Volltexterkennung für historische Sammlungen mit OCR4all-libraries iterativ und partizipativ gestalten

Klaes, Jan Sebastian
klaes@leibniz-gei.de
Georg Eckert Institute for International Textbook Research. Member of the Leibniz Association

Korwisi, Kristof
kristof.korwisi@uni-wuerzburg.de
University of Würzburg Human-Computer Interaction, Germany

Krüger, Katharina
katharina.krueger@gei.de
Georg Eckert Institute for International Textbook Research. Member of the Leibniz Association

Reul, Christian
christian.reul@uni-wuerzburg.de
University of Würzburg Centre for Philology and Digitality, Germany

Towara, Nadine
towara@leibniz-gei.de
Georg Eckert Institute for International Textbook Research. Member of the Leibniz Association

Mit der Initiierung und Durchführung von Massendigitalisierungsprojekten haben Bibliotheken eine wesentliche Grundlage für den Zugang und die Nutzung digitaler Quellen geschaffen. Vor dem Hintergrund der Weiterentwicklung digitaler Forschungsmethoden u.a. im Bereich der Digital Humanities stellt sich nun zunehmend die Frage nach der Qualität der Volltexterkennung (OCR). Die Volltextqualität in den digitalen Sammlungen ist dabei nicht nur abhängig von Materialbesonderheiten, sondern auch von dem eingesetzten Texterkennungssystemen und deren Weiterentwicklungen.

Die Zusammenarbeit zwischen dem Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung (GEI), dem Zentrum für Philologie und Digitalität “Kallimachos” (ZPD) und dem Lehrstuhl für Mensch-Computer-Interaktion (HCI) der Universität Würzburg zielt darauf ab, das Web-GUI-basierte Open-Source-Werkzeug OCR4all (Reul et. al 2017; 2019) so zu erweitern und anzupassen, dass Bibliotheken und Archive bei ihrer Massendigitalisierung die im Rahmen des von der Deutschen Forschungsgemeinschaft (DFG) geförderten OCR-D-Verbundprojekts (Engl, 2020) erarbeiteten Lösungen niederschwellig, flexibel und eigenständig einsetzen können. Als Use Case fungiert die Forschungsbibliothek des GEI mit ihrer digitalisierten Schulbuchbibliothek GEI-Digital.

Die digitale Schulbuchbibliothek umfasst historische deutsche Schulbücher der Fächer Geschichte, Geographie und Politik, Religion/Werteerziehung sowie Realien- und (Erst-)Lesebücher von den Anfängen der Schulbuchproduktion im 17. Jahrhundert bis zum Ende des Ersten Weltkriegs (Hertling/Klaes 2018a ; 2018b) . Mit GEI-Digital ist für die Digital Humanities ein einzigartiges Korpus mit über 6.100 digitalisierten Schulbüchern entstanden, dass die gesamte Epoche der deutschen Schulbücher von deren Entstehung bis 1918 mit hoher Vollständigkeit virtuell zusammenführt. Die Digitalisate und Daten werden in zahlreichen Digital-Humanities-Projekten bereits nachgenutzt, wie z.B. im Projekt „Welt der Kinder“, in dem das Korpus mit Topic Modeling-Verfahren untersucht wurde (Nieländer / Weiß 2018).

Die besonderen Bedarfe nach hochwertiger Texterkennung unter der Gruppe der Forscher*innen und der digitalen Schulbuchbibliothek wurden 2014 in Form einer Befragung ermittelt. Darauf aufbauend erfolgte eine Studie zur Qualität der Texterkennung der seit 2009 eingesetzten kommerziellen und Open-Source-Softwarelösungen für Texterkennungsverfahren bei der Massendigitalisierung. Die Ergebnisse der Studie zeigten auf, das Massenverfahren nicht die Bedarfe der Forscher*innen decken. Der digitale Bestand weist erhebliche Unterschiede in der OCR-Qualität auf, auch weil ein komplexes Layout und uneinheitliche Typographien noch immer große Hürden für eine hochwertige Volltexterkennung darstellen. Um die OCR-Qualität gezielt zu verbessern, soll ausgehend vom konkreten Use Case der Forschungsbibliothek des GEI ein möglichst generisch anwendbares Verfahren implementiert werden, das eine nach Sammlungen mit jeweils ähnlicher Materialgrundlage organisierte Volltexterkennung erlaubt.

Um zunehmende Komplexitäten der so entstehenden OCR-Lösung nutzerorientiert aufzufangen, wird die bestehende grafische Benutzerschnittstelle in enger Kooperation und unter Anleitung der HCI angepasst und weiterentwickelt. Eine zusätzliche visuelle Erklärungskomponente soll darüber hinaus Unterstützung bei der Erstellung und Konfiguration optimaler OCR-Workflows bieten. Alle im Projekt erarbeiteten Lösungen werden schritthaltend mittels umfassender Nutzerstudien evaluiert. Um sicher zu stellen, dass auch nicht-technische Anwender*innen in Bibliotheken und Archiven komfortabel und selbstständig auf OCR-D-Lösungen zugreifen können, fließen die Evaluationsergebnisse stetig in die Weiterentwicklung ein. Dafür werden im Rahmen des Projekts Workshops mit interessierten Anwendern stattfinden, um deren Bedarfe in die Weiterentwicklung der Benutzerführung einfließen zu lassen. Ein besonderer Fokus wird dabei auf die intuitive und selbsterklärende Bedienbarkeit durch ein breites Nutzerspektrum gelegt.

Forscher*innen als wichtige Zielgruppe von Forschungs- und Spezialbibliotheken profitieren mit der Weiterentwicklung von OCR4all von der Möglichkeit OCR-Prozesse *iterativ* und *partizipativ* mitgestalten zu können. Die im Rahmen des OCR-D-Verbundprojekts entwickelten Komponenten können von Forscher*innen dynamisch eingesetzt und bei Bedarf sind Konfigurationen, Trainingsdaten und Modelle zwischen Institutionen und Individuen flexibel austauschbar, was die Optimierung von Texterkennungsprozessen verbessert. OCR-Prozesse und deren Komponenten sollen damit nachnutzbar werden. Ebenso sollen für die Forscher*innen Möglichkeiten geschaffen werden, die Qualität der Texterkennung bewerten zu können und selbst OCR-Prozesse für bestimmte Bestände steuern zu können, da deren spezifischen Bedarfe oftmals an bestimmte Forschungsfragen gebunden sind.

Bibliographie

Engl, Elisabeth. "OCR-D kompakt: Ergebnisse und Stand der Forschung in der Förderinitiative". *Bibliothek Forschung*

und Praxis, vol. 44, no. 2, 2020, pp. 218-230. https://doi.org/10.1515/bfp-2020-0024

Hertling, Anke; Klaes, Sebastian (2018): "Historische Schulbücher als digitales Korpus für die Forschung: Auswahl und Aufbau einer digitalen Schulbuchbibliothek". In: Maret Nieländer und Ernesto William De Luca (Hg.): *Digital Humanities in der internationalen Schulbuchforschung.* Göttingen: V&R unipress, S. 21–44. DOI: 10.14220/9783737009539.21

Hertling, Anke; Klaes, Sebastian (2018): "'GEI-Digital' als Grundlage für Digital-Humanities-Projekte: Erschließung und Datenaufbereitung". In: Maret Nieländer und Ernesto William De Luca (Hg.): *Digital Humanities in der internationalen Schulbuchforschung.* Göttingen: V&R unipress, 45-68. DOI: 10.14220/9783737009539.45

Nieländer, M., Weiß, A. (2018): "Schönere Daten - Nachnutzung und Aufbereitung für die Verwendung in Digital-Humanities-Projekten". In: Maret Nieländer und Ernesto William De Luca (Hg.): *Digital Humanities in der internationalen Schulbuchforschung.* Göttingen: V&R unipress, 91-116. DOI: 10.14220/9783737009539.91

Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., Puppe, F. (2019). "OCR4all - An Open-Source Tool Providing a (Semi-) Automatic OCR Workflow for Historical Printings". In: *Applied Sciences* 9 (22) 4853. https://doi.org/10.3390/app9224853

Reul, C., Springmann, U., Puppe, F. (2017). "LAREX: A Semi-automatic Open-Source Tool for Layout Analysis and Region Extraction on Early Printed Books". In: *Proceedings of the 2Nd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2017, 137–142, New York, NY, USA. ACM. https://doi.org/10.1145/3078081.307809