

Antragssteller:

Dr. Stefan Thiemann
Universität Hamburg, Zentrum für nachhaltiges Forschungsdatenmanagement
Monetastraße 4, 20146 Hamburg
stefan.thiemann@uni-hamburg.de
Tel.: 040 428 38 38 44

Dr. Sylvia Melzer
Universität Hamburg, Centre for the Study of Manuscript Cultures
Warburgstraße 26, 20354 Hamburg
sylvia.melzer@uni-hamburg.de

Titel:	Implementierung der OCR-D-Software zur Volltextdigitalisierung historischer Drucke
Teilvorhaben:	**Automatisierte Auswertung von Digitalisaten historischer Dokumente unter Nutzung der OCR-D-Software**
Geplante Laufzeit:	01.01.2021 – 31.12.2022 (24 Monate)

1 Kurzdarstellung der Projektziele

Nach über 90 Jahren ist *Optical Character Recognition* (OCR) immer noch ein sehr aktuelles Forschungsgebiet im Bereich der Mustererkennung. Die Anwendungsgebiete dieser Technologie sind weitreichend. Dazu gehört unter anderem die Unterstützung der Blindenlese, Interpretation von Morse-Codes, Automatische Erfassung von Postadressen, Auswerten von E-Mails, Einscannen von Preisschildern und Pässen sowie die Textgewinnung von digitalisierten Dokumenten.[1] Die Zahl von Digitalisaten nimmt stetig zu und es entstanden bereits verschiedene Repositorien bzw. Portale mit Suchstrategien.[2] Ein Abgleich von Suchbegriffen mit den vorhandenen Digitalisaten kann jedoch erst erfolgen, wenn vom Digitalisat beispielsweise Text gewonnen wurde, um einen Abgleich mit den gewünschten Suchbegriffen vornehmen zu können.

Zu den klassischen OCR-Verfahren der Texterkennung gehören die Open Source OCR *Tesseract* und *OCRopus*. Eine Weiterentwicklung der OCR-Software erfolgte im Rahmen des DFG-Gesamtvorhabens „Skalierbare Verfahren der Text- und Strukturerkennung für die Volltextdigitalisierung historischer Drucke“ in acht geförderten Teilvorhaben. Im Rahmen dieser Teilvorhaben sind OCR-D-Software-Module z. B. zur Bildvorverarbeitung, Layouterkennung, Textoptimierung, Modelltraining und zur Texterkennung entwickelt worden. Während die Texterkennung dieser OCR-Verfahren von Digitalisaten aus der heutigen Zeit gut funktioniert, besteht noch Verbesserungspotential hinsichtlich der Text- und Worterkennung von historischen Drucken oder auch Dokumenten sowie der Durchsuchbarkeit von generierten Datensätzen. Weiterentwicklungen bestehender OCR-Softwaresysteme liefern zwar bessere Ergebnisse als die Vorgängerversionen[3], doch könnten durch die Integration von Korrekturfunktionen vor, während oder nach der Auswertung von Digitalisaten qualitativ höherwertigere Ergebnisse erzielt werden. Zudem ergeben sich aus den verschiedensten Anwendungsfeldern andere Anforderungen an ein OCR-Softwaresystem, in denen spezifische

[1] https://de.qwe.wiki/wiki/Timeline_of_optical_character_recognition, abgerufen am 21.04.2020
[2] https://staatsbibliothek-berlin.de/fileadmin/user_upload/zentrale_Seiten/katalogsystem_wd/dokumente/e-day/eday_15_HO_Histor.Drucke.pdf, abgerufen am 21.04.2020
[3] http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html, abgerufen am 21.04.2020

Implementierungspakete für die Bild-, Struktur-, Sprach- und Texterkennung zum Einsatz kommen können. Daher ergibt sich der Bedarf eine Bild-, Struktur-, Sprach- und Texterkennung in einem Workflow so zu implementieren, dass eine Nachbesserung in Form einer Interaktion mit den nutzenden Wissenschaftlerinnen und Wissenschaftlern als Feedback erfolgen kann. So soll bei der Volltextdigitalisierung historischer Drucke eine möglichst hohe Genauigkeit erzielt werden, ohne die OCR-D-Module selbst zu verbessern bzw. weiterzuentwickeln. Die Herausforderungen, die es dabei zu lösen gilt, ist es dabei einen Workflow so zu implementieren, dass

- eine intelligente Kombination der einzelnen OCR-D-Implementierungspakete für eine Analyse von Digitalisaten als generischer Ansatz angeboten wird,
- die OCR-D-Software ohne Expertenwissen benutzbar ist,
- die Interaktion mit der Wissenschaftlerin und dem Wissenschaftler als Feedback aktiv angeboten wird, wenn die vollautomatische OCR-Nachkorrektur nicht zu guten Ergebnissen führt,
- interessierte Einrichtungen die OCR-D-Software selbständig betreiben können und
- eigene Workflows mit der OCR-D-Software zusammengestellt werden können.

Diese Herausforderungen sollen in diesem Vorhaben gelöst werden, indem

- in dem Bereich der Manuskriptforschung verschiedene Anwendungsszenarien für die Analyse von Manuskripten erarbeitet werden, für die der zu entwickelnde generische Ansatz angewendet werden kann.
- für die verschiedenen Anwendungsszenarien Workflows, die keine Kenntnisse in der Programmierung der OCR-D-Module zur Benutzbarkeit voraussetzen, entwickelt werden. Dabei gilt es auch, dass nationale und internationale Entwicklungen zu Methoden, Verfahren und Werkzeugen zur Volltextdigitalisierung historischer Drucke berücksichtigt werden.
- die Integration von benutzerfreundlichen Benutzeroberflächen und Handlungsanweisungen in den Workflows unterstützt wird, sodass die Anwendbarkeit der OCR-D-Module ohne vertiefte Kenntnisse der OCR-D-Module erfolgen kann.
- anerkannte Technologien bei der Implementierung der Workflows, Benutzeroberflächen und Handlungsanweisungen berücksichtigt werden.
- eine technische Anschlussfähigkeit der Workflowstruktur gewährleistet wird, damit eigene Workflows erstellt werden können.

Das Vorhaben hat zur Lösung der o.g. Herausforderungen zum Ziel die bereits betriebsfähigen e-Research-Technologien so mit einer zu entwickelten benutzerorientierten Funktion zusammenzuführen, um die Usability und die längerfristige Etablierung der Technologien für mehrere Anwendungsbereiche zu unterstützen. Im Rahmen des Vorhabens würde insgesamt der Schwerpunkt auf die Qualität der Ergebnisse OCR-D-Software gelegt und mit den verschiedenen geplanten Anwendungsfeldern gezeigt werden kann, dass eine technische Anschlussfähigkeit der Infrastrukturen gegeben ist und die Einordnung der Workflows in einer nachvollziehbaren Prozesskette vorgenommen werden kann.

2 Anwendungsszenario und Vorarbeiten

Im Bereich der Manuskriptforschung an der Universität Hamburg am *Centre for the Study of Manuscript Cultures* (CSMC) bieten sich eine Reihe von Anwendungsgebieten, bei denen die OCR-D-Module in verschiedensten Workflows Anwendung finden können und so die Auswertung von Manuskripten und Dokumenten und damit die Forschungsarbeit unterstützen können. Eine automatisierte Zeichen- und Worterkennung der Digitalisate von historischen Manuskripten mittels der entwickelten OCR-D-Module würde die Forschungsarbeit bei der Auswertung der Manuskripte

unterstützen. Die Anwendung würde nicht nur Zeit einsparen, sondern es wäre denkbar, eigene Softwaremodule mit den OCR-D-Modulen zu einer neuen Anwendung zu koppeln.

Zunächst wurden die dokumentierten Tests zur Prüfung der Funktionalität eines vordefinierten OCR-D-Workflows erfolgreich durchgeführt. Ein weiterer Test-Workflow mit ausgewählten OCR-D-Modulen wurde für einen anderen Datensatz (*estor_rechtsgelehrsamkeit02_1758.ocrd.zip*) aus dem Repository https://ocr-d-repo.scc.kit.edu/api/v1/metastore/mets/classification?class=Fachtext erfolgreich ausgeführt. Für diese Digitalisate konnte ebenfalls die Etablierung eines Workflows erfolgreich durchgeführt werden. Die Ergebnisse zeigen, dass es Herausforderungen gibt Text von tabellarischen Seiten zu gewinnen, welches das folgende Anwendungsbeispiel exemplarisch zeigt.

Ein repräsentatives Anwendungsbeispiel für die Textgewinnung aus Tabellen sind die digitalisierten Journals (siehe [4]) aus dem naturhistorischen Museum Godeffroy[5]. Das Museum Godeffroy existierte von 1861 bis 1885 in Hamburg. Die in den Jahren von 1873 bis 1910 erschienenen Journals beinhalten Publikationen zu geographischen, ethnographischen und naturwissenschaftlichen Mitteilungen.[6] Die Artikel beinhalten neben Text, ebenso Tabellen und Abbildungen wie Zeichnungen und Fotographien. Im Rahmen des Vorhabens soll zur Volltextdigitalisierung des „*Journal des Museum Godeffroy*“ die auch Textgewinnung von Tabellen, die in den Journals sehr häufig vorkommen, getestet werden. Es wurden bereits für das Journal des Museum Godeffroy auf der Webseite https://www.biodiversitylibrary.org/item/244246#page/11/mode/1up erste Auswertungen von Tabelleninhalten mit der Tesseract-Software und anderen OCR-Software vorgenommen. Die Ergebnisse zeigen, dass mit der Tesseract-Software keine Tabelle automatisch ausgegeben wird und man daher in diesem Beispiel zurzeit besser beraten ist, die 570 Seiten von Hand abzutippen. In der Pilotierung sollen verschiedene Workflows zur Textgewinnung der Journals ausgeführt werden, um Handlungsempfehlungen über den Umgang mit Tabellen abgeben zu können.

Ein weiteres Anwendungsgebiet der OCR-D-Software ergibt sich unter anderem aus dem Projekt NETamil “Going from Hand to Hand – Networks of Intellectual Exchange in the Tamil Learned Traditions”[7] von März 2014 bis August 2019 aus dem siebten Rahmenprogramm der Europäischen Union für Forschung, technologische Entwicklung und Demonstration. Im Rahmen des Projektes wurden klassische tamilische Manuskripte auf Palmblatt und -papier digitalisiert. Eine automatische Auswertung von Worten in einer Linie würde bei der Analyse von Palmblatt-Manuskripten unterstützen.

In Rahmen des CSMC stehen weitere Projekte zur Manuskriptforschung zur Verfügung, die von den geplanten Arbeiten profitieren würden. Es sind zwar keine große Mengen an Digitalisaten vorhanden, aber die reichhaltigen Anwendungsbeispiele aus den verschiedensten Projekten können zeigen, wie in verschiedensten Workflows die OCR-D-Module angewendet werden können und damit eine größere Menge an Nutzende der OCR-D-Software erreichen und insbesondere die Breite der Anwendungsfälle erheblich vergrößern.

Zudem können die Workflows als Werkzeuge als zentrales Angebot in das Forschungsdatenrepositorium der Universität Hamburg (https://www.fdr.uni-hamburg.de) integriert werden, um dort langzeitgesicherte eingescannte Dokumente für Nutzende mit einem einfach zu nutzenden Workflow in nachnutzbaren Text zu konvertieren.

---

[4] https://www.biodiversitylibrary.org/bibliography/12203#/summary; Abgerufen am 16.04.2020
[5] https://www.wikiwand.com/de/Museum_Godeffroy; Abgerufen am 16.04.2020
[6] https://www.wikiwand.com/de/Journal_des_Museum_Godeffroy; Abgerufen am 16.04.2020
[7] https://cordis.europa.eu/project/id/339470, abgerufen am 19.05.2020