Erweiterte Layout- und Textanalyse von Zeitungsartikeln zur Gewinnung von Meta-Daten

Martin Tölle, Dr. Xia Wang, Dr. Wolfgang Schade Gesellschaft zur Förderung angewandter Informatik e.V. (GFaI)
Forschungsbereich Dokumentenmanagement
Volmerstr. 3, 12489 Berlin
Tel.: 030 814 563 470, Fax: 030 814 563 302
E-Mail: schade@gfai.de, Internet: www.gfai.de

In den letzten Jahren sind in den Archiven und Bibliotheken verstärkt Bestrebungen zu beobachten, vorhandene Bestände einem größeren Nutzerkreis zugänglich zu machen, zumindest durch digitalisierte Findbücher, möglichst jedoch durch vollständige Digitalisierung zeitgeschichtlich wertvoller Dokumente. Hierbei bietet sich die Möglichkeit, die Images der Dokumente und ihren digitalen Inhalt vollständig ins Internet zu stellen, an. Bei dieser Art der Veröffentlichung von Werken des 20. Jahrhunderts ist jedoch das bestehende Urheberrecht zu beachten. Dadurch sind nicht nur Nutzungsbeschränkungen zu berücksichtigen, sondern es fallen i.a. nicht unerhebliche Kosten für die Veröffentlichung an, die an eine Verwertungsgesellschaft abzuführen sind. Eine kostenfreie Lösung besteht darin, nicht den vollen Artikeltext im Internet zu veröffentlichen, sondern nur die META-Daten der entsprechenden Artikel, also z.B.

Zeitungsname,
Erscheinungsdatum,
Artikelüberschrift,
Artikelautor und
Art des Artikels.

In Fortführung bisheriger Arbeiten zur automatisierten Inhaltserfassung von Dokumenten wurde in den letzten Jahren unser System DaCaPo (s. Vortrag EVA-Konferenz 2011 Berlin) insbesondere in diese Richtung erweitert. Als Beispieldokumente wurden dazu Zeitungsausschnitte aus dem Archiv des Herder-Instituts für historische Ostmitteleuropaforschung Marburg herangezogen. Neben Fotos, Karten und archivalischen Materialien besitzt dieses Institut eine bedeutende Sammlung von Zeitungsausschnitten mit thematischem Schwerpunkt Baltikum sowie der historischen deutschen Ostgebiete. Die Besonderheit an diesem Beispielmaterial besteht in der großen Heterogenität des Materials; für die Sammlung wurden Ausschnitte aus über 300 Zeitungen (in- und ausländische) mit unterschiedlichstem Seitenlayout herangezogen. Dadurch bedingt war es notwendig, entweder sehr allgemeine Kriterien für die Extraktion der gewünschten Daten oder alternativ dazu mehrere Methoden dafür zu verwirklichen.

DaCaPo Bestandteile sind:

A. Preparator

Im Preparator werden die als JPEG oder TIFF vorliegenden Images ausgerichtet, eventuelle Ränder beseitigt und binarisiert.

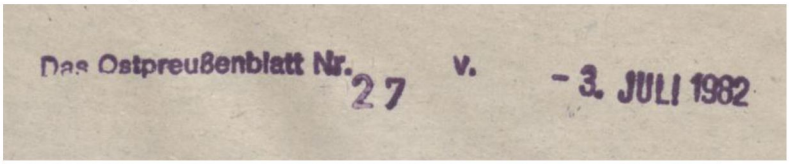
B. Analysator

Für die Gewinnung der oben erwähnten META-Daten ist zunächst eine Separierung von Stempeln, Zeitungsköpfen, Bild- und Textbereichen notwendig. Nahezu vollständig werden auch Artikelüberschriften und Bildunterschriften, bei „normalem“ Layout auch die Zuordnung der Artikeltextblöcke zu der entsprechenden Überschrift automatisch erkannt und entsprechend klassifiziert.

C. Evaluator: Auswertung der separierten Bereiche zur Gewinnung der META-Daten

1. Zeitungsname und Erscheinungsdatum

A. Aus Stempeln

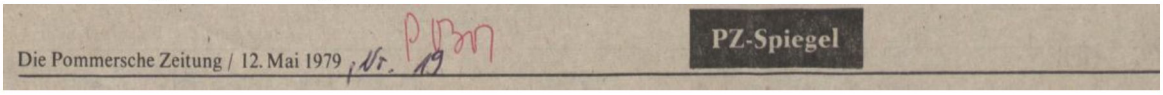


Dazu sind folgende Teilschritte erforderlich

1. Stempelbereich im Image finden
2. Stempel identifizieren
   Dazu müssen die in Frage kommenden Stempel möglichst in größerer Anzahl „angelernt“ werden, so dass durch die so erstellten „Fingerprints“ ein Vergleich möglichst das richtige Ergebnis liefert.
3. Erkennung der numerischen Handschrift im Stempel (Erscheinungsdatum)

B. Aus Zeitungsköpfen der Folgeseiten

Es wird im Kopf des Images nach einem durchgehenden Trennungsstrich gesucht. Der darüber befindliche Text wird nach OCR-Erkennung hinsichtlich Zeitungsnamen und Erscheinungsdatum ausgewertet.



C. Aus der Titelseite

Der Zeitungskopf wird zunächst als Bild klassifiziert. Durch Abgleich mit den angelernten vorkommenden Zeitungsköpfen wird der Zeitungsname ermittelt und aus dem darunter liegenden Text und dem OCR-Ergebnis das Erscheinungsdatum.

2. Artikelüberschriften

Sofern es sich im Image nur um einen einzigen Artikel handelt, ist die Artikelüberschrift z.B. durch die abweichende Buchstabengröße erkennbar. Schwieriger ist der Fall, wenn mehrere derartige Fälle im Image vorkommen. Handelt es sich dabei um Zwischenüberschriften oder um einen gesonderten Artikel?

Überschätzt und unterschätzt: Hans Bredow

Zum hundertsten Geburtstag des „Vaters des deutschen Rundfunks"

Artikel in „Standard"-Layout

sungsgerichts, den Noten der Verbündeten zur Fortgeltung der Berliner Viermächteerklärung gründet und die Lage ohne jede Grenzaner-

vor wenigen Tagen hat mir auf eine Frage das Auswärtige Amt schriftlich mitgeteilt, daß Deutschland weiterhin die Annexion der baltischen

Die Geschichte geht weiter

kennung (vgl. Scheel 9. 2. 1972 vor dem Bundesrat) „beschreibt".
Es wird aber auch das in vollem Gegensatz dazu vom rechtlichen Untergang Deutschlands ausgehende Görlitzer Abkommen bekräftigt, dessen Nichtigkeit die Regierung Adenauer nach vorheriger Erklärung des Bundestages (Löbe) und die drei westlichen Hohen Kommissare als Verstoß gegen die Viermächteverantwortung festgestellt haben. Welche Bindungswirkung kann die Bekräftigung eines so diffusen Vertragsinhaltes haben? Zu Potsdam selbst stellte die Bundesregierung Brandt/Scheel 1972 mit Bezug auf den Text des Warschauer Vertrages fest: „Eine endgültige Festlegung der deutsch-polnischen Grenze blieb (in Potsdam) ausdrücklich einer friedensvertraglichen Regelung vorbehalten". Die Selbstbestimmung des ganzen deutschen Volkes, also der

Staaten wegen gewaltmäßiger Grenzveränderungen nicht anerkenne und der Art. 2 des neuen Partnerschaftsvertrages mit Moskau ebenso wie Art. 3 des Moskauer Vertrages von 1970 nur zum Gewaltverzicht verpflichtet (vgl. Gromyko-Erklärung). Sollte das gar auch für Warschau gelten?

Die Geschichte geht weiter! Noch oft wird versucht werden, Unrecht als endgültig darzustellen. Das heutige Dokument von Warschau ist mehr als brüchig. Der BdV wird weiterhin den friedlichen Wandel zu differenzierten Lösungen im Sinne seiner bisherigen Vorschläge und eines dauerhaften glaubwürdigen historischen Ausgleichs mit den Nachbarn sowie vorerst praktische Maßnahmen zum Wiederaufbau der gestörten Strukturen und zum Schutz der Deutschen vertreten.

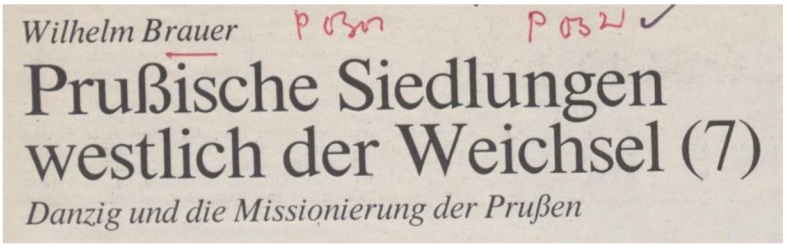
Artikel in „Standard"-Layout oder Text nur durch (Teil-) Überschrift unterbrochen?

An dieser Stelle kann eine **einfache Textanalyse** der OCR-Ergebnisse bereits etwas weiterhelfen. Diese Methoden werden eingesetzt, um die Textblockreihenfolge eines Artikels, der durch spaltenübergreifende Abbildungen oder (Teil-)Überschriften unterbrochen ist, zu bestimmen. Mit anhand des OCR-Textes einfach zu überprüfenden Kriterien können so häufig Vorgänger-Nachfolger-Beziehungen ermittelt werden.

3. Autoren

Wegen der Heterogenität des Materials sind die Autoren bzw. die Textquellen an unterschiedlichen Stellen zu finden.

1. *Oberhalb der Artikelüberschrift*


Wilhelm Brauer

Prußische Siedlungen westlich der Weichsel (7)

Danzig und die Missionierung der Prußen

2. *In der Unterzeile der Überschrift („Von ….“)*


Deutschland:

Egon Bahr schwer belastet

Täuschung des Parlaments über Transitabkommen mit „DDR“?

VON Dr. OTTFRIED HENNIG MdB

Die Bundesregierung steht vor folgender Alternative: Entweder hat Egon Bahr als Staatssekretär im Bundeskanzleramt das Parlament bewußt … truppen, nicht aber für frühere Angehörige der Volkspolizei, des Ministeriums für Staatssicher-

3. *Zu Beginn des ersten Artikeltextblockes :*


Wie es in Moskau wirklich war . . .

H. W. — Mitunter braucht es seine Zeit, bis Vermutungen ihre Bestätigung finden. Etwa die Vermutung, daß die sozialliberale Regierung Brandt/Scheel den Vertrag mit der Sowjetunion deshalb mit besonderem

4. *Am Ende des letzten Artikeltextblockes*


genommen haben.

„Erinnerung ist ein Paradies aus dem man nicht vertrieben werden kann.“

Konrad Werner (SN)

ordens der Bundesrepublik Deutschland. Die Nachfolge Albrecht Baehrs im Funkhaus Stuttgart wird der bisherige Leiter des SDR-Büros für öffentliche Veranstaltungen, Gottfried Wolf, antreten. H.R.F.

Diese Fälle müssen einzeln untersucht werden und stellen deshalb hohe Anforderungen sowohl an die Layoutanalyse wie auch an die Bestimmung der Textblockreihenfolge.

4. Artikelart

Untersucht wird hier hauptsächlich der erweiterte Überschriftenbereich nach Charakterisierungen wie „Reportage“, „Bericht“, „Roman“, „Interview“, „Gespräch“. Interviews liegen auch vor, wenn in den Artikeltextblöcken zwei Parteien abwechselnd vorkommen.

In der Artikelsammlung kommen auch Biografien und Traueranzeigen vor. Hier wird der Text nach bestimmten charakteristischen Wortgruppen abgesucht (z.B. „Trauer“, „Hinterbliebene“, „Beisetzung“, „Friedhof“).

5. **Abbildungsquelle**

   Voraussetzung für das Auffinden der Abbildungsquelle ist, dass durch die Layoutanalyse die Bildungsunterschrift exakt ermittelt wurde. Gewöhnlich steht die Quelle dann in der letzten Zeile rechts, oft hinter „Foto:“, „Photo.“

ert Fröbe und Elisabeth Bergner in Wolfgang Liebeneiners Film **Der Garten** ach einer Komödie von Tim Aspinall (ZDF, 19.30 Uhr) FOTO: URSULA RÖHNERT

D. Korrektor

Die gefundenen Meta-Daten können mit einem Kontroll- und Korrekturinterface ergänzt bzw. korrigiert werden. Das Endergebnis wird in einer MySQL-Datenbank abgelegt. Die Resultate lassen sich sowohl hausintern wie auch für Internet-Präsentationen nutzen.

DaCaPo wurde u.a. auch auf einem Workshop in Marburg vorgestellt und stieß auf großes Interesse sowohl von Archiven und Bibliotheken wie auch von KmU´s. Über eine Verwertung der Ergebnisse laufen mit diesen Interessenten inzwischen teils intensive Verhandlungen.

Die Entwicklung des Programms wurde mit Mitteln des BMWi im Rahmen des Programms INNOKOM-OST gefördert.