Forschungsdaten aus Digitalisaten

Stefan Weil und Jan Kamlah
Universitätsbibliothek Mannheim, DE

Bibliotheken tragen einen wichtigen Teil zur Digitalisierung des kulturellen Erbes bei und ermöglichen Forschenden den Zugang zu diesen Werken weltweit. Digitalisierte Dokumente werden immer öfter nicht nur als Bilder zum Lesen angeboten, sondern zusätzlich durch OCR (optical character recognition) mit dem erkannten Volltext aufgewertet. Dies erlaubt eine Suche nach Begriffen im gesamten Inhalt sowie weitere Analysemöglichkeiten. Während bei der Suche gewisse Fehlerraten toleriert und eine mäßige Layouterfassung akzeptiert werden kann, ist beispielsweise für die Extraktion von Forschungsdaten aus Digitalisaten eine hohe Zeichengenauigkeit und eine eindeutige Erfassung des Seitenaufbaus zwingend.

Gleich mehrere Projekte der Universitätsbibliothek Mannheim beschäftigen sich derzeit mit dieser Thematik. Zwei dieser Projekte werden näher vorgestellt.

Die automatische Texterkennung (OCR) mit Hilfe von Tesseract ist das Ziel des DFG-geförderten Projektes Tesseract als Komponente im OCR-D-Workflow, eines von acht Modulprojekten der DFG-Initiative OCR-D zur Verbesserung der Texterkennung historischer Drucke aus dem 16. bis 19. Jahrhundert. Hier wird die freie OCR-Software Tesseract praxistauglicher und anwendungsfreundlicher gemacht, z. B. durch Dokumentation, Korrektur von Softwarefehlern oder durch verbesserte Performance.

Schwerpunkte im DFG-Projekt Aktienführer-Datenarchiv sind Zeichengenauigkeit und Strukturerfassung zur Extraktion von Forschungsdaten. Der jährlich von 1956 bis 1999 im Buchformat erschienene Aktienführer beinhaltet Daten von börsennotierten Firmen und dient als Referenzwerk für die Forschung in den Bereichen BWL und VWL. Die zusammengestellten Firmenprofile enthalten allgemeine sowie geschäftsjahrspezifische Informationen, unterteilt in einzelne Abschnitte und Tabellen Ein Ziel des Projektes ist die automatisierte Verarbeitung der digitalisierten Daten und Speicherung in strukturierter Form in einer Datenbank. Dafür wurden gescannte Seiten vorverarbeitet, die OCR-Ergebnisse durch die Kombination mehrerer OCR-Ausgaben verbessert, eine automatisierte Strukturerfassung inklusive Tabellen entwickelt und die Daten extrahiert.

Abschließend wird noch ein Ausblick auf zukünftige Projekte mit verwandter Thematik gegeben.

Alle vorgestellte Software ist frei nachnutzbar für eigene Projekte und wird von der Universitätsbibliothek Mannheim auf GitHub unter freien Lizenzen veröffentlicht.

Stefan Weil und Jan Kamlah: Forschungsdaten aus Digitalisaten.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 189. DOI: https://doi.org/10.11588/heibooks.598.c8429