Integrierte Medienerschließung in der Staatsbibliothek zu Berlin: Ein Praxisbericht über die Digitalisierung dreier DDR-Zeitungen

Integrated content analytics at Staatsbibliothek zu Berlin: A report on the digitization of three GDR newspapers

Almut Ilsen[1], Dr. Stefan Paal[2], Dr. Michael Eble[3]

[1]Staatsbibliothek zu Berlin Preußischer Kulturbesitz, Potsdamer Str. 33, 10785 Berlin, WWW: staatsbibliothek-berlin.de
Tel. +49 30 266 433171, Fax. +49 30 266 333171, Email: almut.ilsen@sbb.spk-berlin.de

[2,3]Fraunhofer IAIS, Schloss Birlinghoven, 53754 Sankt Augustin, WWW: www.iais.fraunhofer.de
[2]Tel.: +49 2241 14 3438, Fax: +49 2241 144 3438 E-Mail: stefan.paal@iais.fraunhofer.de
[3]Tel.: +49 2241 14 3406, Fax: +49 2241 144 3406 E-Mail: michael.eble@iais.fraunhofer.de

Zusammenfassung:
Die Staatsbibliothek zu Berlin hat im Rahmen eines von der DFG geförderten Projekts drei DDR-Zeitungen digitalisiert, im Volltext erschlossen und für die wissenschaftliche Forschung frei zugänglich und unentgeltlich zur Verfügung gestellt. Das Projekt wurde im Jahr 2009 begonnen, Anfang 2012 konnten die ersten Jahrgänge und Ende Mai 2013 alle Zeitungen vollständig im Portal „DDR-Presse“ präsentiert werden. Die Verarbeitung der Zeitungsdigitalisate wurde vom Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme geleistet. Mit einer speziellen Fraunhofer-Technologie wurden die Seitenbilder in einzelne Artikel segmentiert (Optical Layout Recogniton, OLR) und mit optischer Zeichenerkennung (Optical Character Recognition, OCR) verarbeitet. Anschließend hat der Dienstleister ArchivInForm die erzeugten Volltext- und Metadaten manuell nachbearbeitet. Durch diese Kombination von automatischen und manuellen Verfahren konnten Ergebnisse mit sehr geringen Fehlerquoten erreicht werden.

Abstract:
In a project of the German Research Foundation (DFG), the Berlin State Library (Staatsbibliothek zu Berlin) digitized and indexed three GDR newspapers and put them online for scientific researchers free of charge. The project started in 2009, in 2012 first issues were presented and in 2013, the whole content was published on the web portal "DDR-Presse". The Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) conducted the automatic indexing of the digitized newpapers. By using a unique technology for optical layout recognition (OLR), the newspaper articles were separated and the text content was extracted by means of optical character recognition (OCR). The project partner ArchivInform performed the manual quality assessment and verified the fulltext and metadata results. As a result, the innovative combination of automatic indexing services and manual quality assessment tools managed the article separation with low failure rates.

1. Motivation und Problemstellung

Zeitungen sind kulturhistorisch eine wichtige Informationsquelle für den gesellschaftlichen Diskurs und die Einordnung von Themen in den geschichtlichen Kontext. Auf dem Weg zur vernetzten Wissensgesellschaft kommt der digitalen Aufbereitung, Bereitstellung und Recherchemöglichkeit von Originalzeitungen eine besondere Bedeutung zu. Die inhaltliche Erschließung von Zeitungsdigitalisaten auf Seitenebene würde der inhaltlichen Strukturierung in Einzelartikel, die wiederum aus verschiedenen Elementen bestehen, nicht gerecht werden. Die Segmentierung der Artikel bzw. die Layouterkennung stellt durch den unterschiedlichen Aufbau und die veränderliche Anordnung von Artikeln eine besondere Herausforderung dar. Die konventionelle Texterkennung ist nicht in der Lage, einzelne Artikel seitenübergreifend zu separieren und ihre Bestandteile wie Überschriften, Autorenangaben, Textkörper, Bilder und Bildunterschriften zu identifizieren. Aber

nur mit diesen Metadaten kann ein zeitgemäßes Zeitungsarchiv aufgebaut werden, das die Suche, Referenzierung und Anzeige auf Artikelebene unterstützt.

2. Projektziel und Ausgangslage

Die Berliner Staatsbibliothek hat im Rahmen des von der Deutschen Forschungsgemeinschaft geförderten Projekts „DDR-Zeitungsportal: Digitalisierung von DDR-Zeitungen und Aufbau eines Portals zur Presse der DDR mit wissenschaftlicher Forschungsumgebung“ drei wichtige Tageszeitungen der SBZ (Sowjetische Besatzungszone) und der DDR digitalisieren sowie einer Layouterkennung bzw. Artikelsegmentierung (Optical Layout Recogniton, OLR) und einer optischen Zeichenerkennung (Optical Character Recognition, OCR) im Volltext erschließen lassen. Ergänzend erarbeitete das Zentrum für Zeithistorische Forschung Potsdam (ZZF) einen wissenschaftlichen Apparat zum Pressesystem der DDR.

Das Projekt umfasst das „Neue Deutschland“ (1946-1990), das sogenannte Zentralorgan der SED (Sozialistische Einheitspartei), die „Berliner Zeitung“, die Bezirkszeitung der SED für Berlin (1945-1990, als Folgeprojekt mit dem Berliner Verlag fortgesetzt bis Ende 1993), und die „Neue Zeit“, die Zeitung der Blockpartei CDU (1945 bis zu deren Erscheinungsende 1994). Diese zeithistorischen Quellen mit insgesamt ca. 400.000 Seiten und ca. 4 Millionen Artikeln stehen Wissenschaftlerinnen und Wissenschaftlern bzw. allen Interessenten nach erfolgter Registrierung frei zugänglich, unentgeltlich und komfortabel recherchierbar innerhalb des Zeitungsinformationssystems der Staatsbibliothek ZEFYS http://zefys.staatsbibliothek-berlin.de/ddr-presse/ zur Verfügung.

Die Zeitungen lagen in gedruckter und gebundener Form vor. Insbesondere die Jahrgänge 1945 bis ca. 1955 befanden sich in einem problematischen konservatorischen Zustand. Zerfallendes Papier bzw. ein unsauberes Druckbild führten zu Text- und damit Informationsverlusten. Weitere Verluste entstanden infolge fehlender Ausgaben, Beilagen, Seiten, Artikel und Bilder. Um einen möglichst vollständigen Bestand zur Verfügung stellen zu können, wurden fehlende und fehlerhafte Ausgaben und Seiten ermittelt und aus anderen Bibliotheken bzw. Archiven beschafft. Dies gestaltete sich als mehrstufiger Prozess, wobei Aufwand-Nutzen-Abwägungen zwangsläufig zum „Mut zur Lücke“ führen mussten. Nichtsdestotrotz konnten insbesondere bei „Berliner Zeitung“ und „Neue Zeit“ nahezu lückenlose virtuelle Bestände zusammengeführt werden. Es wurde die Entscheidung getroffen, die gebundenen Bände aufzutrennen, sowohl um Schriftverzerrungen am Falz und daraus resultierende OCR-Probleme zu eliminieren als auch aus wirtschaftlichen Gründen (Scankosten). Die Scans des „Neuen Deutschland“ wurden vom Verlag Neues Deutschland zur Verfügung gestellt, das Scannen der beiden anderen Zeitungen wurde vom MIK-CENTER Berlin-Blankenburg durchgeführt.

3. Automatische Medienerschließung

Für die Verarbeitung der DDR Zeitungen kamen automatische Verfahren zum Einsatz, die eine strukturelle, inhaltliche und semantische Erschließung effizient und kostengünstig durchführen (Abbildung 1).

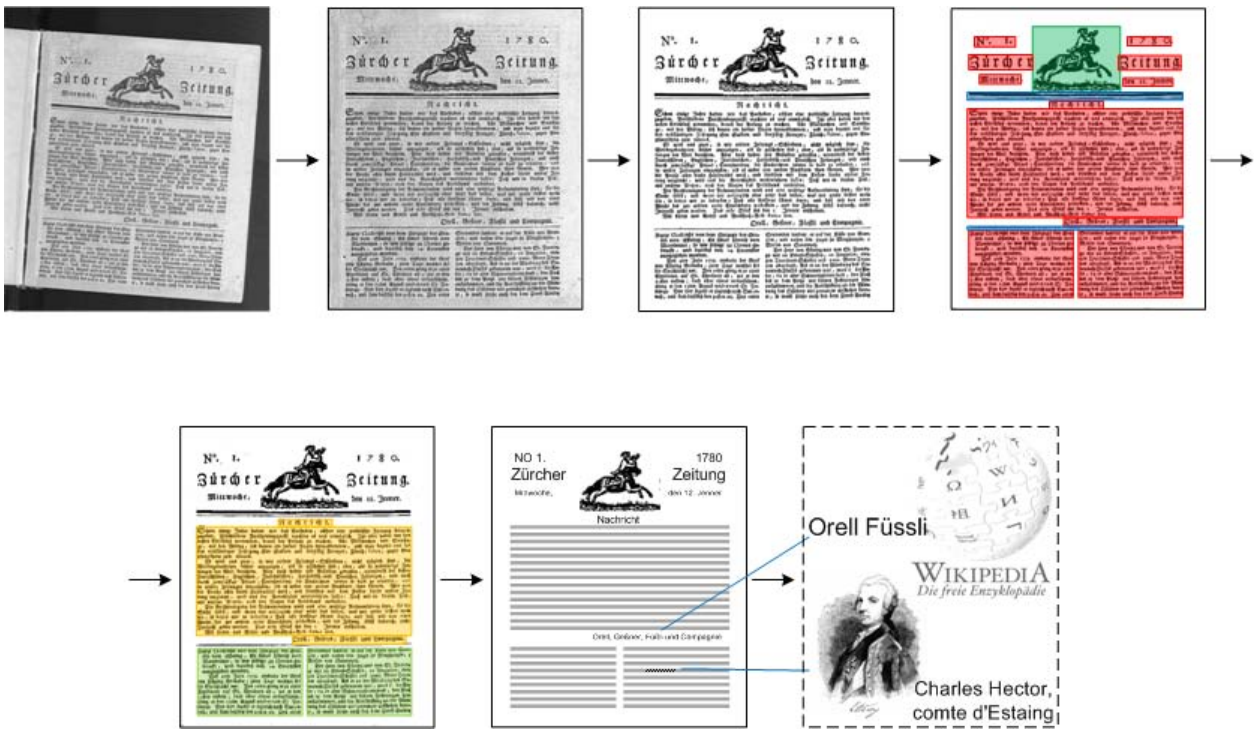

Abbildung 1: Automatische Erschließung von Zeitungsbeständen

In der strukturellen Erschließung wurden die Zeitungsdigitalisate optisch aufbereitet und in Ausgaben zusammengefasst. Dazu zählt zunächst die Freistellung und Rotationskorrektur sowie die Kontrastoptimierung und Schärfung der einzelnen Zeitungsseite. Anschließend wurde eine Seitensegmentierung durchgeführt, die die Seitenelemente wie Überschriften, Bilder und Textblöcke erkennt. Anschließend erfolgten die Klassifikation der resultierenden Layoutelemente und die Zuordnung zu den ursprünglichen Artikeln mittels einer optischen Layouterkennung (Optical Layout Recognition, OLR). Dabei kamen neben probabilistischen Modellen auch regelbasierte Verfahren zum Einsatz, die je nach Layoutformat angepasst und trainiert wurden. In der inhaltlichen Erschließung erfolgten die eigentliche Texterkennung (Optical Character Recognition, OCR) mittels Abbyy FineReader und die Zusammenführung von seitenübergreifenden Artikeln. Damit wurde die Grundlage für die perspektivisch weitere Erschließung mittels semantischer Verfahren, z.B. Eigennamenerkennung und Verlinkung mit externen Datenquellen gelegt. Die komplette Durchführung der automatischen Erschließung erfolgte in einer kontrollierten, verteilten Rechnerumgebung bei Fraunhofer IAIS, so dass keine Zeitungsdigitalisate an Drittanbieter weitergereicht wurden.

4. Manuelle Qualitätssicherung

Da rein maschinelle Ansätze für die Artikelseparierung nach dem heutigen Stand der Forschung nicht immer allen Qualitätsanforderungen in der Praxis genügen, setzt Fraunhofer IAIS auf ein kombiniertes Modell aus automatischer und manueller Verarbeitung. Hierzu werden Softwarewerkzeuge eingesetzt, die eine schnelle Sichtung und effiziente Nachbearbeitung ermöglichen (Abbildung 2).

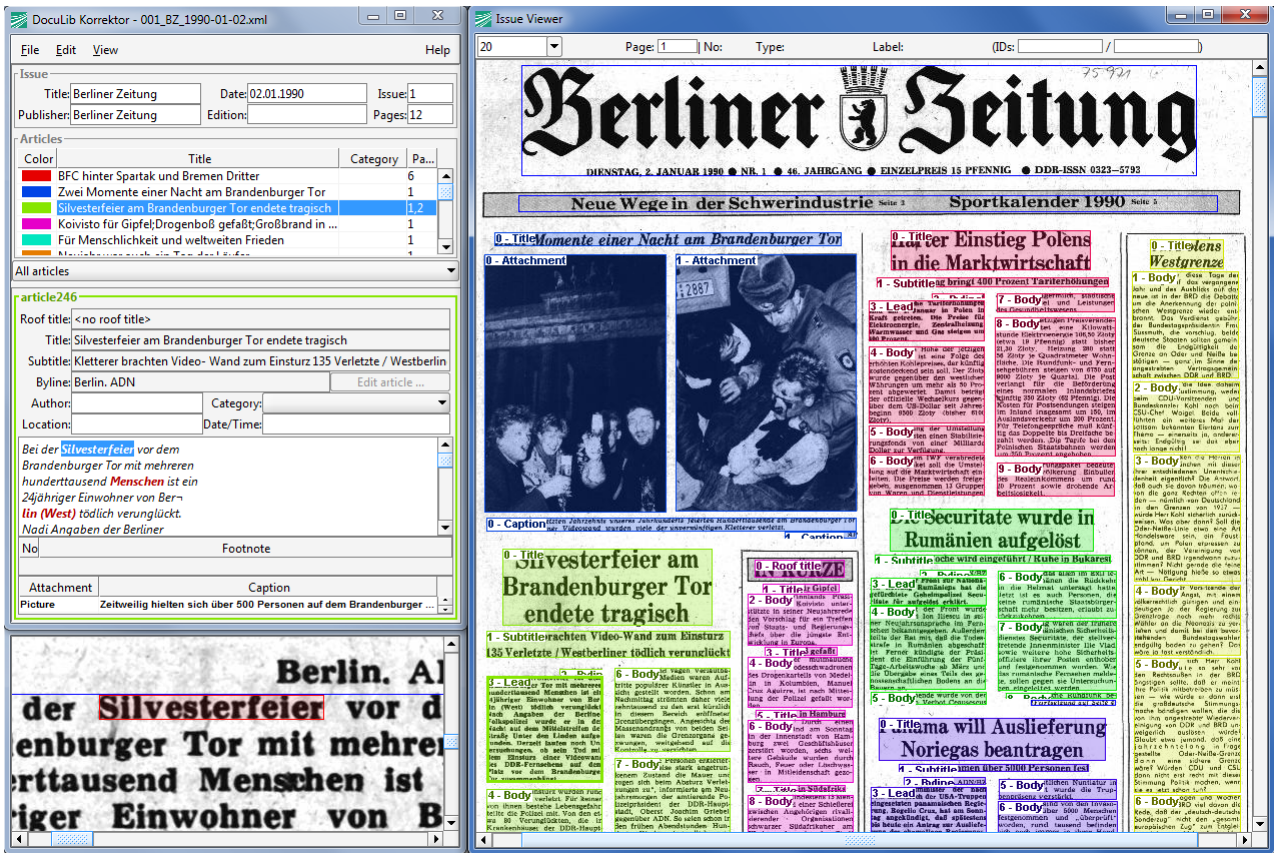

Abbildung 2: Software-Werkzeug für die manuelle Qualitätssicherung der Artikelseparierung

Zur Qualitätssicherung der Artikelseparierung der DDR Zeitungen wurden die Ergebnisse der automatischen Erschließung eingelesen und verschiedene Korrekturmöglichkeiten angeboten. Zum einen konnten die verschiedenen Artikel farblich hervorgehoben und die zugehörigen Seitenelemente neugruppiert und zugewiesen werden. Zum anderen bestand die Möglichkeit der inhaltlichen Annotation von Artikeln, z.B. Autorenangaben, und der Korrektur von OCR-Ergebnissen. Das Software-Werkzeug wurde von Anfang an als Rich Client Anwendung implementiert, so dass auch eine Korrektur über das Internet möglich wurde. Damit konnten in der Praxis manuelle Erschließungsaufwände gezielt an den Dienstleister ArchivInform ausgelagert und die geforderten Qualitätsvorgaben erreicht werden. Die Ergebnisse wurden anschließend an Fraunhofer IAIS zurückübermittelt und gemeinsam mit den Seitenbildern an die Staatsbibliothek zu Berlin ausgeliefert.

5. Präsentation

Die Seitenbilder der digitalisierten Zeitungen sowie die Metadaten aus der Erschließung wurden in das Zeitungsinformationssystem der Staatsbibliothek ZEFYS eingelesen und auf einer von der Staatsbibliothek entwickelten Präsentationsoberfläche als „DDR-Presse“ - Portal für die interessierte Öffentlichkeit frei geschaltet. Den Nutzerinnen und Nutzern stehen damit über 400.000 historische Zeitungsseiten als Faksimile-Ansichten und im Volltext zur Verfügung. (Abbildung 3).

Abbildung 3: Zeitungsportal ZEFYS/DDR-Presse mit Artikelrecherche

Da die DDR-Zeitungen dem Urheberrecht unterliegen, wurden Verträge mit den Zeitungsverlagen und den Verwertungsgesellschaften VG Wort und VG Bild Kunst geschlossen. Daraus resultiert, dass sich die Nutzerinnen und Nutzer des Portals „DDR-Presse“ registrieren müssen. Dies ist möglich über einen Bibliotheksausweis der Staatsbibliothek zu Berlin, über das Deutsche Forschungsnetz DFN oder über den Open-id-Account xlogon.net. Der Rechercheeinstieg kann über eine Kalenderfunktion oder als Volltextsuche erfolgen. Die Volltextsuche bietet auch Suchoptionen über ausgewählte Zeiträume und einzelne Artikelelemente. Die Trefferlisten lassen sich durch Facettierungen (Jahr, Monat, Seite, Zeitungstitel) einschränken. In der Ergebnisanzeige werden Faksimiles der Seiten bzw. Artikel und die Volltexte der Überschriften bzw. Artikeltexte sowie weiterführende Inhalte des Zentrums für Zeithistorische Forschung Potsdam (ZZF) und Verlinkungen zu „Wer war wer in der DDR“, einer Datenbank der Bundesstiftung zur Aufarbeitung der SED-Diktatur angezeigt.

6. Ergebnisse und Ausblick

Durch die integrierte Medienerschließung gelang es, Ergebnisse mit sehr geringen Fehlerquoten zu erreichen. Diese lagen bei der OLR für das „Neue Deutschland“ bei 1,60 %, für die „Berliner Zeitung“ bei 0,70 % und die „Neue Zeit“ bei 0,95 %. Die OCR-Fehlerquote beläuft sich für die „Berliner Zeitung“ auf 0,40 % und die „Neue Zeit“ auf 0,48 %.

Gegenwärtig wird im Rahmen einer Projekterweiterung eine Eigennamenerkennung (Named-Entity-Recognition, NER) durchgeführt. Dabei werden Personennamen, Orte, Länder, Organisationen und Abkürzungen in den DDR-Zeitungen automatisch erkannt und zu externen Datenquellen, z.B. Wikipedia und der Gemeinsamen Normdatei (GND) der Deutschen Nationalbibliothek verlinkt.