Von Handarbeit zur Massenware - OCR als Grundlage für die Forschung in der Wissenschaftsgeschichte

Die Verfügbarkeit von digitalen Volltexten in größerem Rahmen eröffnet neue Möglichkeiten für die historische Forschung. Dieses gilt sowohl für die kultur- und sprachübergreifende als auch für die epochenübergreifende Forschung. Am Max-Planck-Institut für Wissenschaftsgeschichte (MPIWG) werden daher seit der Gründung des Institutes 1994, Tools und Methoden entwickelt, die disziplinübergreifende Textanalysen möglich machen. Technische Lösungen und modulare Workflows erlauben eine effiziente Texterfassung durch Transkription (double-keying). Die sich nun schnell entwickelnden Methoden von Optical Character Recognition (OCR) machen es nun auch möglich, sowohl frühneuzeitliche Texte als auch moderne Texte mit OCR zu erschließen.

Es eröffnen sich damit für die historischen Forschung neue textbasierte Option und neue Strategien. In diesem Kontext wurden am MPIWG zwei neue Forschungsprojekte begonnen, die OCR-Technologien einbeziehen. In dem einen steht die Analyse von Überlieferungstraditionen im Zentrum, die sich an der Verbreitung eines für die Wissenschaftsgeschichte zentralen Werkes der frühen Neuzeit, der *Sphaera* des Sacrobosco, festmachen lassen. Ursprünglich war dies ein Text über Astronomie und Kosmologie, der im Laufe seiner Editionsgeschichte immer wieder durch zusätzliche Texte erweitert und umfänglich kommentiert wurde. Er gehörte zum verbindlichen Wissenskanon der Universitäten seiner Zeit: von 1472 bis 1650 haben wir bisher 363 Editionen identifiziert, die in ganz Europa veröffentlicht wurden. Von der Seite der Digital Humanities stellen wir uns neben netzwerktheoretischen Überlegungen die Frage, ob sich OCR-Technologien zur Analyse der editionsgeschichtlichen Veränderung dieses Werkes eignen. Und welche Informationen müssen zusätzlich extrahiert werden, um die strukturellen Veränderungen des Werkes zu erfassen?

Ein anderes Projekt am MPIWG beschäftigt sich mit mehrere Kilometer umfassenden Aktenbeständen, die bisher noch nicht erschlossen wurden. Diese sollen mit OCR erfasst und mit Methoden des Textmining, wie Topic-Modelling aber auch Named Entity Recognition (NER) bearbeitet werden. Für beide Projekte werden die Ansätze und bisherigen Erfahrungen präsentiert und zur Diskussion gestellt.

Dirk Wintergrün

Max-Planck-Institut für Wissenschaftsgeschichte (MPIWG)