

Universitätsbibliothek Würzburg, Am Hubland, 97074 Würzburg

Öffentlichkeitsarbeit, Dr. Katharina Boll-Becht,

Tel.: +49 931 31-85938

E-Mail: oeffentlichkeitsarbeit@bibliothek.uni-wuerzburg.de

Pressemitteilung (08.06.2017)

Anerkennung im Bereich der Texterkennung (OCR):

Best-Paper-Award für Dr. Uwe Springmann und Münchner Kollegen

Die Konferenzreihe DATeCH (Digital Access to Textual Cultural Heritage) hat sich als internationale Fachtagung für die neuesten Techniken der Digitalisierung etabliert und bringt Forscher und anwendende Bibliotheken zusammen, die innovative Ansätze entwickeln, um historische Dokumente in digitale Formen zu transformieren. In diesem Jahr richtete die Staats- und Universitätsbibliothek (SUB) Göttingen die Veranstaltung aus. Der Schwerpunkt lag auf der automatischen Texterkennung (Optical Character Recognition, OCR) sowie der Nachkorrektur und Weiterverarbeitung der Daten.

Im Rahmen der Tagung wurde auch der „Best-Paper-Award“ verliehen. Der neue Leiter der Abteilung Digitalisierung / Kallimachos-Zentrum für Digital Humanities der Universitätsbibliothek Würzburg, Dr. Uwe Springmann, wurde zusammen mit seinen Münchner Kollegen Florian Fink und Klaus U. Schulz für die Arbeit im Bereich nachträgliche Textverbesserungen der OCR-Ergebnisse ausgezeichnet. Ihr Paper „Profiling of OCR’ed Historical Texts Revisited“ ist über <https://arxiv.org/> verfügbar: <https://arxiv.org/abs/1701.05377>

Neben Uwe Springmann waren weitere Mitarbeiter der Universität Würzburg an Papers beteiligt, die auf der Konferenz präsentiert und diskutiert wurden:

Christian Reul, Mitarbeiter am Lehrstuhl Informatik VI, stellte ein neues Tool für die halbautomatische Segmentierung von Digitalisaten vor. Dieser Vorverarbeitungsschritt erleichtert und verbessert OCR-Ergebnisse: „LAREX – A semi-automatic open-source Tool for Layout Analysis and Region Extraction on Early Printed Books“ (<https://arxiv.org/abs/1701.07396>). Die neue Software kam in Zusammenarbeit mit der Universitätsbibliothek schon zum Einsatz: In einer Anwendungsstudie wurde ein Werk aus dem Zeitalter des Frühdrucks ge-

scannt, mit LAREX semantisch segmentiert und anschließend mittels einer speziell auf diesem Werk trainierten OCR erfasst. Dabei wurde eine Zeichenerkennungsrate von mehr als 97 Prozent erzielt, die seit 2014 aufgrund der Arbeiten von Herrn Springmann und seinen Würzburger Kollegen möglich ist, während vor 2014 Inkunabeln für nicht OCR-fähig gehalten wurden. Das Paper von Christian Reul, Marco Dittrich und Martin Gruner ist ebenfalls auf <https://arxiv.org/> einsehbar:

“Case Study of a highly automated Layout Analysis and OCR of an incunabulum: ‘Der Heiligen Leben’ (1488)”: (Link: <https://arxiv.org/abs/1701.07395>)

Die Tagung unterstrich einmal mehr die Bedeutung der Digitalisierung für die Forschung. Digitalisierung des kulturellen Erbes ist notwendiger denn je und das nicht nur aus konservatorischen Gründen. Sie liefert einen digitalen Datenschatz, der automatisiert weiterverarbeitet werden kann. Eine Führung im Digitalisierungszentrum der SUB Göttingen rundete das Tagungsprogramm ab. Fast 200 Terabyte umfasst mittlerweile der Bestand an gescannten Büchern, die in Kooperation mit dem Rechenzentrum der Universität Göttingen jedem Nutzer weltweit kostenfrei zur Verfügung gestellt werden. Ein Ziel, das auch die UB Würzburg anstrebt.

Kontakt:

Dr. Uwe Springmann, Universitätsbibliothek, Abteilung Digitalisierung / Kallimachos-Zentrum für Digital Humanities

Tel.: +49 931 31-85697, E-Mail: uwe.springmann@bibliothek.uni-wuerzburg.de



vlnr: Dr. Marco Büchler (Organisator der Tagung), Dr. Uwe Springmann (UB Würzburg), Prof. Dr. Klaus-U. Schulz (UB München), Florian Fink (UB München), Prof. Dr. Apostolos Antonacopoulos (Organisator der Tagung)
Foto: Greta Franzini, Digital-Humanities-Center Göttingen