Almut Ilsen

Ein Schatz wird gehoben – digitalisierte volltexterschlossene Zeitungen sowie ein Werkstattbericht zum Portal »DDR-Presse«

Digitalisierung und Volltexterschließung werten Zeitungen zu attraktiven Forschungsobjekten auf. Die im Vergleich zu analogen Zeitungen bedeutend komfortableren Recherchemöglichkeiten und die Nutzung im Rahmen der Digital Humanities werden erläutert. Es folgt ein kurzer Abriss über den derzeitigen Stand der Zeitungsdigitalisierung im internationalen und nationalen Kontext. Der Hauptteil umfasst einen detaillierten Werkstattbericht zu dem an der Staatsbibliothek zu Berlin realisierten DFG-Projekt »DDR-Zeitungsportal« von der Rechteproblematik über das Scannen, die Layout-, Text- und Entitätenerkennung bis zur Präsentation und der Verknüpfung mit weiteren Inhalten. Abschließend wird auf bereits laufende bzw. geplante Folgeprojekte hingewiesen, in denen die Zeitungskorpora als Forschungsdaten für computerlinguistische Analysen dienen.

Digitisation and full text indexing are turning newspapers into attractive research resources. Their significantly more user-friendly search possibilities in comparison to analogue newspapers and their use within the field of digital humanities are described. This is followed by a brief outline of the current state of newspaper digitisation in the international and national context. The main section includes a detailed workshop report on the DFG »DDR-Zeitungsportal« project carried out by the Staatsbibliothek zu Berlin (Berlin State Library) covering an array of topics ranging from legal issues, scanning, layout, text and entities recognition through to presentation and links to other content. Finally, ongoing and planned future projects are mentioned in which newspaper corpora serve as research data for computer linguistic analyses.

Einleitung

Wurden Begriffe wie »Ausreiseantrag«, »Waldsterben« oder »Mauertote« in der DDR-Presse verwendet? Und wenn ja, in welchen Zusammenhängen? Wie spiegelt sich der Aufstieg und Fall Wolf Biermanns vom jungen gefeierten Lyriker zum Verfemten und Ausgebürgerten wider? Mit welchen zeitlichen Häufungen und in welchen Zusammenhängen werden die Begriffe »Revanchisten«, »deutscher Friedensvertrag« oder »Entspannungspolitik« verwendet?

Fragen dieser Art können durch Recherchen in gedruckten und mikroverfilmten, also analogen Zeitungen allenfalls mit unverhältnismäßigem Aufwand und oft nur rudimentär beantwortet werden. Denn Zeitungen haben mit Ausnahme der »Times« keine Indizes, sie sind nicht erschlossen im Hinblick auf ihre Nutzung als Quelle für die Forschung. Ihr originärer Zweck ist, Momentaufnahme und ein aus vielen Elementen bestehendes Kaleidoskop eines einzelnen Tages zu sein. Wenn analoge Zeitungen zum Forschungsobjekt werden und keine entsprechende Zeitungsinhaltsbibliografie zur Verfügung steht, ist ein Rechercheeinstieg nur über ein Erscheinungsdatum bzw. über eine bereits bekannte Chronologie möglich. In Anbetracht des immensen Umfangs des Materials kann die Suche nur mit großem Zeitaufwand und ebensolcher Beharrlichkeit zum Erfolg führen. Dies ändert sich grundlegend, wenn Zeitungen digital transformiert werden. Über eine Volltextsuche und mit der richtigen Suchstrategie können die eingangs gestellten Fragen mühelos beantwortet werden. Historische Zeitungen mit ihrer thematischen Fülle und Vielfalt aus Politik, Gesellschaft und Kultur werden zu einer für Forschende durch nichts Adäquates zu ersetzenden Primärquelle.

Foto: Janna Brechmacher

Almut Ilsen

Was ist das Besondere an digitalisierten Zeitungen?

Für die Recherche am Original muss sich die Nutzerin oder der Nutzer zum Objekt des Interesses, der gedruckten oder mikroverfilmten Zeitung, begeben und nicht selten mehr als eine Bibliothek aufsuchen. Denn Zeitungen sind in der Regel nicht vollständig vorhanden – es fehlen Jahrgänge, Ausgaben, Seiten, oder diese sind beschädigt und nicht mehr benutzbar.

virtuell vollständiger Bestand

Bei einer Digitalisierung hingegen können Lücken ergänzt werden. Es entsteht ein virtuell vollständiger Bestand, der online recherchierbar ist. Darüber hinaus sind alle Zeitungen eines Zeitungsportals gemeinsam durchsuchbar. Anstatt aufwändig an verschiedenen Orten arbeiten zu müssen, kann ortsunabhängig, breiter und schneller recherchiert werden. Außerdem ermöglichen digitalisierte Zeitungen nach entsprechender Weiterverarbeitung eine bedeutend tiefere und differenziertere Nutzung als ihre analoge Form. Dies trifft allerdings nicht zu, wenn lediglich Images bzw. Faksimiles online präsentiert werden. Dann beschränkt sich der einzig mögliche Rechercheeinstieg – wie bei analog vorliegenden Zeitungen auch – auf das Datum. Wird eine Zeitung aber digital transformiert und einer OCR (Optical Character Recognition) unterzogen, revolutioniert dies die Recherchemöglichkeiten. Mit einer Stichwortsuche kann der gesamte Inhalt eines digitalen Zeitungsarchivs auf Wortebene durchsucht werden. Zum chronologischen Sucheinstieg über die Kalenderfunktion kommt der inhaltliche über die Volltextsuche hinzu. Bob Nicholson konstatiert: »We are potentially on the cusp [...] of a ›digital turn‹ in humanities

umfassende Recherchemöglichkeiten

scholarship driven by the creative use of online archives and a willingness to imagine new kinds of research. [...] the development of keyword search technology has made it possible to trace the development and movement of ideas and discursive formations in ways that were once impossible. [...] Though a digitized text may look familiar, it is not the same source; we are able to access, read, organize, and analyse it in radical new ways.«[1] Und Adrian Bingham weist darauf hin: »It is far easier, for example, to find out when a subject was first discussed in the press, or when a term was coined.«[2] Erst OCR und Volltextsuche ermöglichen es, den Schatz der Tageszeitungen zu heben!

»Culturomics«

Die einzige Einschränkung besteht darin, dass bei der OCR nicht oder falsch erkannte Begriffe bei der Volltextrecherche nicht gefunden werden können. Wird es den Nutzern eines Portals jedoch ermöglicht, diese Fehler nachträglich zu korrigieren, verbessern sich OCR-Genauigkeit, Datenqualität und damit die Rechercheergebnisse.

optische Layout-Erkennung

Wird vor der OCR eine optische Layout-Erkennung und Artikelsegmentierung (Optical Layout Recognition – OLR) durchgeführt, kann noch differenzierter recherchiert werden. Eine OLR strukturiert zum einen jeden einzelnen Artikel bis auf die Ebene seiner Bestandteile, erkennt Überschriften, Bildunterschriften und Lesereihenfolge eines Artikels und verknüpft zum anderen jene Artikelteile miteinander, die auf verschiedenen Seiten platziert sind. Damit kann die Granularität der Erkennung von der Seitenebene bis auf die Ebene der Artikel bzw. Artikelbestandteile erhöht werden. Die Volltextsuche setzt dann auf dieser Feingranularität auf und kann entsprechend differenzierte Ergebnisse liefern. So wird beispielsweise eine Recherche nur über Artikelbestandteile wie Bildunterschriften oder Überschriften möglich.

Aber damit sind die Nutzungsmöglichkeiten digital transformierter Zeitungsbestände noch nicht erschöpft: Da Zeitungsbestände sehr umfangreich sind, entstehen bei ihrer Retrodigitalisierung sehr große Datenmengen, die eine hervorragende Datenbasis für

Data-Mining-Methoden

Data-Mining-Methoden bieten. Mittels statistischer Methoden werden Häufungen und Muster erkannt. Über Jahrzehnte erschienene Zeitungen laden zu Longitudinaluntersuchungen über lange Zeitverläufe ein. Kontinuitäten, Veränderungen oder Entwicklungsbrüche können in unterschiedlichsten Zusammenhängen erforscht werden.

»Distant Reading«

Franco Moretti entwickelte im Jahr 2000 das Konzept des »Distant Reading«[3] und realisiert im Stanford Literary Lab eine Reihe interessanter Projekte.[4] Im Gegensatz zum »Close Reading«, bei dem aus einer Vielzahl von Texten relevante Texte ausgewählt und diese klassisch gelesen und qualitativ erforscht werden, wendet »Distant Reading« quantitativ-statistische Verfahren an. Bisher verborgene Strukturen werden sichtbar, neue Blickwinkel entstehen. Dazu Adrian Bingham: »[...] historians can situate and contextualize newspaper content far more efficiently and effectively. They can make comparisons over time, between papers and between different media forms, to isolate innovations and specificities and to identify similarities, continuities and borrowings«[5]

Jean-Baptiste Michel u. a. prägten den Begriff »Culturomics« für die quantitative Analyse großer digitaler Archive zur Untersuchung kultureller Entwicklungen und Phänomene. Dabei wird die Sprachnutzung in Abhängigkeit ihres zeitlichen Verlaufs untersucht, und es können sehr differenzierte Aussagen z. B. über Frequenzverläufe, Kollokationen, semantische Netze und Prä- und Suffixe von Komposita getroffen werden. Michel u. a. analysierten beispielsweise über 5 Millionen Monografien – ca. 4 % aller gedruckten Bücher. Ihre Fragestellungen reichten von Sprache, Lexikografie und Grammatik über Wandlungen des kollektiven Gedächtnisses, die Einführung von Technologien bis hin zu Zensur und Unterdrückung.[6]

Volltexterschlossene Zeitungen wurden im letzten Jahrzehnt in den Geistes- und Sozialwissenschaften zu attraktiven Forschungsobjekten: Literaturwissenschaftler können den kulturellen Diskurs nach Erscheinen eines Werkes nachvollziehen, Genealogen und Biographen den Spuren ihrer Subjekte folgen. Linguisten erforschen die Verwendung, Verbreitung und Entwicklung von Sprache und Kulturhistoriker spüren der Entwicklung und Bewegung von Ideen und Diskursen nach.

Wo steht die Zeitungsdigitalisierung international und in Deutschland?

Bisher wurden und werden vor allem Zeitungen und Zeitschriften des 17. bis 19. Jahrhunderts in größerem Umfang digitalisiert (wichtige internationale und länderspezifische Zeitungs- und Zeitschriftenportale s. Abb. 1). Retrodigitalisierungen von Zeitungen des 20. Jahrhunderts gestalten sich aufgrund der Urheberrechtslage als schwierig und stehen deshalb bisher nur in geringem Umfang zur Verfügung.

Bei den großen internationalen Portalen wie »Europeana Newspapers«[7] und den kommerziellen Portalen Proquest[8], Readex[9] und Gale Digital Collections[10] werden bereits viele Millionen Seiten aus tausenden von Titeln angeboten.

Länderspezifische Portale basieren zumeist auf nationalen Digitalisierungsprogrammen und sind in der Regel bei den Nationalbibliotheken angesiedelt.

Region / Land	Portal	Inhalte	Titel-Anzahl	Seiten-Anzahl (Mio)	OCR (Mio)
Europa	Europeana Newspapers	Zeitungen aus 45 europäischen Ländern	Ca. 3.480	18	10
USA + international	Proquest Historical Newspapers	Zeitungen 1764 bis 20. Jhd.	Ca. 3.000	35	35
USA + international	America's Historical Newspapers (Readex)	Zeitungen aus USA, Afrika, Lateinamerika, Südasien 1690 bis 21.Jhd.	Ca. 2.500	k. A.	ja
Großbritannien + USA	Gale Digital Collections	Zeitungen und Zeitschriften	k. A.	10	10
USA	Chronicling America: Historic American Newspapers	Zeitungen 1836–1922	1.904	10	10
Großbritannien	The British Newspaper Archive	Zeitungen 1710–1959	533	12	12
Australien	Trove	Zeitungen 1803–2007	Ca. 1.000	19,2	19,2
Neuseeland	Papers Past	Zeitungen und Zeitschriften 1839–1948	Ca. 120	3	3
Österreich	Anno	Zeitungen und Zeitschriften 1568–1944	Ca. 800	14	Für die Jahre 1689–1918 1938–1944
Spanien	Hemeroteca Digital	Zeitungen und Zeitschriften	1.452	25,1	25,1
Frankreich	Gallica	Zeitungen und Zeitschriften	73.272 Ausgaben	1,5	0,007
Niederlande	Delpher	Zeitungen und Zeitschriften	80	1,5	1,5
Finnland	DIGI – National Library's Digital Collections	Alle Zeitungen 1771–1910	k. A.	3,46 frei: 1,96	3,46
Färöer Inseln, Grönland, Island	Timarit	Ziel: alle Zeitungen und Zeitschriften 1815 bis 2–4 J. vor lfd. Jahr	980	5	5
Belgien	BelgicaPress	Zeitungen 1831–1918	9	4 frei: 1,2	4

Abb. 1: Internationale und länderspezifische Zeitungs- und Zeitschriftenportale. Stand November 2015

Hier sind insbesondere Australien, Neuseeland[11], die Niederlande[12], Österreich, Finnland, Spanien[13], Großbritannien, Frankreich[14] und die USA[15] zu nennen. So plant die British Library, bis zum Jahr 2021 ca. 40 Millionen Seiten zu digitalisieren.[16] Besonders hervorzuheben ist das australische Trove-Portal[17], das sich durch hervorragende Suchfunktionalitäten und differenzierte Filtermöglichkeiten auszeichnet. Erwähnenswert ist, dass die Nutzer des Portals Korrekturen von OCR-Textfehlern vornehmen können und dadurch die Genauigkeit der Volltexte ständig verbessert wird. Auch das ANNO-Portal[18], das die Österreichische Nationalbibliothek aufbaut, beeindruckt durch schnelles Wachstum und seine Strategie. Die in einem ersten Schritt gescannten Zeitungen werden in einem weiteren Schritt einer Volltexterschließung unterzo-

länderspezifische Portale

gen. Derzeit stehen die Zeitsegmente 1689–1918 und 1938–1944 bereits volltexterschlossen zur Verfügung.

Kleinere Länder mit einem überschaubaren Zeitungsbestand digitalisieren im Rahmen der Sicherung ihres kulturellen Erbes teilweise ihren gesamten historischen Zeitungsbestand. Dazu gehört Finnland mit den im Portal »DIGI – National Library's Digital Collections«[19] zwischen 1771 und 1900 erschienenen Zeitungen, wobei die OCR-Fehler durch die Einbeziehung der Nutzer korrigiert werden konnten. Auch das Portal Timarit[20] präsentiert den vollständigen historischen Zeitungsbestand aus Grönland, Island und den Färöer Inseln. Hier können die Zeitungen und Zeitschriften durch rechtliche Vereinbarungen mit den Verlagen bis in die unmittelbare Gegenwart mit einer Moving Wall von zwei bis vier Jahren präsentiert werden.

drei zeithistorisch relevante Tageszeitungen

In Deutschland werden bei den großen Digitalisierungsprojekten von Beständen des 17. bis 19. Jahrhunderts die Zeitungen und Zeitschriften bisher leider vernachlässigt. Es gibt eine Reihe von regionalen und thematischen Portalen, aber sowohl eine umfassende Digitalisierung innerhalb eines nationalen Zeitungsdigitalisierungsprogramms als auch die Präsentation unter einer Oberfläche bleiben bisher ein Desiderat.[21] Zu den größeren Zeitungsdigitalisierungsprojekten gehörte die Mitarbeit im Projekt »Europeana Newspapers«, bei dem an der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (SBB) ca. 1,44 Millionen Seiten aus vier Berliner Tageszeitungen digitalisiert wurden. Auch das DFG-Projekt »Zeitungsdigitalisierung«[22] der SLUB Dresden, der ULB Halle, der BSB München und der SuUB Bremen wird zu einem Zuwachs von ca. 450.000 digitalisierten Zeitungsseiten in Deutschland führen und in einem Masterplan für die Zeitungsdigitalisierung in Deutschland zudem Standards und Strukturen erarbeiten, die auch Anpassungen des DFG-Viewers und der Zeitschriftendatenbank beinhalten.

nationales Zeitungsdigitalisierungsprogramm – in Deutschland ein Desiderat

rund 400.000 Seiten und 4 Millionen Artikel im Portal »DDR-Presse«

Thematische Portale wie z. B. Compact Memory[23] mit 172 deutschsprachigen jüdischen Periodika und die »Amtspresse Preußens«[24] präsentieren Zeitungen mit hohem Quellenwert. Leider können die an der Deutschen Nationalbibliothek (DNB) aufgebauten Portale »Exilpresse digital« und »Jüdische Periodika aus NS-Deutschland«, die schwer zugängliches und wertvolles Quellenmaterial enthalten, aus rechtlichen Gründen derzeit nicht im Internet bereitgestellt werden. Den thematischen Portalen ist auch das Portal »DDR-Presse« zuzuordnen, das sich mit drei wichtigen Tageszeitungen der SBZ/DDR aus dem insgesamt bisher kaum abgedeckten Zeitraum 1945–1994 zu einem zentralen Portal für die zeitgeschichtliche Forschung entwickelt hat.

thematische Portale

Die Realisierung des Portals »DDR-Presse« – ein Werkstattbericht

Allgemeines

Für das DFG-Projekt »DDR-Zeitungsportal – Digitalisierung von DDR-Zeitungen und Aufbau eines Portals zur Presse der DDR mit wissenschaftlicher Forschungsumgebung« wurden drei zeithistorisch relevante Tageszeitungen der SBZ und DDR ausgewählt. Dies sind die Zeitungen »Neues Deutschland«, das Zentralorgan der Sozialistischen Einheitspartei Deutschlands (SED), die »Berliner Zeitung«, die Zeitung der SED für Berlin, sowie die »Neue Zeit«, die Zeitung der CDU in der DDR. Sie war die auflagenstärkste Zeitung einer Blockpartei, richtete sich an konfessionell-kirchliche Bevölkerungsgruppen und wurde exemplarisch für die Presse der Blockparteien in der DDR ausgewählt. Die drei Zeitungen wurden von Beginn ihres Erscheinens 1945/46 bis zum Tag der deutschen Wiedervereinigung, dem 3. Oktober 1990, bzw. die »Neue Zeit« bis zu ihrem Erscheinungsende im Juli 1994 vom Papieroriginal gescannt, einer optischen Layout- und Texterkennung (OLR und OCR) sowie einer automatischen Entitätenerkennung (Named Entity Recognition – NER) für Personen, Orte und Organisationen unterzogen.

Durch ein Folgeprojekt mit dem Berliner Verlag konnten weitere 40.000 Seiten der Berliner Zeitung für den Zeitraum vom 4.10.1990 bis 31.12.1993 in das Portal aufgenommen werden. Da die Berliner Zeitung ab 1994 digital publiziert wird, steht sie damit für ihren gesamten Erscheinungsverlauf online zur Verfügung.

Die drei Zeitungen werden mit ca. 400.000 Seiten und ca. 4 Millionen Artikeln im Portal »DDR-Presse« innerhalb von ZEFYS, dem Zeitungsinformationssystem der SBB präsentiert.[25] Sie können nach Anmeldung von jedem Interessierten weltweit und kostenfrei genutzt werden. Flankierend wurde eine wissenschaftliche Forschungsumgebung durch den Projekt-Kooperationspartner, das Zentrum für Zeithistorische Forschung Potsdam (ZZF), erarbeitet. Zusätzlich werden biografische Informationen aus einschlägigen Datenbanken sowie Personen, Organisationen und geografische Begriffe mit ihren Einträgen in Wikipedia und GND verknüpft. Sowohl bei der Layout- als auch bei der Entitätenerkennung wurde Neuland betreten, da nur teilweise auf kommerziell verfügbare automatisierte Verfahren zurückgegriffen werden konnte. Das Projekt wurde von Juni 2009 bis Mai 2013 realisiert. Die daran anschließende und im Frühjahr 2014 abgeschlossene projektergänzende NER verleiht zusätzlichen Mehrwert.

Rechtliche Regelungen

Da die DDR-Zeitungen noch nicht urheberrechtsfrei sind, mussten vor der eigentlichen Projektrealisierung fünf Verträge abgeschlossen werden: mit den drei Zeitungsverlagen zur Klärung der Nutzungsrechte und mit den Verwertungsgesellschaften VG Wort und VG Bild-Kunst als Vertreter der Text- und Bildautorenrechte. Wegen der schwierigen rechtlichen Situation und des Fehlens von Präzedenzfällen bei der digitalen Präsentation nicht urheberrechtsfreier Tageszeitungen gestalteten sich die juristischen Klärungen sehr aufwändig.

In den Verträgen mit den Verlagen wurde sowohl die Notwendigkeit einer Nutzeranmeldung als auch ein detailliertes Stufenverfahren für die Annahme, Bearbeitung und Klärung von Rechtsansprüchen seitens der Autoren vereinbart. Einsprüche können über ein Kontaktformular im Portal »DDR-Presse« online gemeldet werden. Nach einer juristischen Prüfung der Rechtmäßigkeit des Einspruchs werden gegebenenfalls sowohl Artikel als auch Autorennamen im Index gelöscht. Bisher hat jedoch nur ein einziger Autor von dieser Regelung Gebrauch gemacht.

Vorarbeiten sowie Ermittlung und Beschaffung von Ersatzausgaben

Mit der Zeitung »Neues Deutschland« (ND) begann der Aufbau des Portals »DDR-Presse«, da die Scans vom Verlag zur Verfügung gestellt werden konnten. Als Dienstleister für die OLR und OCR wurde das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) ausgewählt. Die MIK-Center GmbH Berlin übernahm das Scannen der Papieroriginale von »Berliner Zeitung« (BZ) und »Neue Zeit« (NZ) aus dem Bestand der Zeitungsabteilung der SBB.

Um Textverluste bei der OCR durch die Wölbung im Falz zu vermeiden, wurden die Zeitungsbände von BZ und NZ aufgetrennt, die Einzelseiten separiert und kartoniert. Dabei diente die Dokumentation von Fehlendem und Fehlerhaftem in Kollationierungslisten als Grundlage für die Beschaffung von Ersatzausgaben. Fehlerhaftes resultierte beispielsweise aus schlechter Papier- und Druckqualität insbesondere für den Zeitraum von 1945 bis 1955. Auch zerfallendes Papier oder Beschädigungen durch ausgerissenen bzw. ausgeschnittenen Text bzw. Verschmutzungen, Falten, Knitter, Wasserschäden und Klebeband verhindern die Weiterverarbeitung. Ein zusätzlicher Bedarf an Ersatzausgaben wurde während nachfolgender Workflowstationen ermittelt: während des Scannens, bei der stichprobenartigen Qualitätskontrolle als auch durch die maschinelle Kontrolle der Metadaten auf Vollständigkeit (fortlaufende Ausgaben- und Seitenzählung).

Ein eigens entwickelter Workflow regelte die mehrstufige Identifikation und Beschaffung von Ersatzexemplaren. Da letzteres organisatorisch wie zeitlich sehr aufwändig ist, wurde die Lückenergänzung auf zwei Einrichtungen beschränkt. Im Ergebnis liegen die BZ und NZ nahezu vollständig vor. Nichtbeschaffbares und später ermittelte Lücken werden im Portal dokumentiert.

Identifikation und Beschaffung von Ersatzexemplaren

Da beim »Neuen Deutschland« der Verlag die Scans zur Verfügung stellte, konnte der für die BZ und NZ entwickelte Workflow zur Ermittlung von Ersatzausgaben nicht greifen. Daher konnte Fehlendes beim ND erst bei der maschinellen Vollständigkeitsprüfung der Daten nach Abschluss der OLR/OCR sowie bei der nachfolgenden Qualitätskontrolle festgestellt werden. Fehlende Ausgaben und Seiten konnten teilweise ergänzt, auf fehlende Beilagen musste verzichtet werden. Insgesamt sind die Lücken und Metadatenfehler beim ND umfangreicher als bei den anderen Zeitungen, da Kontroll- und Steuermechanismen vor, während und nach dem Scanprozess nicht greifen konnten.

Scannen und Qualitätskontrolle

Eines der Auswahlkriterien für den Dienstleister war die OCR-Tauglichkeit der Scans. Das Scannen vom Original erfolgte gemäß der DFG-Praxisregeln »Digitalisierung« (DFG-Vordruck 12.151): Graustufenscan mit 256 Stufen pro Pixel, Auflösung 300 dpi, Ausgabeformate der Master: TIFF (LZW-komprimiert). Die Metadaten wurden im METS-ALTO-Format gespeichert. Beim Scannen kamen Großformat-Buchscanner für fragiles Papier und Großformat-Duplex-Scanner für stabile Zeitungsseiten zum Einsatz.

Da die während des Scannens als fehlerbehaftet eingeschätzten Ausgaben vom Dienstleister direkt in gesonderte Fehlerordner abgelegt wurden, konnten diese zu 100 % kontrolliert werden. Auf diese Weise konnte die Wahrscheinlichkeit erhöht werden, fehlerhafte Scans aus der Gesamtmenge zu ermitteln. Scans aus Reklamationen wurden ebenfalls zu 100 % kontrolliert. Die Qualitätskontrolle beinhaltete die Ermittlung von fehlenden Ausgaben, Seiten und Beilagen sowie von fehlendem bzw. unlesbarem Text als auch die Kontrolle der Metadaten z. B. auf Diskrepanzen bei Datumsangabe, Seitenzuordnung oder Seitenzählung.

Dokumentation von Fehlendem und Fehlerhaftem

Optische Layout- und Texterkennung

Der erste Schritt bei OLR und OCR bestand in der optischen Aufbereitung der Scans. Dies umfasste deren Freistellung, Rotationskorrektur, Kontrastoptimierung und Schärfung. Anschließend wurden die Scans bina-

risiert. Auf Grundlage einer von der SBB vorgegebenen Erfassungsanweisung erfolgte dann die automatische OLR mit Erkennung der Elemente und deren Klassifikation und Zuordnung zu den einzelnen Artikeln. Die Strukturierung erfolgte in drei Ebenen: Die erste Ebene umfasst die Kategorien, die zweite den Typ und die dritte die Label. In der ersten Ebene werden die beiden Kategorien »Nachricht« – die eigentlichen Artikel – und »Nichtredaktionelles« voneinander getrennt. Letzteres umfasst Listen, Tabellen und Vermischtes wie z. B. Sportergebnisse, Lottozahlen, Radio- und Fernsehprogramme, Wetterberichte, Rätsel, Inhaltsverzeichnisse, das Impressum sowie Familienanzeigen, Werbung und Kleinanzeigen. Für die Kategorie »Nichtredaktionelles« erfolgte keine tiefere Strukturierung. Die zweite Ebene, die sich nur auf die Kategorie »Nachricht« bezieht, beinhaltet den Typ »Text« für die Artikel und »Bild« für Fotografien, Grafiken und Diagramme. In der dritten Ebene erfolgt eine weitere Strukturierung, bei der für die Typen »Text« und »Bild« unterschiedliche Label vergeben werden. Die Label für den Typ »Text« enthalten verschiedene Titelarten wie z. B. Ober- und Untertitel sowie die »Verfasserzeile« mit Angabe von Autor, Verlag, Nachrichtenagentur u. ä. Für die Darstellung der Lesereihenfolge werden die Abschnitte jedes Artikels durchnummeriert. Die Label für den Typ »Bild« erfassen das eigentliche Bild und die Bildunterschrift.

Strukturierung in drei Ebenen

Nach der OLR wurden die OCR mittels ABBYY FineReader und die Zusammenführung von seitenübergreifenden Artikeln durchgeführt. Bei der OLR kamen komplexe geometrische und statistische Modelle zum Einsatz. So wurden beispielsweise über höhere Weißanteile im Schriftbild die Abgrenzungen zwischen den einzelnen Artikeln ermittelt. Das IAIS stellte das Softwaretool »Korrektor« sowohl zur Ansicht der OLR-Ergebnisse als auch zur Korrektur zur Verfügung (s. Abb. 2).

manuelle Nachbearbeitung

Nach Bearbeitung der ersten Jahrgänge zeichnete sich ab, dass mittels der maschinellen OLR nur eine unzureichende Genauigkeit erzielt wurde. Daraufhin fiel die Entscheidung, eine manuelle Nachbearbeitung durchführen zu lassen. Sie wurde von der Firma ArchivInForm (AI) Potsdam, Forschungspartnerin des IAIS und einzige Anwenderin der prototypischen Korrekturanwendung durchgeführt. Dabei musste jede einzelne zuvor maschinell bearbeitete Seite manuell kontrolliert und nachbearbeitet werden.

Abb. 2: Korrektor

Um die OLR-Software zu optimieren, wurde sie um regelbasierte Komponenten erweitert. Dafür wurden für einige Jahrgänge der BZ und NZ stichprobenartige Layout-Analysen durchgeführt, wobei u. a. Spaltenanzahl, die Struktur von Überschriften, abweichende Schrifttypen und Hinweise auf Artikelfortsetzungen dokumentiert wurden. Die Anpassung und das Training der regelbasierten OLR-Komponente für das jeweilige Layout verbesserten die Genauigkeit der maschinellen OLR und verringerten damit den Aufwand für die manuelle Nachbearbeitung. Dies könnte für ähnliche Projekte differenzierter ausgebaut werden.

Die Kombination von maschineller und manueller OLR führt zu sehr guten Ergebnissen, ist jedoch bei einer Massenverarbeitung sehr kostenintensiv. Bei diesem Projekt war der Seitenpreis für die manuelle Nachbearbeitung doppelt so hoch wie der für die maschinelle OLR. Eine effiziente maschinelle OLR würde Zeitungsdigitalisierungen bedeutend kostengünstiger gestalten.

Qualitätskontrolle der optischen Layout- und Texterkennung

Bei der Qualitätskontrolle der maschinell und manuell durchgeführten OLR lag die Stichprobenmenge anfangs bei 30 %. Die hohen OLR-Fehlerquoten einiger Jahrgänge zwangen zu deren Nachbearbeitung und zur Anpassung der Erfassungsanweisung. Nach Reduktion der Fehlerquote konnte der Stichprobenumfang auf 10 % und nach Stabilisierung der OLR-Qualität auf 7,5 % bzw. 24 Ausgaben pro Jahr gesenkt werden. Dies entspricht Prüfniveau I nach ISO 2859.

Zur Beschleunigung der sehr zeitintensiven Qualitätskontrolle wurde ein Fehlererfassungstool entwickelt, das einfach zu bedienen war und die Fehlerquote maschinell berechnete. Das Tool kam ab Jahrgang 1964 des ND und für BZ und NZ vollständig zum Einsatz. Unterstützend wirkte auch ein weiteres Tool für die zufällige Auswahl der zu prüfenden Zeitungsausgaben. Durch Einsatz der beiden Tools konnte der Zeitbedarf für die Qualitätskontrolle erheblich reduziert werden. Beim ND lag der Zeitbedarf vor Einsatz des Tools für einen Jahrgang bei 20 bis 24 Arbeitsstunden, während er sich danach und bei einem Stichprobenumfang von 7,5 % auf 7,5 bis 15 Stunden verringerte. Die Spanne bei der Bearbeitungsdauer resultiert aus der unterschiedlichen Seitenanzahl pro Ausgabe, die zwischen 4 bis 16 Seiten variiert.

Die häufigsten OLR-Fehler bezogen sich auf die falsche Zuordnung sowohl von Textblöcken zu Artikeln als auch auf die des Labels. Insgesamt konnte bei der Qualitätskontrolle eine sehr gute OLR-Genauigkeit von ca. 99 % festgestellt werden.

Zeitung	Genauigkeiten (in %)	
	OCR	OLR
ND	94,10	98,43
BZ	99,46	99,33
NZ	99,52	99,00

Abb. 3: Fehlergenauigkeiten

Die im Portal vorliegende OLR-Genauigkeit liegt noch über den angegebenen Werten, da die während der Qualitätskontrolle festgestellten Fehler sofort korrigiert wurden. Die OCR-Genauigkeit ist insgesamt gut mit Ausnahme der jeweils ersten zehn Jahrgänge. Dies ist dem schlechten Druck (zu stark, zu schwach, unsauber) bzw. Nicht- Antiqua-Schrifttypen insbesondere in den Überschriften geschuldet.

Präsentation und Nutzung

Da der aktuelle DFG-Viewer Zeitungen nicht in geeigneter Form präsentieren konnte, entwickelte die SBB für ZEFYS einen eigenen Zeitungsviewer. Im Januar 2012 wurden die ersten neun Jahrgänge des ND online gestellt. Weitere Jahrgänge und Zeitungen folgten bis zur vollständigen Präsentation im Mai 2013. Das Portal »DDR-Presse« umfasst:

eigener Zeitungsviewer

Neues Deutschland 1946–1990	135.020 Seiten
Berliner Zeitung 1945–1993	177.359 Seiten
Neue Zeit 1945–1994	135.655 Seiten

Eine sukzessive Erweiterung der Funktionalitäten bei der Präsentation fand während der gesamten Projektlaufzeit statt. Um eine offene Nutzung des Portals zu ermöglichen, werden verschiedene Anmeldeoptionen angeboten: über einen Bibliotheksausweis der SBB, für Angehörige von universitären bzw. außeruniversitären Forschungseinrichtungen über das DFN-Netz oder über einen Open-ID-Account bei xlogon.net. Darüber hinaus gibt es für einzelne Institutionen die Möglichkeit, auf Antrag IP-Adressen bzw. IP-Ranges für die Nutzung frei schalten zu lassen.

Fehlererfassungstool

Für die Recherche werden zwei Einstiege angeboten:

Rechercheeinstiege

- Eine Kalendernavigation, wobei bei der tagesbezogenen Anzeige im jeweiligen Jahreskalender alle vorhandenen Zeitungen bzw. Beilagen angezeigt werden.
- Eine Volltextsuche, wobei zwischen einfacher Suche mit Booleschen Verknüpfungen bzw. Phrasensuche sowie erweiterter Suche gewählt werden kann. Letztere ermöglicht die Recherche nach einzelnen Zeitungen, nach verschiedenen Struktur-

Abb. 4: Artikelanzeige

elementen wie Artikeltext, Bildunterschriften und flexiblen Datumsbereichen. Eine Facettierung gestattet die Eingrenzung der Treffermenge.

Image und Volltext

Bei der Volltextsuche zeigt die Trefferliste auch ein Vorschaubild der Zeitungsseite an, in dem der betreffende Artikel markiert ist. Dies ermöglicht eine Vorauswahl unter Berücksichtigung des Artikelumfangs. Die Trefferliste ist nach weiteren Kriterien wie Zeitungstitel, Jahr, Monat und Seitenangabe facettiert. So können – quasi als Nebenprodukt – anhand der Trefferzahlen der einzelnen Jahre zeitliche Verläufe bzw. das Auftreten und Verschwinden von Begriffen und damit politische, gesellschaftliche oder wirtschaftliche Entwicklungen verfolgt werden. Beispielsweise werden die auf der 9. Tagung des SED-Zentralkomitees geplanten Großforschungszentren im Jahr 1968 in 15 Treffern genannt, die bis 1970 auf 113 anwachsen und nach 1971 versiegen aufgrund der Nichtfinanzierbarkeit dieser Zentren. Auch der Trefferverlauf bei der Recherche nach »Kinderkrippe« gibt Hinweise auf die Bewertung der Kleinkindbetreuung und damit über die Berufstätigkeit der Frauen: Der Trefferverlauf bewegt sich vom einstelligen Bereich in den Jahren 1945 bis 1948 bis zum dreistelligen Bereich von 1952 bis 1990.

Kontaktformular

Nach Auswahl eines Treffers werden Image und Volltext des Artikels sowie weiterführende Inhalte angezeigt (s. Abb. 4). Wenn die Textverständlichkeit durch OCR-bedingte Fehler in der Volltextansicht eingeschränkt ist, kann dies durch das ebenfalls sichtbare Faksimile kompensiert werden. Für die Imageanzeige stehen eine stufenlose Zoomfunktion, die freie Bildpositionierung und die Möglichkeit der Bilddrehung zur Verfügung. Es kann nicht nur innerhalb der Ausgabe geblättert, sondern auch zur Vorgänger- oder Nachfolger-Ausgabe oder zu den Ausgaben der anderen Zeitungstitel navigiert werden. Sowohl einzelne Seiten als auch komplette Ausgaben sind als PDF-Datei herunterladbar. Nutzer können über ein Kontaktformular Fragen, Kommentare oder Anregungen senden sowie Fehler bei der Artikelsegmentierung melden, wobei die Fehlermeldung die PURL des Artikels bzw. der Seite referenziert.

Bei den weiterführenden Inhalten wird verlinkt auf:

- die Inhalte der wissenschaftlichen Forschungsumgebung des ZZF (Portal »Presse in der DDR«),
- biographische Angaben aus den Datenbanken »Wer war wer in der DDR« und »Deutsche Kommunisten: Biographisches Handbuch 1918 bis 1945«,
- Personen, Orte und Organisationen und ihre Einträge in Wikipedia und GND.

Es wird eine Linkliste zu relevanten Institutionen, Archiven, Zeitschriften sowie zu allgemeinen, thematischen und Biografien-Portalen angeboten.

Die Nutzung des Portals hat sich nach hohen Zugriffszahlen infolge des Medienechos bei Abschluss des Projekts bei einer stabilen Nutzung eingepegelt: täglich wird auf ca. 1.700 Seiten zugegriffen, wobei die drei Zeitungen unterschiedlich stark genutzt werden (s. Abb. 5). Insgesamt wurden bisher ca. 2,15 Millionen Seiten aufgerufen (Stand Mitte November 2015). Bei den Nutzerzahlen lassen sich nur die xlogon-Nutzer statistisch erfassen, wobei Mitte November 2015 fast 6.000 Nutzer angemeldet waren. An der SBB rangiert ZEFYS bei der Nutzung aller Online-Angebote hinter dem StabiKat an dritter Stelle, wobei der Anteil des »DDR-Presse«-Portals bei ca. 25 % liegt.

	Durchschnittliche tägliche Seitenzugriffe	Gesamte Seitenzugriffe (auf Tausend gerundet)
ND	826	1.248.000
BZ	572	609.000
NZ	301	290.000

Abb. 5: Nutzungsstatistik. Stand 19. November 2015

Technische Details der Präsentation und Langzeit-Archivierung

Da das Dokument-Management-System Goobi noch nicht für Zeitungen zur Verfügung stand, erfolgten die Realisierung von Workflow und Präsentation unabhängig von Goobi. Die bestehenden Daten werden nachträglich in die Goobi-Umgebung überführt werden. Zum Zweck der Langzeit-Archivierung werden die originalen Graustufen-TIFFs und die XML-Dateien (METS/ALTO) in einem replizierten Speicherbereich gehalten.

Named Entity Recognition (NER)

Eine maschinelle Entitätenerkennung von Personen, geografischen Namen, Organisationen und Abkürzungen wurde von der Firma IntraFind München durchgeführt. Um die Qualität der NER zu verbessern, stellten SBB und ZZF Listen mit Personennamen und DDR-Organisationen zur Verfügung. Insgesamt liegen die Ergebnisse leider unter den aufgrund der Testergebnisse erwarteten. Dies betrifft sowohl den Anteil der erkannten Entitäten im Verhältnis zu den nicht erkannten als auch die Erkennung von Organisationen und Personen. Bei den Organisationen konnten vor allem die 800 aus der gelieferten Liste verlässlich erkannt werden. Die erkannten Entitäten wurden automatisch mit den entsprechenden Eintragungen in Wikipedia und GND verknüpft.

Beiträge des Zentrums für Zeithistorische Forschung

Begleitend zur Präsentation der Zeitungsbestände hat das Zentrum für Zeithistorische Forschung Potsdam (ZZF) eine wissenschaftliche Forschungsumgebung erstellt. Diese enthält zeithistorische Fachartikel zur Geschichte der einzelnen Zeitungen, zu Aspekten der Pressegeschichte der DDR wie die politische Steuerung der Printmedien, zu Kontrolle und Zensur, zur Rolle der Journalisten und zur Bildpolitik. Zusätzlich kann auf ca. 185 kurze Glossar-Texte zu Begriffen der DDR-Geschichte, Ausschnitte aus »Erinnerungen« wichtiger Zeitzeugen sowie Dokumente zur Pressepolitik zugegriffen werden. Die Bereitstellung der Inhalte erfolgt im Open Access.

wissenschaftliche Forschungsumgebung

Die Beiträge werden auf einer eigens für das Projekt eingerichteten MediaWiki/SemanticMediaWiki (MW/SMW) Plattform veröffentlicht. Dafür konnten teilweise die für das Projekt Docupedia-Zeitgeschichte entwickelten Publikationsformate und Wiki-Funktionen nachgenutzt werden durch eine Weiterentwicklung und Ergänzung um neue Datenschemata. Ziel war die konsistente Bereitstellung aller Inhalte wie Texte und Sachbegriffe über die RDF-Schnittstelle der MW/SMW Installation.

Langzeit-Archivierung

Eine erste Umsetzung erfolgte dann als RSS-Feed, der sowohl einen definierten Schlagwortbegriff als auch einen Verweis auf den Inhalt in der Forschungsumgebung enthielt. Die Schlagwortbegriffe wurden an der SBB in die Darstellung der Digitalisate integriert. Auf Seiten des ZZF wurde ein Eingabeformular zur Erstellung und Änderung der Suchbegriffe entwickelt.

Bei der »semantischen« Verzahnung der Angebote ist zu berücksichtigen, dass dies nicht als automatisierter Prozess zu realisieren ist, der auf der fachwissenschaftlichen Seite »Themen« als Suchwortfelder definiert, die dann zur Verlinkung der Angebote im Präsentationssystem führen. Dies scheitert an der Inkongruenz zwischen dem heutigen Themenfeld und dem Sprachgebrauch der DDR-Presse. Möglich ist dies nur bei der Bezeichnung eindeutiger Entitäten wie

in Wikipedia und GND verknüpft

WASSERZEICHEN –
SCHREIBER – PROVENIENZEN
Neue Methoden zur Erforschung und Erschließung von Kulturgut im digitalen Zeitalter: zwischen wissenschaftlicher Spezialdisziplin und »catalog enrichment«
Hrsg. von Wolfgang Eckhardt, Julia Neumann, Tobias Schwinger und Alexander Staub
2016. 322 Seiten, gebunden, Fadenheftung
ISBN 978-3-465-04257-0
ZfBB Sonderband 118
Auch als E-Book erhältlich

Die Erforschung der musikalischen Quellen gehört zu den Grundlagen des Faches Musikwissenschaft. Aufschlussreich für alle ihre Gebiete ist eine erweiterte Datenbasis mit tief erschlossenen Musikquellen. Mit dem Pilotprojekt »Kompetenzzentrum Forschung und Information Musik« (KoFIM) an der Staatsbibliothek zu Berlin – Stiftung Preußischer Kulturbesitz soll jetzt die musikalische Quellenforschung durch eine Bibliothek mit großem Musikalienbestand entscheidend vorangebracht werden. Dabei sollen neue Methoden erprobt und entsprechende Geschäftsgänge für die Erschließung entwickelt werden. Das Projekt dient der Tiefenerschließung von bislang noch nicht ausreichend katalogisierten Musikhandschriftenbeständen und der Etablierung eines *catalog enrichment* auf dem Gebiet der digitalen Dokumentation von Schreiberhänden und Wasserzeichen. Im Rahmen dieses von der DFG geförderten Projekts fand im Oktober 2014 an der Staatsbibliothek zu Berlin ein wissenschaftliches Kolloquium statt, das seitens des Fachpublikums auf großes Interesse stieß. Dieser ZfBB-Sonderband dokumentiert die anlässlich des Kolloquiums entstandenen Beiträge.

VITTORIO KLOSTERMANN

Personen, Objekte und Organisationen, die auch als Suchbegriffe in einem zeitgenössischen Text eindeutig sind. Die Organisation einer solchen inhaltlichen »Verzahnung« von zwei Web-Angeboten beinhaltet eine laufende Kontrolle der Ergebnisse der »Verlinkung«.

Die ZZF-Forschungsumgebung ist als eigenständiges Portal über eine Suchfunktion sowie Sachbegriffe und Personennamen erschlossen.[26] Personennamen und Suchbegriffe des ZZF-Angebots werden außerdem zur Verlinkung aus der Präsentation der Zeitungen an der SBB verwendet. Die Online-Redaktion des Instituts verstetigte das Angebot, entwickelte es technisch weiter und ergänzte es durch weitere Beiträge.

... und weiter zu den Digital Humanities

Um in den Digital Humanities repräsentative Ergebnisse zu erhalten, werden große Textkorpora mit mindestens 5 Millionen Textwörtern benötigt. Das Zeitungskorpus der drei DDR-Zeitungen umfasst über 1 Milliarde und dient im Projekt »Kuration des ›DDR-Presseportals‹ und Evaluierung der CLARIN-D-Services als Grundlage für die zeithistorische Forschung« als Forschungsobjekt für computerlinguistische Analysen. Dieses Projekt wird derzeit in Kooperation mit der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) als CLARIN-D-Zentrum, dem Institut für Geschichtswissenschaften, Bereich Historische Fachinformatik der Humboldt-Universität Berlin (HU), und dem Zentrum für Zeithistorische Forschung Potsdam (ZZF) realisiert und soll im Frühjahr 2016 abgeschlossen sein.

Innerhalb von CLARIN (Common Language Resources and Technology Infrastructure), einem europaweiten Netzwerk, entsteht eine nachhaltige webbasierte Forschungsinfrastruktur insbesondere für die Geistes- und Sozialwissenschaften. CLARIN-D[27], das deutsche Teilprojekt, verfügt über computerlinguistische Tools, die für verschiedene Anwendungsgebiete evaluiert werden sollen. Die CLARIN-Facharbeitsgruppe 10 »Zeitgeschichte« wählte dafür die Textkorpora der drei DDR-Zeitungen. Um die Daten für computerlinguistische Untersuchungen verwendbar zu machen, werden sie von der HU und der BBAW in das TEI-XML-Format überführt und am CLARIN-Servicezentrum im »Digitalen Wörterbuch der Deutschen Sprache« (DWDS) in die Zeitungskorpora des 20. Jahrhunderts integriert. Eine systematische Auswertung des Nutzens des CLARIN-D-Tools findet im Rahmen des ZZF-Forschungsschwerpunktes zur »Historischen Semantik des Politischen im 20. Jahrhundert« statt. Es werden Wortfrequenzen, Worthäufigkeiten, Neologismen und Wortprofile der DDR-Sprache in ihren Entwicklungen untersucht. Die ebenfalls im DWDS enthaltene

bundesdeutsche Wochenzeitung »Die Zeit« ermöglicht darüber hinausgehende Analysen zum Vergleich von Ost- und Westsprache sowie zu Sprachtransfers. Dabei verspricht die computerlinguistische Analyse von synchronen Zeitungskorpora aus verschiedenen politischen Kontexten interessante Forschungsergebnisse.

In einem weiteren geplanten Projekt soll für vergleichende computerlinguistische Analysen der in den DDR-Zeitungen verwendeten Herrschaftssprache die Sprache des politisch autonomen Denkens gegenübergestellt werden. Diese in Opposition und Widerstand der DDR verwendete Sprache ist im Korpus des »Politischen DDR-Zeitschriftensamisdat« dokumentiert. Dieses Korpus entstand im Rahmen eines von der Bundesstiftung Aufarbeitung der SED-Diktatur und der DFG geförderten Projekts, bei dem unter Leitung der Umweltbibliothek Großhennersdorf acht Aufarbeitungsinitiativen 152 Zeitschriften bzw. Einzelausgaben mit fast 13.000 Seiten digitalisiert und volltexterschlossen haben. Das Mitteleuropazentrum für Staats-, Wirtschafts- und Kulturwissenschaften der Technischen Universität Dresden war Kooperationspartner und stellt auch die Hard- und Software-Plattform zur Verfügung.[28]

Diese Projekte können als prototypische Beispiele für das engere Zusammenrücken von Geisteswissenschaften, Informatik und Bibliotheken gewertet werden, wie auch Meister und Veit in ihrem Aufsatz »Digital Humanities – Neue Netzwerke für die Geisteswissenschaften« ausführen: »So wie Geisteswissenschaften und Informatik rücken künftig auch Geisteswissenschaften und Bibliotheken mit Kontakt ›auf Augenhöhe‹ enger zusammen bzw. wird das ›Ermöglichen‹ von Forschung zum aktiven Bestandteil von Forschungsprozessen.«[29] Bibliotheken und Archive sollten diese Möglichkeiten nutzen und in engem Kontakt mit den Netzwerken der Digital Humanities ihre Schätze heben.

computerlinguistische Analyse

»Politischer DDR-Zeitschriftensamisdat«

1 Nicholson, Bob: The digital turn, In: Media History 19 (2013) No. 1, S. 59–73, S. 63/64. Doi:http://dx.doi.org/10.1080/13688804.2012.752963

2 Bingham, Adrian: The Digitization of Newspaper Archives – Opportunities and Challenges for Historians, In: Twentieth Century British History, vol. 21 (2010) No. 2, S. 228. Doi:10.1093/tcbh/hwq007

3 Moretti, Franco: Conjectures on World Literature, In: New Left Review, 1 (2000), S. 57–58; Moretti, Franco: Graphs, Maps, Trees – Abstract Models for a Literary Theory. London: Verso, 2005.

4 http://litlab.stanford.edu/current-projects/

5 Bingham, Adrian: ebenda, S. 229.

6 Michel, Jean-Baptiste et al.: Quantitative Analysis of Culture Using Millions of Digitized Books, In: Science 14, Jan 2011, vol. 331, no. 6014, pp 176–182. Doi:10.1126/science.1199644

7 www.theeuropeanlibrary.org/tel4/newspapers

8 www.proquest.com/products-services/pq-hist-news.html

9 www.readex.com/content/americas-historical-newspapers

10 http://gdc.gale.com/products-by-medium/

11 https://natlib.govt.nz/collections/a-z/papers-past

12 www.delpher.nl/nl/tijdschriften

13 http://hemerotecadigital.bne.es/index.vm

14 http://gallica.bnf.fr/html/presse-et-revues/presse-et-revues

15 www.loc.gov/ndnp/ ; http://chroniclingamerica.loc.gov

16 www.britishnewspaperarchive.co.uk/

17 http://trove.nla.gov.au/newspaper

18 http://anno.onb.ac.at/index.htm

19 DIGI – National Library's Digital Collections: http://digi.kansalliskirjasto.fi/?language=fi

20 http://timarit.is/

21 Siehe auch Seiderer, Birgit: Die Digitalisierung von Zeitungen im deutschsprachigen Raum – ein Zustandsbericht, In: ZfBB 57 (2010) Nr. 3–4, S. 165–171 und Hagenah, Ulrich: Retrodigitalisierung von Zeitungen durch Regionalbibliotheken – Gedanken zu einer Momentaufnahme, in: ZfBB 57 (2010) Nr. 3–4, S. 183–189.

22 http://gepris.dfg.de/gepris/projekt/227800404

23 http://sammlungen.ub.uni-frankfurt.de/cm/nav/index/all

24 http://zefys.staatsbibliothek-berlin.de/amtspresse/

25 http://zefys.staatsbibliothek-berlin.de/ddr-presse

26 http://pressegeschichte.docupedia.de/

27 www.clarin-d.de/de/

28 www.ddr-samisdat.de

29 Meister, Jan Christoph; Veit, Joachim: Digital Humanities – Neue Netzwerke für die Geisteswissenschaften, In: ZfBB 61 (2014) Nr. 4–5, S. 266. Doi: http://dx.doi.org/10.3196/18642950146145174

Die Verfasserin

Almut Ilsen, Projektleiterin DDR-Zeitungsportal bis 2014, Fachreferentin für Chemie, Physik und Astronomie bis 2016, Benutzungsabteilung Referat Wissenschaftliche Dienste (II D 2), Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Potsdamer Straße 33, 10785 Berlin,
E-Mail: almut-ilsen@gmx.de