Martina Tumulla

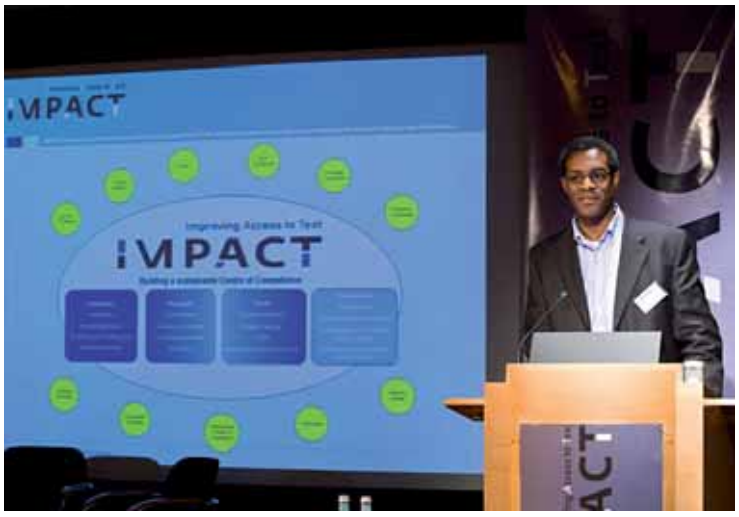
IMPACT-Konferenz 2009

OCR in Mass Digitisation – Challenges between Full Text, Imaging and Language

Vom 6. bis 7. April 2009 fand die erste IMPACT-Konferenz in Den Haag statt (Improving Access to Text). Themenschwerpunkt lag auf der Texterkennung in Massendigitalisierungsprojekten, speziell in Bezug auf historischen Text.

Mehr als 130 Teilnehmer aus über 30 Ländern kamen zur IMPACT-Konferenz in die Königliche Bibliothek (KB), die Nationalbibliothek der Niederlande. Vertreter von Bibliotheken, Forschungseinrichtungen, Dienstleister und Wirtschaftsunternehmen nutzten die Gelegenheit zur Kontaktaufnahme und des gegenseitigen Informationsaustausches rund um die Inhalte Texterkennung und Massendigitalisierung. Besonders die Themenvielfalt der Vorträge führte zu einer erfolgreichen Konferenz. Auch die gelungene Mischung von Teilnehmern und Vortragenden aus den verschiedenen Ländern und Sparten trug zu einer interessanten Konferenz und einer angenehmen Atmosphäre bei. Die Tagung wurde auch dazu genutzt, um das Projekt IMPACT einem größeren Publikum vorzustellen und einen Ausblick auf die kommende Projektlaufzeit zu geben.

Rund 130 Konferenzteilnehmer aus über 30 Ländern



Aly Conteh, The British Library
Foto: Königliche Nationalbibliothek der Niederlande / Jacqueline de Kort, Jos Uljee

Was ist IMPACT?

Am IMPACT-Projekt[1] sind 15 internationale Bibliotheken, Forschungseinrichtungen und Wirtschaftsunternehmen unter der Leitung der KB beteiligt. Das von der Europäischen Kommission im Rahmen des 7. EU-Forschungsrahmenprogramms (FP7) geförderte Projekt ist auf insgesamt vier Jahre angelegt und endet im Dezember 2011. Hauptziele von IMPACT sind u. a. der Aufbau eines Kompetenzzentrums für Massendigitalisierung von historischen Quellen und die Verbesserung der Texterkennung für historische Texte.[2]

Automatische Texterkennung

Automatische Texterkennung – auch Optical Character Recognition (OCR) genannt – ist für historische Textquellen sehr komplex und auch heutzutage noch nicht ausgereift. Hierbei werden die Seiten gescannt und durch ein Softwareprogramm analysiert. Die Erkennungsrate liegt weit unter den Ergebnissen für moderne Texte. Um eingescannte historische Textquellen elektronisch durchsuchbar zur Verfügung zu stellen, ist eine gut funktionierende Texterkennung mit sehr hoher Erkennungsrate unentbehrlich.

Nach der Konferenzeröffnung durch Hans Jansen, KB, dem Vorsitzenden der IMPACT General Assembly, ordnete Pat Manson (European Commission's Cultural Heritage and Technology Enhanced Learning Unit) in ihrem Eröffnungsvortrag mit dem Titel »Digitisation of Cultural Resources: European Actions and the Context of IMPACT« das Projekt in den Zusammenhang der europäischen Digitalisierungsbestrebungen ein, die das Ziel haben das europäische Erbe digital zur Verfügung zu stellen. Im Anschluss stellte Hildelies Balk, KB, das Projekt IMPACT näher vor, indem sie den Hintergrund und die Herausforderungen von IMPACT und die Bedeutsamkeit der Nachhaltigkeit betonte. Astrid Verheusen, KB, stellte die Probleme und Herausforderungen der Bibliotheken bezüglich der Massendigitalisierung dar. Sie erörterte einige Möglichkeiten, um die Effizienz von Digitalisierungsprojekten zu erhöhen anhand von neuen Veränderungen im Digitalisierungsworkflow der Königlichen Bibliothek der Niederlande.

In zwei Sessions wurden einige IMPACT-Tools vorgestellt, die sich mit der Verbesserung der Texterkennung und der Planung von Digitalisierungsprojekten beschäftigen.

Impact Tools

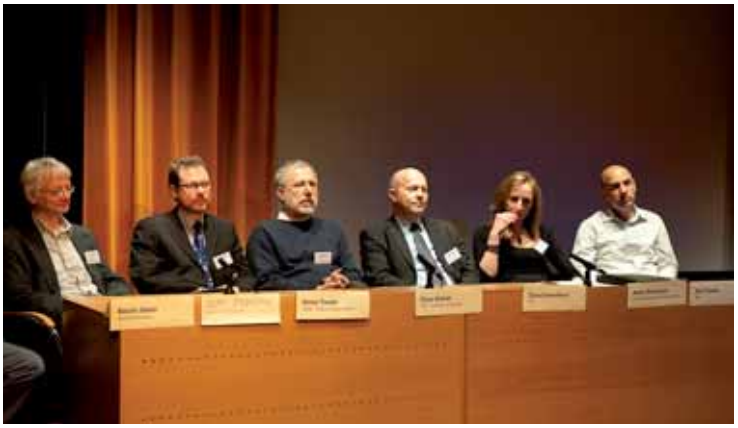
Asaf Tzadok, IBM Haifa Research Laboratory, erklärte das Prinzip der »Adaptive OCR« – eine anpassungsfähige Texterkennung, die sich selber an die Besonderheiten der jeweiligen Schriftart adaptiert. Zusammen mit einem gemeinschaftlichen Korrekturprogramm, in dem Internetbenutzer mitwirken können, sollen die Ergebnisse der OCR deutlich verbessert werden. Basilis Gatos, National Center for Scientific Research »Demokritos«, stellte die »Enhancement and Segmentation Platform« vor, die die derzeitigen Entwicklungen und die neuen Erkenntnisse im Bereich Entzerrung (dewarping), Entfernung der Rahmen (border removal) und Buchstabensegmentierung (character segmentation) interaktiv vergleichend darstellt. Mit dem Vortrag »Language Technology for Improving OCR on Historical Texts« präsentierte Klaus Schulz, Centrum für Informations- und Sprachverarbeitung CIS, München, die Notwenigkeit der Verbesserung der Texterkennung durch Sprachtechnologie. So ist es für historische Testquellen außerordentlich wichtig, die Rechtschreibungsvarianten der Zeitperiode zu berücksichtigen, in der das gescannte Werk verfasst wurde. Der Aufbau von Wörterbüchern für einzelne spezifische verschiedene Zeitperioden und Sprachen ist in diesem Zusammenhang besonders relevant.

Anhand von Beispielen erklärte Apostolos Antonacopoulos, University of Salford, in seinem Vortrag »Digital Restoration and Layout Analysis« Problemstellungen bezüglich der Scanqualität und verwies auf Forschungsaktivitäten im Bereich geometrische Korrekturen, Entfernung von Rahmen und Binarisierung. Ein weiteres Forschungsfeld ist die Segmentierung, die automatische Erkennung von Textbereichen.

Anschließend legte Katrien Depuydt, Institute for Dutch Lexicology, Leiden, den Fokus auf das Thema »Historical Lexicon Building and How it Improves Access to Text«. Der Einfluss von Historischen Lexika auf die Texterkennung, um historische Sprachbarrieren zu minimieren, wurde verdeutlicht und die Bedeutung für das Retrieval veranschaulicht. Ein Toolpaket für den Aufbau dieser historischen Lexika wird im Rahmen von IMPACT entwickelt. Neil Fitzgerald, The British Library, London, stellte die IMPACT »Decision Support Tools« vor, welche Entscheidungshilfen für Digitalisierungsprojekte anhand von Fallstudien und Dokumenten anbieten werden.

Im Rahmen einer Podiumsdiskussion bestand die Gelegenheit Fragen an einige der Vortragenden und andere Experten zu stellen. Moderiert wurde dieser Teil der Veranstaltung von Günter Mühlberger, Universitätsbibliothek Innsbruck. Die Hauptdiskussionsthemen waren u. a. die Fragestellung, ob Inhouse-Lösungen oder dem Outsourcen der Vorzug gegeben werden sollte sowie die Qualität von OCR Ergebnissen.

Podiumsdiskussion



Teilnehmer der Podiumsdiskussion
Foto: Königliche Nationalbibliothek der Niederlande / Jacqueline de Kort, Jos Uljee

Simon Tanner, King's College London, beschäftigte sich in seinem Vortrag mit der Messung der OCR-Qualität am Beispiel des Zeitungsarchivs der British Library. Er unterstrich nicht nur die Bedeutung der Buchstabengenauigkeit, sondern vor allem die Worterkennungsrate und die Richtigkeit der signifikanten Worte, die die Suchgenauigkeit beeinflussen.

Sehr viel Resonanz erlangte Rose Holley, National Library of Australia, mit ihrem Vortrag »Many Hands Make Light Work: Collaborative OCR Text Correction in Australian Historic Newspapers«, indem sie das Digitalisierungsprojekt der historischen australischen Zeitungen vorstellte. Bei der »Collaborative Correction« können Internetbenutzer die Texterkennung der australischen Zeitungen über eine Plattform verbessern. Die besonders

positiven Erfahrungen des »gemeinschaftlichen Korrigierens«, wie z. B. die ausdauernde Bereitschaft der Internetnutzer, das Fehlen von mutwilligen Falscheinträgen und die wachsende Zahl an Freiwilligen wurden herausgestellt.

Der Vortrag von Claus Gravenhorst (CCS Content Conversion Specialists GmbH, Hamburg) befasste sich mit den zukünftigen technischen Herausforderungen der Texterkennung. Der Referent erläuterte, dass obwohl große Fortschritte im Bereich der OCR für historische Quellen erreicht wurden, in Zukunft besonders »next-level OCR« (wie die Berücksichtigung der Dokumentstruktur) gebraucht wird, um die Prozesse der OCR zu verbessern und zu beschleunigen.

Bevor die Konferenz mit einem Rückblick zu Ende ging, hatten die Teilnehmer die Möglichkeit an einer von drei parallel stattfindenden Veranstaltungen teilzunehmen. Zur Auswahl standen, neben einer Führung durch die Digitalisierungsabteilung der KB, die Diskussionen über das geplante IMPACT Kompetenzzentrum oder über die technischen Fragestellungen, die Bibliotheken bezüglich Massendigitalisierung und Texterkennung beschäftigen.

Die Präsentationen und weitere Informationen zur Konferenz sind auf den Webseiten des Projektes zu finden.[3] Die zweite IMPACT-Konferenz wird voraussichtlich im Jahr 2011 stattfinden.

Ausblick

Anmerkungen

1 <http://www.impact-project.eu>

2 Weitere Informationen: Tumulla, Martina: IMPACT Improving Access to Text. In: Dialog mit Bibliotheken, 20 (2008) 2, S. 39 – 41.

3 <http://www.impact-project.eu/news/ic2009/presentations/>