OCR-D

KOORDINIERUNGSPROJEKT ZUR WEITERENTWICKLUNG VON OCR-VERFAHREN

Das Koordinierungsprojekt *OCR-D* bereitet die Transformation der deutschsprachigen Drucke des 16.-19. Jahrhunderts in maschinenlesbaren Text vor. Dabei werden Verfahren der automatischen Texterkennung untersucht, beschrieben und ggf. optimiert.

Für die Verzeichnisse der deutschsprachigen Drucke werden seit 2006 über 1 Mio. historische Bestände digitalisiert und online der Forschung zur Verfügung gestellt. Mit Hilfe von OCR-Verfahren können aus diesen Bilddaten durchsuchbare Volltexte automatisch generiert werden. Das zusätzliche Vorliegen von Volltexten wird mittlerweile auf breiter disziplinärer Front als Schlüssel zu einer ganzen Reihe von geistes- und kulturwissenschaftlichen Forschungsfragen gesehen und gilt zunehmend als elementare Voraussetzung für die Weiterentwicklung der transdisziplinär arbeitenden Digital Humanities.

Um die OCR-Behandlung für diese große Menge historischer, digitalisierter Drucke vorzubereiten sucht das Projekt OCR-D Antworten auf aktuelle technische, informationswissenschaftliche und organisatorische Herausforderungen. So fehlt es z.B. an historischen Textkorpora und lexikalischen Ressourcen; Bedarf besteht auch bei Weiterentwicklungen im Bereich der Nachkorrektur sowie im Bereich der Genauigkeitsberechnung.

Die Präsentation auf dem 13. PhilTag-Workshop an der UB Würzburg gibt einen Überblick über den aktuelle Erkenntnisstand und die weitere Projektarbeit.

Projektpartner sind die Herzog August Bibliothek Wolfenbüttel, die Bayerische Staatsbibliothek in München sowie die Berlin-Brandenburgische Akademie der Wissenschaften, im Besonderen das Deutsche Textarchiv (DTA) in Berlin. Unterstütz wird das Vorhaben durch Experten, Wissenschaftler und Bibliotheken. Das Projekt wird durch die Deutsche Forschungsgemeinschaft (DFG) gefördert.

Elisa Herrmann

Projekt OCR-D

Herzog-August-Bibliothek Wolfenbüttel