Auf dem Weg zum digitalen Volltext – ein Konferenzbericht

Von Doris Škarić, Mark-Oliver Fischer und Fedor Bochow

Aus dem ganzen deutschsprachigen Raum kamen am 11. und 12. Oktober 2011 Interessenten aus Bibliotheken, Archiven und Unternehmen an die Bayerische Staatsbibliothek, um an zwei Veranstaltungen teilzunehmen, die aus verschiedenen Blickwinkeln und mit unterschiedlicher Akzentuierung das übergreifende Thema „Historische Dokumente auf dem Weg zum digitalen Volltext“ behandelten.

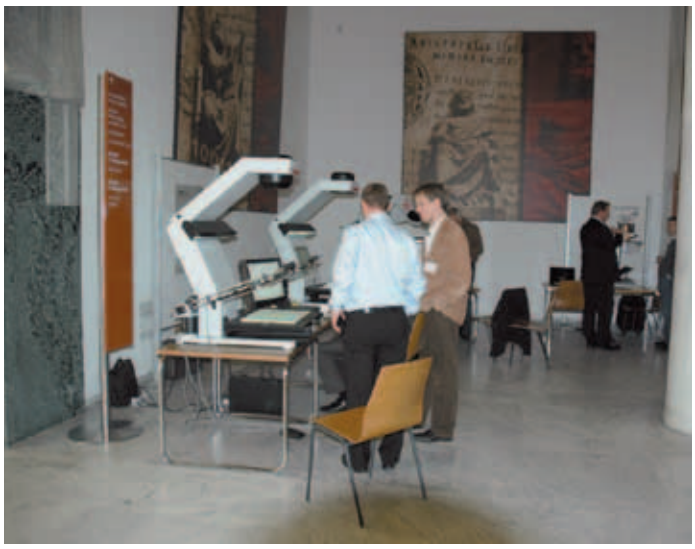
Die BSB ist mit ihrem Münchener Digitalisierungszentrum (MDZ), das auf über ein Jahrzehnt an reichhaltiger Projekterfahrung in Sachen Retrodigitalisierung zurückblicken kann, ein idealer Ort für den Informations- und Wissensaustausch zum komplexen Thema der Digitalisierung bibliothekarischer Bestände. Während bei der Retrodigitalisierung bisher überwiegend die möglichst effiziente und kostengünstige Erstellung hochwertiger digitaler Abbildungen im Mittelpunkt stand, ist eines der nächsten großen Ziele die flächendeckende Bereitstellung von Volltexten, durch die die Suche und das Auffinden einschlägiger Texte erheblich erleichtert wird. Auf dem Weg dorthin sind allerdings noch viele Hürden zu nehmen. Kommerzielle Texterkennungssoftware war bisher vor allem auf moderne Gebrauchstexte (Firmenkorrespondenz, Formulare etc.) ausgerichtet, das Verbesserungspotential im Bereich der bibliothekarischen Nutzung ist entsprechend groß. Vermutlich zog die Veranstaltung auch deshalb so viele interessierte Besucher an. Weil die Kapazitätsgrenze des Veranstaltungssaals mit seinen 90 Sitzplätzen schnell erreicht war, musste die Anmeldeliste sogar vorzeitig geschlossen werden. In der Nähe des Vortragssaals konnten sich die Teilnehmer an den Ständen diverser Sponsoren – darunter Beratungsfirmen, Software-Anbieter und Scanner-Hersteller – über deren Produkte und Dienstleistungen informieren.

IMPACT Demo Day

Der erste Tag – gemeinsam mit der Österreichischen Nationalbibliothek (ÖNB) veranstaltet – legte als Teil einer Reihe von europaweit stattfindenden sogenannten „IMPACT Demo Days“ den Fokus auf „Ergebnisse aus der aktuellen OCR-Forschung“. IMPACT (IMProving ACcess to Text, www.digitisation.eu) ist ein von der Eu-

Das Organisationsteam

FOTOS: MÜNCHENER DIGITALISIERUNGSZENTRUM/BAYERISCHE STAATSBIBLIOTHEK



V. l. n. r.: Sponsoren im Marmorsaal, Prof. Dr. Manfred Thaller, Dr. Annette Gottscharek, Dr. Sven Slarb, Dr. Henning Pahl

ropäischen Kommission seit 2008 gefördertes Forschungsprojekt zur Verbesserung der computerbasierten Erkennung von historischen Texten, das 26 Partner aus 13 Ländern zusammenbringt. Das gemeinsame Ziel der beteiligten Bibliotheken, Forschungsinstitute und kommerziellen Partner ist es, die Digitalisierung und Volltexterstellung historischer Drucke maßgeblich zu verbessern, einschlägige Kompetenz in diesem Bereich zu bündeln und beides europaweit zu verbreiten. Die BSB und die anderen im Projekt vertretenen Bibliotheken sind in erster Linie für die Erstellung bibliotheksspezifischer Anforderungsprofile sowie für die Bereitstellung von geeignetem Testmaterial und für das Testen der von den anderen Projektpartnern entwickelten Programme zuständig. Die beteiligten Forschungsinstitute beschäftigen sich unter anderem mit der Entwicklung historischer Wörterbücher, die von entscheidender Bedeutung für die Verbesserung der Erkennungsergebnisse sind, während die kommerziellen Partner ihre bestehenden Software-Lösungen an die Bedürfnisse von Bibliotheken und deren historische Bestände anpassen.

Im Folgenden sollen einige der spannenden und lehrreichen Vorträge kurz vorgestellt werden. Projektleiterin Hildelies Balk-Pennington de Jongh (Koninklijke Bibliotheek der Niederlande in Den Haag) schilderte anfangs kurz die Hintergründe, Ziele und Ergebnisse des Projekts.

Annette Gotscharek von der LMU München sprach über die Entwicklung historischer Speziallexika. Da sich OCR-Software bei der Texterkennung auch auf Wörterbücher stützt und sich die Texte früherer Jahrhunderte in Rechtschreibung und Wortschatz teils deutlich von der heute verwendeten Sprache unterscheiden, müssen diese entsprechend angepasst bzw. neu erstellt werden. Einem weiteren wichtigen Aspekt für die Erstellung korrekter Volltexte widmeten sich Doris Škarić (BSB) und Ulrich Reffle (LMU), die von zwei verschiedenen Ansätzen zur Nachkorrektur von OCR-Ergebnissen berichteten, während Stefan Platschacher von der University of Salford Methoden zur Messung der Qualität von OCR-Ergebnissen präsentierte. Einige der vorgestellten Softwarelösungen konnten von den Teilnehmern auch in den Pausen zwischen den Vorträgen an eigens dafür bereitgestellten Computern ausprobiert werden. Am Ende stellte Frau Balk-Pennington de Jongh das „IMPACT Centre of Competence“ vor, das die Erkenntnisse und Entwicklungen des Projekts über dessen Laufzeit hinaus zur Verfügung stellen und als europaweiter Ansprechpartner für alle Fragen der Digitalisierung und Erstellung von Volltexten dienen wird – hieran wird sich auch die BSB voraussichtlich beteiligen.

Erfahrungen aus der Digitalisierungspraxis

Den „Erfahrungen aus der Digitalisierungspraxis: OCR, Volltexte und Präsentationsformen“ war der zweite Tag gewidmet. Mehrere Referenten präsentierten Herausforderungen und Lösungsansätze aus den an ihren Institutionen betriebenen Digitalisierungsprojekten. Der Wunsch nach Bereitstellung von durchsuchbaren Volltexten prägt dabei die Konzeption des gesamten Ablaufs eines Digitalisierungsprojekts nachhaltig – beginnend bei der Projektdefinition und der Auswahl der optimalen Hardware bis hin zur abschließenden Präsentation der historischen Dokumente im Internet.

Manfred Thaller (Universität Köln), ein Pionier der Retrodigitalisierung in Deutschland, stellte seinen Einführungsvortrag unter den von Friedrich Schiller inspirierten Titel „Was heißt und zu welchem Ende betreiben wir Volltextdigitalisierung?“.

Vom Ideal zum Konkreten ging es dann in den folgenden Vorträgen. Die Vorstellung von Projekten, die sich der Digitalisierung von Monographien, Zeitungen, Archivmaterial und Nachlässen widmeten, zeigten, welche Herausforderungen diese sehr unterschiedlichen Ausgangsmaterialien an Digitalisierung und Volltexterstellung stellen. Henning Pahl (Bundesarchiv Berlin) etwa wies in seinem Beitrag darauf hin, dass die Nutzer des Bundesarchivs momentan weniger an der Digitalisierung einzelner Akten als vielmehr an digitalisierten Findmitteln, also den Verzeichnissen des Archivbestands, interessiert seien. Maria Federbusch von der Staatsbibliothek zu Berlin berichtete von einem Praxistest, bei dem zwei verschiedene OCR-Software-Lösungen anhand von Funeralschriften – protestantischen Leichenpredigten des 17. und 18. Jahrhunderts – einander gegenübergestellt wurden. Marco Büchler vom Lehrstuhl für Automatische Sprachverarbeitung an der Universität Leipzig präsentierte die Möglichkeiten der computergestützten Korrektur und Rekonstruktion von Texten durch sogenanntes „Text Mining“. Constanze Hofmann zeigte das Potential ehrenamtlicher Arbeit in Zeiten des Internets am Beispiel von Distributed Proofreaders, einer Plattform, die Inhalte für Project Gutenberg erstellt. Nach diesem Überblick über die verschieden Methoden der Volltexterstellung widmeten sich Dirk Scholz (BSB) und Christa Müller (ÖNB) den von ihren Institutionen beschrittenen Wegen der Nutzung und Bereitstellung der Volltexte.

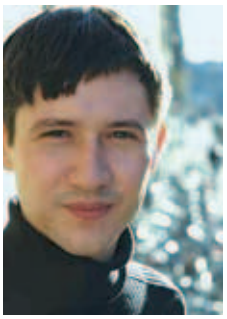
Abschließend berichtete Matthias Leopold von der Deutschen Zentralbücherei für Blinde zu Leipzig über das immer mehr in den Fokus rückende Thema Barrierefreiheit von digitalen Angeboten und darüber, wie Volltexte hierbei helfen könnten.

Auch diese kurz vorgestellten Vorträge stellen natürlich nur eine Auswahl dar, mehr Informationen zu allen Referaten finden Sie am Ende des Artikels.

Die Rückmeldungen zu den beiden Veranstaltungstagen waren insgesamt sehr erfreulich: „Sehr gut, dass auch mehrere Archivare referierten“, „Sehr positive Mischung der Vortragsthemen“ und „Veranstaltung diente als erste Orientierung für mögliches Projekt“ sind nur eine Auswahl der eingegangenen Reaktionen.

Wenn Sie sich weiter über die Inhalte der beiden Veranstaltungstage informieren wollen, finden Sie diese frei zugänglich im Internet. Einen Überblick über das Programm gibt es unter www.muenchener-digitalisierungszentrum.de/~lza/impact/. Videos aller, auch der im Artikel nicht vorgestellten, Vorträge des „IMPACT Demo Day“ sowie Materialien weiterer Veranstaltungen des Projekts finden Sie unter http://impactocr.wordpress.com; die Vorträge der „Erfahrungen aus der Digitalisierungspraxis“ können auf http://mdzblog.wordpress.com nachverfolgt werden.

Blick ins Publikum



DIE AUTOREN
Doris Škarić, Fedor Bochow und Mark-Oliver Fischer sind Mitarbeiter des EU-Projekts IMPACT im Referat Münchener Digitalisierungszentrum/ Digitale Bibliothek der Bayerischen Staatsbibliothek.