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

Teil 1: Scanprozess, Texterkennung und Metadatenanreicherung

Olaf Guercke

Der vorliegende Text ist der erste Teil einer zweiteiligen 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 beleuchtet. Teil 1 bietet zunächst eine Einleitung mit Informationen über den Gegenstand der Digitalisierung und die Ziele sowie den derzeitigen Stand des Projekts. Anschließend werden die verschiedenen Aspekte des Scanprozesses, der maschinellen Texterkennung und der Metadaten-Anreicherung in Augenschein genommen. Ein zweiter Teil, der sich ausführlich mit der Präsentation der Zeitung im Web und den im Projekt entwickelten Suchfunktionalitäten beschäftigt, wird Anfang 2017 folgen.
Der Verfasser versucht in beiden Teilen, kritische Blicke auf die eigene Arbeit zu werfen, so dass der ext 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 erste Teil eines Werkstattberichtes, in dem das seit Anfang 2015 laufende und bis Ende 2017 datierte Projekt in all seinen Aspekten und Entwicklungsperspektiven beschrieben wird, damit Praktiker aus dem Digitalisierungsbereich an unseren Erfahrungen teilhaben können.
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. 100.000 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.

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 Suchtreffen in der Präsentation.

Ergebnis dieses mittlerweile etablierten Workflows

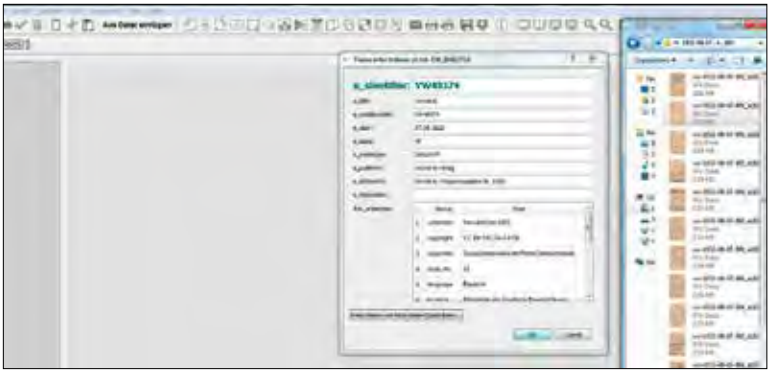

Abb. 1-3: Vorwärts-Ausgabe vom 07.08.1932 beim Import in BCS2

sind Daten-Container, die von der Präsentations-Software MyBib-eL[2] zur Darstellung der Web-Präsentation verwendet werden können. Während der hier vorliegende Teil 1 der Arbeit sich mit der Produktion von Scans und Daten-Containern beschäftigt, wird der in der nächsten b.i.t.online folgende zweite Teil die derzeit laufende Entwicklung der Präsentation mit ihren Suchfunktionalitäten in Zusammenarbeit mit der Herstellerfirma ImageWare Components GmbH (IWC)[3] abbilden. 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 soll. Zum gegenwärtigen Zeitpunkt hoffen wir, im Frühjahr 2017 einen Teilbestand des Vorwärts öffentlich präsentieren zu können.

Scanprozess

Das Scannen einer Zeitungsseite scheint auf den ersten Blick eine banale Angelegenheit zu sein. Wenn man es jedoch mit einer Inhouse-Digitalisierung von großem Umfang und mit hohem Qualitätsanspruch zu tun hat, müssen grundsätzliche Entscheidungen getroffen werden, die eine nicht unerhebliche Auswirkung auf das Projektergebnis haben. Die verschiedenen Arbeitsschritte von der Papierausgabe zum digitalisierten Image wurden im Vorwärts-Projekt zu einem recht frühen Zeitpunkt festgelegt.

- Der Vorwärts liegt in 174 Quartalsfolianten von bis zu 1500 Seiten Umfang vor. Diese werden zunächst komplett durchgesehen, wobei Nummer, Erscheinungsdatum, Seitenzahl und Zustand jeder Ausgabe in Excel-Tabellen festgehalten und Bestandslücken identifiziert werden. Die so entstehenden Daten sind die Grundlage für den eigentlichen Scanprozess und für die Metadatenanreicherung.
- Es hat sich herausgestellt, dass die Digitalisierung der gebundenen Vorlagen wegen starker Unebenheiten und teils sehr enger Bindungen zu unbefriedigenden Ergebnissen führt. Um die Erzeugung von ebenen und textverlustfreien Scans zu ermöglichen, werden die durchgesehenen Folianten daher durch einen Buchbinder geöffnet. Die Einzelseiten werden quartalsweise in säurefreien Archivkisten gelagert.
- Nun werden die Einzelseiten mit einem Din A1-Aufsichtsscanner[4] digitalisiert. Der Scan-Operator hält hierbei einige strukturelle Metadaten (Titel,

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.

4 Hierzu kommt ein Scanner der Marke Zeutschel OS 12000 DIN A1 zum Einsatz, der eigens für das Projekt angeschafft wurde. Vgl.: https://www.zeutschel.de/de/produkte/scanner/farbscanner/os-12000-din-a1.html

Jahrgang, Erscheinungsdatum, Nummer der Ausgabe) fest, auf deren Grundlage die Images in verschiedenen Formaten (Tiff-Master, Jpeg 80%, PDF ganzer Ausgaben) in eine logische Ordnerstruktur ausgeworfen werden.

- Ergebnis sind nach Titel, Jahrgang, Erscheinungsdatum und Ausgabennummer sortierte Scans, wobei pro Ausgabe jeweils ein Unterordner generiert wird. Diese logische Struktur drückt sich auch in den automatisch erstellten Dateinamen der einzelnen Images aus, die eine Identifikation jedes Images auch außerhalb der Ordnerstruktur ermöglichen. Für die Bedienung einer steigenden Zahl von Nutzeranfragen im Rahmen des Projekts erweist sich diese Struktur als sehr hilfreich.
- Anschließend finden die Qualitätskontrolle und der Upload von Master und Derivaten an ihre jeweiligen Speicherorte statt.
- Bestandslücken und stark beschädigte Seiten werden vermerkt und in bestimmten Abständen bei der ULB-Bonn nachgescannt.[5, 6]
- All diese Arbeitsschritte laufen parallel. Stand des Projekts ist folgender:
  - Digitalisiert sind die Jahrgänge 1914–1933 (Insgesamt 100.000 Images)
  - Geöffnet sind die Jahrgänge 1906–1933
  - Durchgesehen und in Tabellen festgehalten sind die Jahrgänge 1899–1933

Texterkennung und Metadaten-Anreicherung

Der Workflow OCR- und Metadatenanreicherung wurde ab Mitte 2016 entwickelt und funktioniert mittlerweile stabil. Zur Weiterverarbeitung für die Präsentation wird das auf Ausgabenebene strukturierte Ordnerarchiv mit den JPEG 80%-Derivaten verwendet. Diese werden Ausgabe für Ausgabe in die Software BCS2-Professional[7] importiert, wo sie zunächst auf Job-Basis mit weiteren Metadaten[8] versehen werden. Die Metadaten werden hierbei nach händischer Eingabe eines ausgabenspezifischen Identifiers durch eine hinterlegte CSV-Datei importiert, die auf Basis der bei der Bestandsichtung erhobenen Daten erstellt wurde. Einzig die Ausgabennummer muss zum gegenwärtigen Zeitpunkt noch per Hand eingegeben werden, was sich jedoch voraussichtlich in der nächsten BCS2-Version ändern wird. In der Praxis hat sich gezeigt, dass es ca. 1 Stunde Arbeit kostet, 100 Ausgaben mit insgesamt ca. 1000 Zeitungsseiten auf diese Weise in BCS-Professional zu importieren. Die OCR-Verarbeitung findet dann per Batch-Operation mittels einer in BSC2-Professional integrierten ABBYY-Finereader-Version für Frakturschrift[9] statt. Diese kostenpflichtige Lösung zeigte im Vergleich zur kostenlosen Tesseract-OCR deutlich bessere Er-

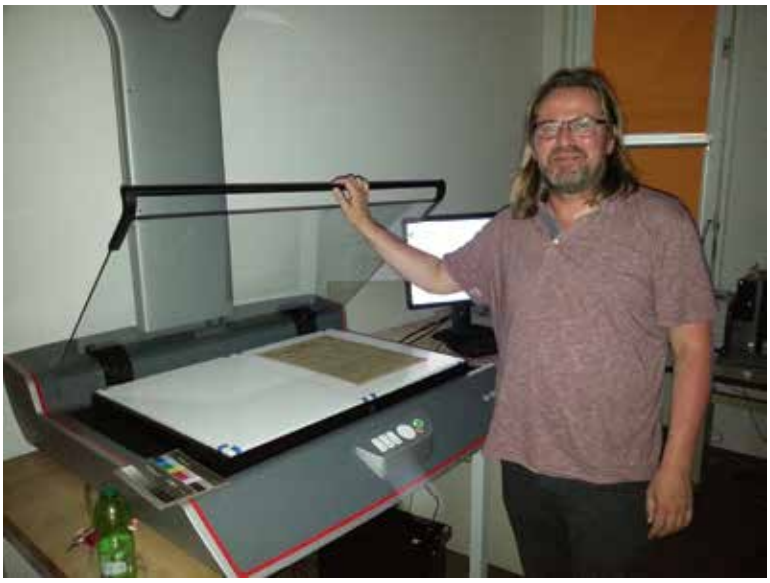

Abb. 4: Inhouse Digitalisierung in der Friedrich-Ebert-Stiftung

Abb. 5: Tabakwerbung im Vorwärts vom 15.06.1929

5 Die Zeitungsabteilung der ULB besitzt ein relativ vollständiges Papierexemplar des Vorwärts und hat sich auf höchst dankenswerte Weise dazu bereit erklärt, das Vorwärts-Projekt bei der Bestandslückenfüllung zu unterstützen. Hierfür möchte sich der Verf. an dieser Stelle herzlich bedanken.

6 Das Austauschen einzelner Images beschädigter Zeitungsseiten mit den Nachscans der ULB hat sich als sehr zeitraubende, kleinteilige Arbeit herausgestellt. Dies führt dazu, dass im Projekt kleinere Textverluste in Kauf genommen werden müssen. Wünschenswert wäre ein Workflow, der den Austausch der Images für Master und alle Derivatversionen automatisiert umsetzt.

7 Vgl.: http://www.imageware.de/loesungen/scansoftware/bcs2_professional/

8 Das Metadatenschema orientiert sich an den DFG-Richtlinien für Digitalisierungsprojekte.

9 Es handelt sich um eine im Rahmen von ABBYY SDK speziell für BSC2-Professional entwickelte OCR-Engine. Wie beim ABBYY Recognition Server entstehen hier Kosten pro verarbeitete Seite.

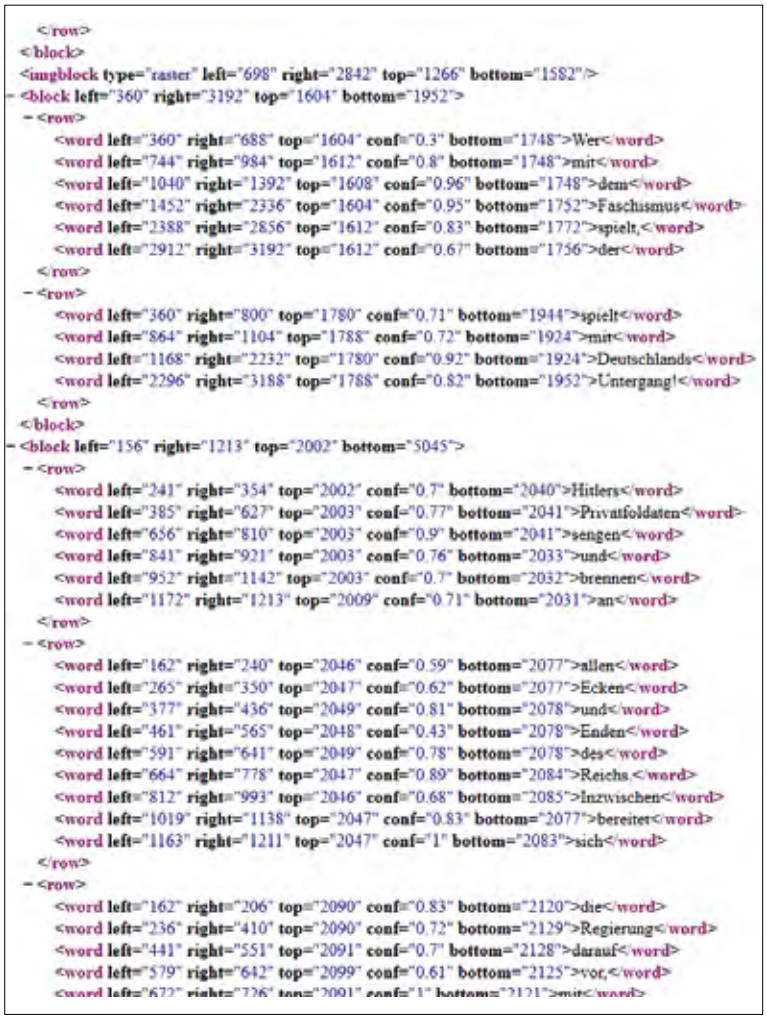

Abb. 6: Ausschnitt aus einer XML-Datei mit zonalen OCR-Ergebnissen zur Ausgabe vom 07.08.1932

gebnisse im Hinblick auf Texterkennungsrate und Performance. Dennoch muss auch hier ein Zeitfaktor beachtet werden. Auf dem hierfür eingesetzten Rechner benötigt die Software ca. 30 Sekunden pro DIN A3 Seite.

Ergebnis dieses Arbeitsgangs sind Container-Dateien, die für jeweils eine Ausgabe JPEGs, zonale OCR-XML-Dateien und eine Metadata-XML-Datei enthalten[10]. Diese Zip-Container umfassen sämtliche Informationen, die zur Präsentation der entsprechenden Ausgaben in MyBib eL benötigt werden und können in eine entsprechend konfigurierte Instanz der Software importiert werden. Zum gegenwärtigen Zeitpunkt sind mit den Jahrgängen 1928–1933 insgesamt 3300 Ausgaben mit ca. 35000 Zeitungsseiten verarbeitet.

10 OCR-XML und Metadata-XML sind proprietäre Formate, die nur von MyBib-eL verarbeitet werden können. Lt. Hersteller soll BCS2 dahingehend entwickelt werden, dass es ALTO-XML und METS/MODS-Metadaten erzeugen kann. Für Digitalisierungsprojekte, die BCS2 als Gesamtlösung ohne den MyBib eL nutzen wollen, könnte es sich lohnen, diese Entwicklung abzuwarten, die auch im Hinblick auf kooperative Projekte, die ALTO-XML und METS/MODS erfordern, absolut wünschenswert wäre.

Der gesamte Korpus wird voraussichtlich Ende 2017 verarbeitet sein.

Ausblick

In der Projektplanung stehen aktuell die Implementierung der Präsentationsplattform auf dem hbz Server und das Testen von Performance und Suchfunktionalitäten mit einem größeren Bestand an Vorwärts-Ausgaben an. Anschließend hoffen wir, im Frühjahr 2017 bei laufendem Scan- und Verarbeitungsprozess mit einem Teilbestand an die Öffentlichkeit gehen zu können. Ein Zwischenfazit vor der Bewältigung dieser entscheidenden Projektphase könnte wie folgt lauten:

- Inhouse-Digitalisierungsprojekte von Papierausgaben historischer Zeitungen sind zeit- und kostenintensiv. In unserem Fall beschäftigen 200.000 Zeitungsseiten einen Bibliothekar (25 Stunden) und zwei Hilfskräfte (insgesamt 23 Stunden) für drei Jahre. Dieser Weg empfiehlt sich daher für Zeitungen, deren Relevanz sehr hoch eingeschätzt wird und deren Umfang einigermaßen überschaubar ist. Bei sehr viel umfangreicheren Massendigitalisierungen wird man nicht umhinkommen, vom Mikrofilm zu scannen und den Verlust an Authentizität hinzunehmen.
- Diesen Kosten und Mühen steht ein nicht hoch genug einzuschätzender Nutzen gegenüber. Die massive Aufwertung der Quelle durch die erleichterte Zugänglichkeit einer strukturierten Web-Präsentation wird nochmals potenziert durch die Texterkennung, die neben der erweiterten Volltextsuche auch ganz neue Nutzungsmöglichkeiten des Corpus, etwa in der Linguistik und verschiedenen Bereichen der Digital Humanities eröffnet. Um die in diesem ersten Teil des Textes geschilderten aufwändigen Vorarbeiten für die Nutzer fruchtbar zu machen, ist eine leistungsfähige Web-Präsentation unabdingbar. Inwiefern das Vorwärts-Projekt diesen Anspruch erfüllen kann, wird im 2. Teil dieses Textes gezeigt werden.

Website des Projekts: http://library.fes.de/inhalt/digital/vorwaerts/vorwaerts.html

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