Das Projekt „Vorwärts bis 1933“: Digitalisierung und elektronische Präsentation einer historischen Zeitung – Ein Werkstatt-Report

Teil 2: Präsentation der Zeitung im Web

Olaf Guercke

Der vorliegende Text ist der zweite und letzte Teil einer Arbeit, die aus Sicht des Praktikers die Digitalisierung einer historischen Zeitung von der Papiervorlage bis zur im Volltext durchsuchbaren Web-Präsentation beschreibt. Das Projekt „Vorwärts bis 1933“ wird dabei in Form eines Werkstatt-Reports in einem recht fortgeschrittenen, jedoch noch nicht abgeschlossenen Stadium umfassend illustriert. Teil 1 (b.i.t online 2016,6) bot zunächst eine Einleitung mit Informationen über den Gegenstand der Digitalisierung und die Ziele sowie den derzeitigen Stand des Projekts. Anschließend wurden die verschiedenen Aspekte des Scanprozesses, der maschinellen Texterkennung und der Metadaten-Anreicherung in Augenschein genommen. Mit dem vorliegenden Text folgt nun ein zweiter Teil, der sich mit der Präsentation der Zeitung im Web und den im Projekt entwickelten Suchfunktionalitäten beschäftigt. Zur Erleichterung des Einstiegs wurde die Einleitung aus Teil 1 leicht verändert in den 2. Teil übernommen.
Der Verfasser versucht, kritische Blicke auf die eigene Arbeit zu werfen, so dass der Text für andere Digitalisierungs-Praktiker im Hinblick auf ihre Projekt-Strategien möglichst aufschlussreich sein kann.

Einleitung

Der „Vorwärts – Berliner Volksblatt“ ist seit seinem Bestehen das zentrale Presseorgan der deutschen Sozialdemokratie. Als bis zu zwei Mal täglich erschienenes Periodikum bietet er für die Zeit des deutschen Kaiserreichs und der Weimarer Republik eine Fülle von Quellenmaterial, welches jedoch gegenwärtig nur schwer zugänglich ist. Relevant ist der Vorwärts vor allem für historische Forschung, die sich mit der Arbeiterbewegung innerhalb und außerhalb der SPD und den mit ihr verbundenen politischen Auseinandersetzungen befasst. Darüber hinaus ist die Zeitung eine Fundgrube für die kulturwissenschaftliche Forschung, die sich auf die Arbeiter-Milieus in dieser Zeit bezieht, für die Lehre in Schule und Universität, für die biografische Literatur, für die Ahnenforschung und für historisch interessierte Bürgerinnen und Bürger. Das mögliche Themenspektrum reicht hier, um nur zwei Beispiele zu nennen, von detaillierten Texten aus den Anfängen der proletarischen Frauenbewegung bis zu Kleinanzeigen, die überraschende Aufschlüsse über das tägliche Leben in den proletarischen Milieus des späten 19. und frühen 20. Jahrhunderts geben können.

Das Ziel des Projekts „Vorwärts bis 1933“, das seit Januar 2015 in der Bibliothek der Friedrich-Ebert-Stiftung Bonn verwirklicht wird, ist es, sämtliche Ausgaben des „Vorwärts“ von dessen Gründung 1876 bis zum Verbot im Februar 1933 in hoher Qualität zu digitalisieren und der Öffentlichkeit in einer mittels OCR durchsuchbaren Web-Präsentation zur Verfügung zu stellen. Insgesamt werden ca. 200.000 Zeitungsseiten vom Papier-Original digitalisiert, die sich auf ca. 19.000 Ausgaben verteilen. Die vorliegende Arbeit ist der zweite Teil eines Werkstatt-Berichtes, in dem das seit Anfang 2015 laufende und bis Ende 2017 datierte Projekt in all seinen Aspekten und Entwicklungsperspektiven beschrieben wird.

Die Phase, in der sich das Projekt derzeit befindet, lässt sich folgendermaßen skizzieren: Der Produktionsprozess von der in Folianten vorliegenden Zeitungsseite zum fertigen Scan läuft kontinuierlich im Rahmen eines funktionierenden Workflows. Bisher sind ca. 120.000 Zeitungsseiten mit Hilfe von zuvor

angeschaffter Technik bei uns im Hause digitalisiert worden. Seit August 2016 werden die Zeitungsseiten mittels der Software BCS2 Professional sowie einer ABBYY-OCR-Engine weiter verarbeitet und mit Metadaten sowie zonalen OCR-Daten[1] angereichert. Ergebnis dieses mittlerweile etablierten Workflows sind Daten-Container, die von der Präsentations-Software MyBib-eL[2] zur Darstellung der Web-Präsentation verwendet werden können. Während der bereits erschienene Teil 1 der Arbeit sich mit der Produktion von Scans und Daten-Containern beschäftigt, beschreibt der hier vorliegende zweite Teil die derzeit laufende Entwicklung der Web-Präsentation mit ihren Suchfunktionalitäten in Zusammenarbeit mit der Herstellerfirma ImageWare Components GmbH (IWC)[3]. Der hierzu aufgesetzte Pilot-Lesesaal befindet sich auf einem Server des Hochschulbibliothekszentrums NRW (hbz), wo auch die als Projektziel angestrebte öffentliche Präsentation des Vorwärts gehostet werden wird. Zum gegenwärtigen Zeitpunkt hoffen wir, ab Ende März 2017 einen Teilbestand des Vorwärts öffentlich präsentieren zu können.

Anforderungen verschiedener Nutzergruppen an ein Zeitungsdigitalisierungsprojekt

Es liegt nahe, Überlegungen zur Web-Präsentation einer digitalisierten Zeitung mit einem detaillierten Blick auf die Bedürfnisse der verschiedenen Nutzergruppen zu beginnen, die mit dem Angebot angesprochen werden sollen. Da im Laufe des Vorwärts-Projekts fortlaufend zahlreiche Anfragen von Nutzern eingehen, aus deren Inhalt sich Rückschlüsse bezüglich dieser Bedürfnisse ziehen lassen, haben wir uns bei der Entwicklung der Suchfunktionalitäten an diesen Informationen orientiert.

Etwa die Hälfte der eingehenden Anfragen kommen aus dem universitären Bereich, wobei es sich häufig um Studierende oder Doktoranden der Fächer Geschichte oder Politikwissenschaften handelt, die den Vorwärts im Rahmen der in der Regel eng umgrenzten Themen ihrer Bachelor- und Masterarbeiten oder Dissertationen zu Rate ziehen möchten. Von Bedeutung ist auch die Gruppe der etablierten Wissenschaftler, die in ihrer Forschungsarbeit, etwa beim Verfassen umfangreicher Monografien, auf den Vorwärts angewiesen sind. Auch in der Lehre zeigt sich ein Bedarf an der Quelle Vorwärts, wobei hier neben der Konzeption universitärer Seminare auch die Suche nach Materialien für den Geschichtsunterricht an Schulen eine große Rolle spielt. Darüber hinaus bekommen wir Anfragen von Museen, denen wir bei der Konzeption von Ausstellungen behilflich sein konnten sowie aus dem Bereich der Erwachsenenbildung. Schließlich sind Nutzer zu erwähnen, die eine auf persönlichem Interesse basierende Forschung zu bestimmten Themen oder Personen betreiben.

Abbildung 1:
Zeitlich eingegrenzte Phrasensuche nach „Kurt Singer“ mit Trefferliste

Die Anfragen aus all diesen Bereichen lassen sich wie folgt gruppieren:

a) Anfragen nach bestimmten Personen oder Institutionen, die mutmaßlich im Vorwärts vorkommen, wobei nicht oder nur ungefähr bekannt ist, wo genau sie zu finden sind.

- **Beispiele:**

Eine Anfrage nach mehreren weniger bekannten Berliner Kabarettgruppen aus der Endphase der Weimarer Republik.

Eine Anfrage nach Artikeln eines bestimmten Autoren, die dieser über einen Zeitraum von 10 bis 12 Jahren hinweg im Vorwärts veröffentlicht haben könnte.

1 Es handelt sich um von der OCR-Engine erkannten Text, dem die jeweiligen Koordinaten der einzelnen Wörter auf dem Image beigefügt werden. Diese Daten sind die Grundlage für die farbliche Hervorhebung von Suchtreffern in der Präsentation.

2 Vgl.: http://www.imageware.de/produkte/mybib-el/ [17.03.2016]

3 Die Zusammenarbeit findet auf der Basis eines kooperativen Forschungsprojekts statt. In dessen Rahmen wird u. a. kontinuierlich die verwendete Software im Hinblick auf die speziellen Anforderungen von Zeitungsdigitalisierungen weiter entwickelt.

Abbildung 2: Ansicht des ersten Treffers auf der Ausgabenebene

• **Anforderungen:**

Eine OCR-basierte Stichwortsuche über den Volltext. Ohne diese Möglichkeit sind solche Anfragen praktisch nicht zu beantworten, da sie, je nach der Größe des Zeitfensters, das Durchsuchen tausender Zeitungsseiten erfordern würden. Eine unscharfe Suche sollte möglich sein, um auch bei OCR-Fehlern Treffer erzielen zu können.

b) Anfragen nach Artikeln zu bestimmten Themenbereichen oder Personen im Zusammenhang mit historischen Ereignissen, die sich zeitlich gut eingrenzen lassen.

• **Beispiele:**

Eine Anfrage nach Artikeln im Zusammenhang mit dem Spartakus-Aufstand, dem KPD-Aufstand und dem Kapp-Putsch, die sich auf Ereignisse in der Stadt Halle beziehen.

Eine Anfrage nach Artikeln (Prozessberichte etc.) über verschiedene Serienmörder der Weimarer Republik.

• **Anforderungen:**

Hier ist zusätzlich zur Stichwortsuche eine Möglichkeit der zeitlichen Eingrenzung von Bedeutung, so dass ganz bestimmte Zeiträume mit Stichworten durchsucht werden können. Notwendig ist außerdem eine Kalenderfunktion, mit der man die Ausgaben bestimmter Tage und Wochen ohne Stichwortsuche finden und durchsuchen kann.

c) Verifizierung von Zitationen in wissenschaftlichen Arbeiten, die aufgrund der schlechten Zugänglichkeit der Zeitung in der Vergangenheit bisweilen widersprüchlich oder ungenau sind.

• **Anforderungen:**

Hier ist zunächst ein Zugang über Jahrgang und Datum wichtig. Wird das Zitat am angegebenen Ort nicht vorgefunden, kann eine zeitlich begrenzte Stichwortsuche zum Einsatz kommen.

Suchfunktionalitäten der Vorwärts-Präsentation

Wir haben versucht, die Funktionalitäten der Vorwärts-Präsentation im Rahmen der vorhandenen Möglichkeiten so zu gestalten, dass sie den oben beschriebenen Anforderungen genügen und den Nutzern die Möglichkeit eröffnen, sich die Quelle nach ihren Bedürfnissen selbständig zu erschließen. Es folgt eine kurze Beschreibung der Funktionalitäten:

• **Einfache Stichwortsuche:**

Das Web-Frontend des Vorwärts-Projekts bietet zunächst eine so genannte Einschlitz-Suche, mit der der gesamte Bestand anhand von Stichworten durchsucht werden kann. Möglich sind hier auch eine Phrasensuche, bei der in Anführungszeichen gesetzte Wortfolgen gefunden werden und der Einsatz von Variabeln (* für mehrere Zeichen, ? für ein Zeichen). Die Eingabe mehrerer Suchworte hintereinander führt zur Anzeige von Seiten, auf denen beide Suchworte vorkommen.

• **Erweiterte Suche mit Kalenderfunktion:**

Die erweiterte Suche bietet das Feld „Titel“, in der die Suche auf bestimmte selbständige Beilagen des Vorwärts eingegrenzt werden kann. Wichtiger sind die beiden mit Kalendern versehenen Felder „Datum von“ und „Datum bis“ die zum einen eine Stichwortsuche auf einen tagesgenauen Zeitraum eingrenzen können und zum anderen ein Navigationsinstrument darstellen, mit dem der Bestand auch ohne Stichwortsuche erschlossen werden kann.

• **Verarbeitung von Treffern:**

Jede erfolgreiche Suche ergibt zunächst eine Liste von Ausgaben, in denen Treffer vorkommen. Diese Liste kann nach Datum und nach Relevanz sortiert werden. Neben den Ausgaben wird für den jeweils ersten Treffer der OCR-Text in der unmittelbaren Umgebung angezeigt, um eine erste Orientierung zu ermöglichen. Ein Klick auf die einzelnen Ausgaben bietet die Möglichkeit, sich unter dem Menüpunkt „Treffer im Dokument“ die farblich hervorgehobenen Treffer auf den Einzelseiten anzuschauen. Hinzu kommen verschiedene Downloadmöglichkeiten einzelner Seiten, Ausgaben oder des reinen OCR-Texts. Außerdem werden Direktlinks zu Ausgaben und Einzelseiten angeboten.

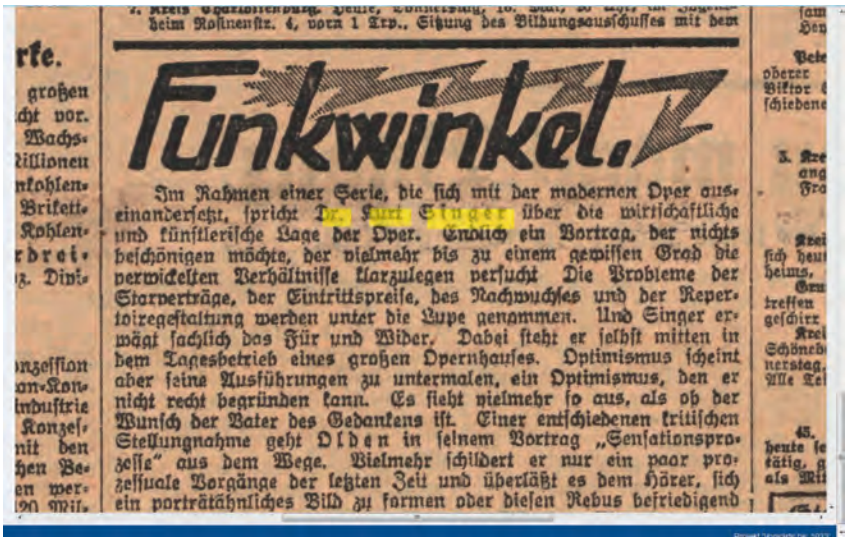
• **Umgang mit fehlerhafter Texterkennung:**

Die Qualität der Texterkennung im Vorwärts-Projekt ist zwar für eine in Fraktur gedruckte Zeitung recht gut, jedoch nicht perfekt. Es besteht immer die Möglichkeit, dass eigentlich relevante Treffer nicht gefunden werden, weil einzelne Buchstaben von der Software falsch erkannt wurden. Hier schafft die Verwendung der Platzhalter * und ? Abhilfe, mit denen eine unscharfe Suche durchgeführt werden kann. Besonders bei sehr langen Suchbegriffen oder den in Fraktur einander sehr ähnelnden Buchstaben f und s empfiehlt sich diese Methode:

Gesucht: Donaudampfschiffartskapitänspatent; Suchwort: Donauda*
Gesucht: Jakob Wassermann; Suchwort: Wa??ermann
Das Web-Frontend des Projekts befindet sich zum gegenwärtigen Zeitpunkt in einer weit fortgeschrittenen Phase der Entwicklung. Die meisten der hier beschriebenen Funktionalitäten sind bereits umgesetzt, dennoch handelt es sich um ein „Work in Progress“, so dass hinsichtlich der bald zu erwartenden Veröffentlichung noch ein gewisser Vorbehalt besteht.

Umgang mit urheberrechtlich relevantem Material

Ein Problem bei Zeitungsdigitalisierungen aus dem frühen 20. Jahrhundert stellt das urheberrechtlich relevante Material dar, welches die Zeitungen in der Regel enthalten. Als Beispiel wären hier etwa Fortsetzungsromane zu nennen, die von nach 1947 verstorbenen Autoren stammen und teils noch heute im Handel erhältlich sind. Hier bietet der MyBib-eL die Möglichkeit, im Rahmen seines Rechtemanagements einzelne Seiten innerhalb von Ausgaben zu sperren. Im Projekt nutzen wir diese Möglichkeit, um etwa oben erwähnte Fortsetzungsromane von der Publikation auszuschließen. Der Nutzer erhält anstelle der betreffenden Seite einen Sperrvermerk mit einer Kontaktadresse, so dass im Einzelfall geklärt werden kann, ob und zu welchen Zwecken die gewünschte Seite zur Verfügung gestellt werden kann.
Von Vorteil ist hierbei, dass die Seiten flexibel gesperrt und – etwa wenn Inhalte im Laufe der Zeit gemeinfrei werden – wieder entsperrt werden können und dass die Notwendigkeit der Manipulation am Image selbst (z.B. Schwärzung von Inhalten) entfällt. Von Nachteil ist, dass bisher nur auf Seitenebene gesperrt werden kann. Daher sind häufig auch unproblematische Inhalte von der Sperrung betroffen. Verbesserungspotenzial besteht auch beim recht aufwändigen Workflow der Seitensperrung im Admin-Bereich des MyBib eL. Konkret erfordert die Sperrung von 50 Einzelseiten aus verschiedenen Ausgaben einen Arbeitsaufwand von ca. 1 Stunde hochkonzentrierter und sehr repetitiver Arbeit. Insgesamt gesehen bietet die Funktion jedoch eine gute Möglichkeit, mit den Anforderungen retrodigitaler Publikationen an das Urheberrecht umzugehen.



Funkwinkel.

Im Rahmen einer Serie, die sich mit der modernen Oper auseinandersetzt, spricht Dr. Kurt Singer über die wirtschaftliche und künstlerische Lage der Oper. Endlich ein Vortrag, der nichts beschönigen möchte, der vielmehr bis zu einem gewissen Grad die verwickelten Verhältnisse klarzulegen versucht. Die Probleme der Starverträge, der Eintrittspreise, des Nachwuchses und der Repertoiregestaltung werden unter die Lupe genommen. Und Singer erwägt sachlich das Für und Wider. Dabei steht er selbst mitten in dem Tagesbetrieb eines großen Opernhauses. Optimismus scheint aber seine Ausführungen zu untermalen, ein Optimismus, den er nicht recht begründen kann. Es sieht vielmehr so aus, als ob der Wunsch der Vater des Gedankens ist. Einer entschiedenen kritischen Stellungnahme geht Olden in seinem Vortrag „Sensationsprozesse“ aus dem Wege. Vielmehr schildert er nur ein paar prozessuale Vorgänge der letzten Zeit und überläßt es dem Hörer, sich ein porträtähnliches Bild zu formen oder diesen Rebus befriedigend

Abbildung 3: Gefundener Artikel über Kurt Singer mit farblich hinterlegten Suchwörtern

Fazit

Das Vorwärts-Projekt steht nun kurz vor der Veröffentlichung eines ersten großen Datenbestandes, der die Jahrgänge 1924 bis 1933 umfasst und einen Umfang von ca. 55.000 Seiten hat. Wir hoffen, mit diesem Meilenstein unserem Ziel, die reichhaltige historische Quelle bis Ende des Jahres in Gänze einer möglichst breiten wissenschaftlichen und gesellschaftlichen Öffentlichkeit auf einfache Weise zugänglich zu machen, einen großen Schritt näher zu kommen.

Olaf Guercke
Bibliothek der
Friedrich-Ebert-Stiftung
Godesberger Allee 149
D-53175 Bonn
olaf.guercke@fes.de
vorwaerts-projekt@fes.de

Info

Für Kritik und kollegialen Erfahrungsaustausch sind wir jederzeit sehr dankbar und bitten ausdrücklich darum, in diesem Sinne Kontakt mit uns aufzunehmen. Gelegenheit zum persönlichen Gespräch böte sich auch im Rahmen der Tagung „Vorwärts bis 1933“, die wir am 17.05.2017 in Bonn veranstalten werden.
Olaf Guercke, Bibliothek der Friedrich-Ebert-Stiftung, Godesberger Allee 149, D-53175 Bonn
olaf.guercke@fes.de, vorwaerts-projekt@fes.de, http://library.fes.de/inhalt/digital/vorwaerts/vorwaerts.html