Martina Tumulla

IMPACT

Improving Access to Text

Schon heute gibt es viele ambitionierte Digitalisierungsprojekte der unterschiedlichsten Ausprägungen. Zeitaufwändig und kostenintensiv ist nicht nur die Anschaffung der Geräte und der Scanvorgang selbst, sondern vor allem die Nachbereitung beziehungsweise die Korrektur der Scans.

Durch einen normalen Scanvorgang werden Dokumente nur als Bilder ohne zusätzliche Informationen abgespeichert. Um einen besseren Zugang zum Text der Dokumente zu finden, muss eine Software vorangeschaltet werden. Durch die so genannte »Optical Character Recognition« (OCR)-Software werden die von den gescannten Seiten hergestellten Bilddateien durch die automatische Texterkennung in durchsuchbare Volltexte verwandelt. Bei modernen Texten werden durch die automatische Zeichenerkennung fast 100 % aller Buchstaben und Zeichen korrekt erkannt. Anders ist die Situation bei historischen Dokumenten, wo die Schriftzeichenerkennung durch verschiedene Faktoren erschwert wird. Zurzeit gibt es keine OCR-Software, die mit akzeptablem Ergebnis die Schriftzeichen von historischen Quellen erkennt.

OCR-Texterkennung

Vor allem die unterschiedlichen Rechtschreibvarianten, komplexen Layouts, Schreibmaschinen- oder Handschriften und ältere Schriftarten stellen ein Problem dar. Ein Beispiel für eine ältere Schriftart ist die Frakturschrift, die eine besondere Herausforderung an die OCR stellt. Nicht nur, dass diese Schriftart unterschiedliche Ausprägungen aufweisen kann, sie besticht auch durch die Ähnlichkeit der einzelnen Buchstaben untereinander, die es manchmal selbst dem geübten menschlichen Auge erschwert, den Inhalt richtig zu deuten. Die andersartigen Druckprozesse, Alterungs- und Gebrauchsspuren durch jahrelange Benutzung beziehungsweise Brand- oder Wasserschäden erschweren den Prozess der OCR. Zusätzlich birgt jede Sprache andere Problemstellungen im Bereich der automatischen Texterkennung, wie die unterschiedlichen grammatikalischen Sprachstrukturen und -entwicklungen.

Was ist IMPACT?

An dieser Stelle setzt das von der Europäischen Kommission im Rahmen des »Siebenten Forschungsrahmenprogramms« (FP7) geförderte Projekt IMPACT ein. Die Projektdauer ist von Januar 2008 bis Dezember 2011 angesetzt.

Die Ziele von IMPACT gehen weit über die einfache Zeichenerkennung von gescannten historischen Quellen und deren Verbesserung hinaus. Die OCR-Technologie soll durch unterschiedliche Aspekte weiterentwickelt werden, um eine schnellere und bessere Massendigitalisierung zu erreichen. Ziel hierbei ist es, dass eingescannte Texte die gleichen Funktionalitäten und Eigenschaften wie original digitale Quellen aufweisen.

Die neuen Softwareprogramme beziehungsweise verschiedenartige Softwarekomponenten sollen den heutigen Standard in Form von Qualität und Schnelligkeit übertreffen. Ferner sollen Best-Practice-Grundsätze für die Massendigitalisierung der historischen Quellen formuliert werden, um eine Vereinheitlichung von Digitalisierungsprojekten herbeizuführen. Mit der technischen Weiterentwicklung gehen Aufbau und Erweiterung von sprachlichen Werkzeugen und Lexika für verschiedene Sprachen einher. Englisch, Deutsch und Niederländisch sollen derzeit aufgebaut werden, gegebenenfalls werden weitere Sprachen im Rahmen dieses Projektes angegliedert. Durch spracheigene Lösungswege zur Texterkennung werden historisch bedingte Rechtschreibungs- und Vokabularvarianten berücksichtigt.

Formulierung von Best-Practice-Grundsätzen

Darüber hinaus soll die freiwillige Mitarbeit der Internetnutzer verstärkt werden, indem eine neue Softwareplattform die »collaborative correction« – Bewertung und Korrektur – erleichtert und die Popularität dieses Web 2.0-Angebots erhöht. Ein

weiteres Ziel besteht darin, die Kosteneffizienz durch eine Verbesserung der Automatisierung zu steigern und so gegebenenfalls Digitalisierungsprogramme zu ermöglichen.

Nicht zuletzt sollen die gewonnenen Erkenntnisse anderen Interessenten aus dem Bereich der Text-Digitalisierung zur Verfügung gestellt werden, um ihnen den Einstieg in die Digitalisierung zu erleichtern. Hierzu werden in Zukunft zahlreiche Informationen über das Projekt und deren Ergebnisse online über die Internetpräsenz des Projektes[1)] und sicherlich über Publikationen angeboten. Ein zentraler Anlaufpunkt soll für alle Arten von Interessensgruppen bereitgestellt werden, in dessen Hintergrund ein Netzwerk an Kompetenzzentren steht. Auch durch Schulungen und Vorführungen sollen Interessierte gewonnen werden, Digitalisierungsvorhaben weiter voranzutreiben.

Schon allein die Kombination der an diesem Projekt beteiligten National- und Staatsbibliotheken, Forschungsinstitute und Wirtschaftsunternehmen verspricht verschiedene Herangehensweisen an das Problem. Die Zusammenarbeit der Projektmitglieder soll Experten aus den zum Teil fachlich sehr unterschiedlich ausgerichteten Institutionen zusammenführen und so die Sichtweisen jedes einzelnen erweitern. Außerdem können die entstandenen Synergieeffekte genutzt werden, um nicht nur für die zahlreichen Sprachen Wörterbücher und Erschließungsmöglichkeiten auf- und auszubauen, sondern auch das »Mehrfachentwickeln« zu vermeiden.

Projektpartner

Folgende 15 Projektpartner stehen unter der Leitung des Projektkoordinators, Koninklijke Bibliotheek der Niederlande, Den Haag:

- Koninklijke Bibliotheek, Den Haag, Niederlande,
- The British Library, London, Großbritannien,
- Österreichische Nationalbibliothek, Wien
- Universität Innsbruck, Österreich,
- Deutsche Nationalbibliothek, Leipzig, Frankfurt am Main, Berlin, Deutschland
- Bayerische Staatsbibliothek, München, Deutschland,
- Staats- und Universitätsbibliothek Göttingen, Deutschland,
- ABBYY Production LLC, Moskau, Russland,
- IBM Israel – Science and Technology Ltd., Haifa, Israel,
- Instituut voor Nederlandse Lexicologie, Leiden, Niederlande,
- National Centre for Scientific Research »Demokritos«, Athen, Griechenland,
- Centrum für Informations- und Sprachverarbeitung, LMU München, Deutschland,
- University of Bath, Großbritannien,
- University of Salford, Großbritannien,
- Bibliothèque nationale de France, Paris.

Projektstruktur

Das Projekt ist neben der Koordinierung in vier Sub-Projects (SP) untergliedert. Die erste Einheit ist der »Operational Context« (SP-OC). Hierbei wird der Fokus auf interne Anforderungen, Arbeitsabläufe und Evaluationsverfahren gelegt. Als weiteres Sub-Project beschäftigt sich »Text Recognition« (SP-TR) mit dem technischen Teil der Texterkennung. Das dritte Sub-Project »Enhancement & Enrichment« (SP-EE) betrifft die Erstellung der Lexika, die Verbesserung des Textzuganges u. a. durch Nutzung nationaler Normdaten und das gemeinschaftliche Korrektur-Projekt. Ergänzend befasst sich »Capacity Building« (SP-CB) mit der Informationsvermittlung an interne und externe Interessensgruppen, sei es durch Internetpräsenz, Helpdesk, FAQs oder durch Publikationen, Workshops und Vorführungselemente. Jedes der Unterprojekte unterteilt sich wiederum in verschiedene Arbeitspakete.

Projektaufgaben DNB

Die Deutsche Nationalbibliothek (DNB) mit ihren Abteilungen Informationstechnik und Digitale Dienste und der Arbeitsstelle für Standardisierung (AfS) beteiligt sich an verschiedenen Arbeitspaketen und ist federführend bei der Implementierung des Projekt-Helpdesks. Hierbei werden die Anfragen von Interessenten durch eine Kommunikationsplattform aufgenommen und zu den geeigneten Experten zur Beantwortung weitergeleitet. Ein anderes Arbeitspaket entwickelt das interne Anforderungsforum, während sich ein weiterer Arbeitsbereich um die technischen Grundvoraussetzungen beziehungsweise um die übergeordnete Softwarearchitektur kümmert. Ferner ist die DNB an der Anreicherung der Wörterbücher durch die Normdaten beteiligt.

Als Europäisches Projekt soll IMPACT den jetzigen und zukünftigen Digitalisierungsprojekten hinsichtlich Kosteneffizienz und Qualität der Voll-

Ausblick

texterkennung nutzen. All diese Bemühungen sollen abschließend auch der Europäischen Digitalen Bibliothek zugutekommen und diese mit Leben füllen, um die historischen Quellen des europäischen Erbes zu digitalisieren, so wie es die Europäische Union in ihrer i2010 Vision der Europäischen Digitalen Bibliothek als Ziel definiert hat.

Anmerkungen

1 <www.impact-project.eu/>