

## NZZ: 225 Jahre Old Economy vernetzt – Realisierung des digitalen Archivs der Neuen Zürcher Zeitung

Stefan Eickeler, Lars Bröcker,  
Fraunhofer-Institut Medienkommunikation  
Schloss Birlinghoven  
53754 Sankt Augustin, Deutschland  
{stefan.eickeler,lars.broecker}@imk.fraunhofer.de

Ruth Haener  
Neue Zürcher Zeitung  
Falkenstrasse 11  
8021 Zürich, Schweiz  
r.haener@nzz.ch

**Abstract:** Die Speicherung von Zeitungsausgaben in einem digitalen Archiv bieten gegenüber der traditionellen Archivierung der Zeitungsausgaben in Bänden oder auf Mikrofilm durch schnelle Such- und Zugriffsmöglichkeiten einen deutlich höheren Komfort für den Benutzer. In diesem Beitrag wird der Aufbau des digitalen Archivs der Neuen Zürcher Zeitung beschrieben.

### 1 Einleitung

Information muss fließen können. Wird sie gedruckt, gebunden, gestapelt und magaziniert, so bedeutet dies zunächst Stillstand, dann unmerkliche Sickerverluste, die schliesslich zum Austrocknen führen können. Die rasche Entwicklung von digitalen Archiven - Internet inklusive - und effizienten Suchmaschinen im letzten Jahrzehnt hat das Selbstverständnis für den freien, wenn auch oftmals noch zu wenig kostenpflichtigen Zugang zu Informationen gestärkt. Die NZZ wird im Jubiläumsjahr 2005 ihre Berichterstattung über 225 Jahren Weltgeschehen in einem digitalen Archiv zugänglich machen. Im Fraunhofer-Institut Medienkommunikation hat sie den geeigneten Partner für die Realisierung gefunden.

Aus archivischer Sicht sprechen zusätzlich zum Nutzen des freien und raschen Zugriffs für die Digitalisierung auch buchstäblich handfeste Gründe. Der Papierbestand wird geschont und damit die Langzeitsicherung von Weltkulturgut qualitativ verbessert. Die NZZ gehört zu den ältesten Tageszeitungen der Welt, ist vollständig erhalten und zusätzlich auf Mikrofilmen gesichert. Zwar gelten analoge Datenträger allgemein noch immer als stabiler als digitale. Sie sind aber durch sorglosen Umgang, Umwelteinflüsse, Feuer, Wasser, Pilz, Diebstahl, Kriegs- und anderen Katastrophenfolgen zerstörbar. Bleiben transparent organisierte digitale Daten in aktiven Systemen, wie dies hier vorgesehen ist, so überleben sie durch Migration und parallele Sicherung langfristig. Dies trifft besonders dann zu, wenn Primärdaten (digitale Originale) und Sekundärdaten (technische Information und Layout-Daten) strikt getrennt organisiert sind und dies langfristig nachvollziehbar dokumentiert ist. Nur dann können technische Fortschritte in dynamischer Umgebung effizient genutzt werden, was wiederum den Wert des Archivs erhöht. Das aktuelle interne digitale Archiv der NZZ offeriert Daten ab 1993 und ist längst zu einem Produk-

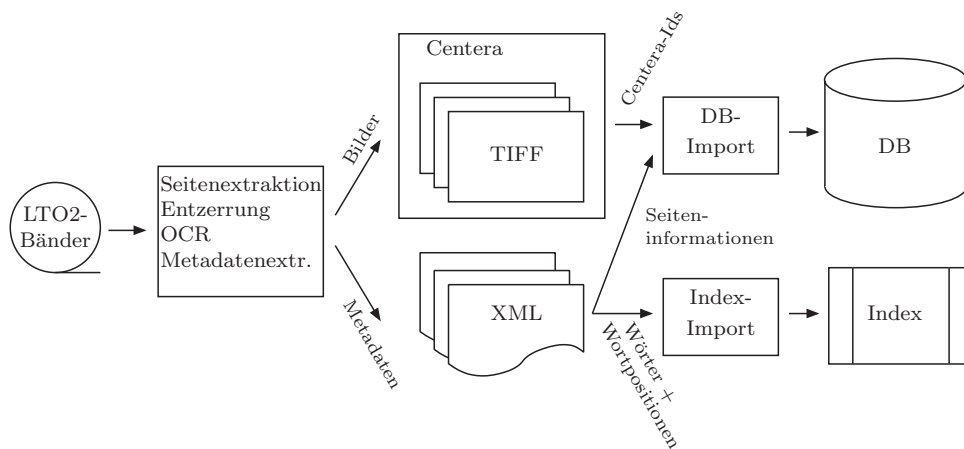


Abbildung 1: Workflow zum Erstellen des Archivs

tionsfaktor im Tagesgeschäft des Zeitungsmachens geworden. Deshalb bleibt die Bestandespflege und -erhaltung des Digitalisats ab 1780 gewährleistet. Das NZZ-Archiv sieht Letzteres im Sicherheitsbereich als Ergänzung zum Papier und wird auch die vollständig automatisierte Mikroverfilmung weiterführen.

Angelsächsische Zeitungen sind den Weg der Digitalisierung vorausgegangen, so *The Christian Science Monitor* als Pionier (ab 1908), dann *The New York Times* (ab 1851), *The Washington Post* (ab 1877), *The Wall Street Journal* (ab 1889). Sie haben sich entschieden, zusätzlich zur Erstellung der PDF-Dateien der einzelnen Seiten ihre Digitalisate im asiatischen Raum nachbearbeiten zu lassen, die Seiten in Artikel zerlegen und die betreffenden Artikelbestandteile systematisch verlinken zu lassen etc.. Dadurch erreichen sie eine hohe Qualität bei den Suchresultaten. Allerdings ist die Nachbearbeitung zu ca. zwanzig Prozent manuell und somit ausgesprochen ressourcenintensiv.

Die NZZ geht hier neue und andere Wege. Schätzungsweise zwei Millionen Zeitungsseiten sind eine quantitative Herausforderung, die nach technischen Verfahren ruft, die mit unter neu zu entwickeln sind. Datenbasis ist der Mikrofilmbestand im Umfang von ca. 1525 Filmen unterschiedlicher Qualität, die gescannt, nachbereitet und maschinenlesbar verarbeitet werden. Resultat ist ein ab September vorliegendes Seitenarchiv mit einem automatisierten faksimilierten Transkript für die Texte in Frakturschrift.

## 2 Verarbeitung

Ausgehend von dem Mikrofilm werden die Bilder der Zeitungsseiten mit einem Mikrofilmscanner in einer Auflösung von 300 ppi in 256 Graustufen digitalisiert. Dieses resultiert in einer Datenmenge von ca. 25 MB für eine Einzelseite.

Die Digitalisierung wird von einem externen Dienstleister durchgeführt und die

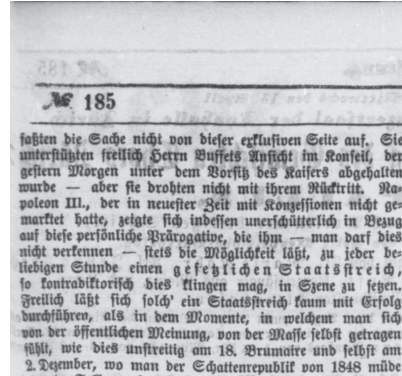
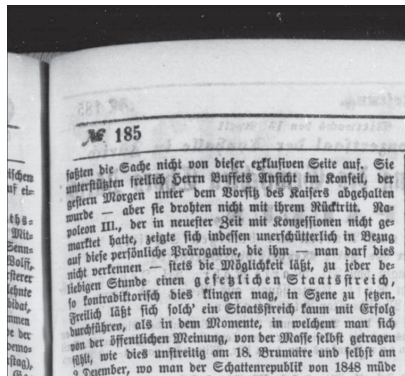


Abbildung 2: Ausschnitt einer Seite vor und nach der Entzerrung

digitalisierten Mikrofilme auf LTO-2 Magnetbändern zum Fraunhofer IMK transportiert.

In der Verarbeitungskette (siehe Abbildung 1) werden die Seiten aus den gescannten Bildern extrahiert, eine Zeichenerkennung (OCR) durchgeführt und die relevanten Metadaten herausgefiltert. Das Bild der Seite wird im TIFF-Format gespeichert, die Metadaten in einem XML basierten Format.

Die Seitenextraktion generiert die Bilder einer Zeitungsseite aus den gescannten Aufnahmen eines Mikrofilms. Das Ziel ist hierbei, die Zeitungsseite mit möglichst wenigen Verzerrungen zu extrahieren und ein Bild der Seite zu erhalten, das dem Erscheinungsbild der Originalseite entspricht.

Bis zum Jahrgang 1960 wurden die Zeitungsseiten aus den gebundenen Zusammenstellungen doppelseitig verfilmt. Hierbei entstehen starke Verzerrungen, die durch die zweidimensionale Abbildung der gekrümmten Buchseite entsteht. Diese Verzerrungen sind besonders an der Buchfalz zu sehen. Die Verzerrung lässt sich unter der Annahme, daß der obere und der untere Rand der Buchseite eine gerade Linie darstellen, leicht automatisch bestimmen und korrigieren. Abbildung 2 zeigt einen Ausschnitt einer Seite vor und nach der Entzerrung.

Die Zeichenerkennung wird mit dem Abby Finereader 7.0 SDK durchgeführt. Diese Software kann Antiqua- und Frakturschrift erkennen und erreicht eine hohe Erkennungsgenauigkeit. Das SDK lässt sich gut in das Programm zur Entzerrung und Metadatenextraktion integrieren und ermöglicht daher eine effiziente Verarbeitung auf einem Cluster aus 20 Computern.

### 3 Digitales Archiv

Die Neue Zürcher Zeitung verfügt bereits über ein umfangreiches Intranet-Angebot, das unter anderem ein digitales Archiv der Zeitungen ab 1993 enthält. Dieses Archiv

wird im Haus intensiv genutzt, so daß sich viele Anforderungen in Sachen Design und Funktionalität an das neu zu schaffende Seitenarchiv ab 1780 daraus ableiten ließen. Die Kernanforderungen an das Archiv lassen sich wie folgt zusammenfassen:

- Suche nach Einzelbegriffen und Phrasen, optionale Einschränkung auf einzelne Tage oder Zeiträume
- Suche nach ganzen Zeitungen durch Datumseingabe
- Ausliefern der Suchergebnisse als PDF, darin Hervorhebung der gesuchten Begriffe
- Anbieten eines automatischen Transkripts von in Fraktur gesetzten Zeitungen

Die zur Erfüllung dieser Anforderungen benötigten Arbeitsschritte werden in den folgenden Abschnitten beschrieben. Abbildung 1 zeigt als Überblick den Workflow zur Erstellung des Archivs aus den Rohdaten.

### **3.1 Datenbankerstellung**

In der Datenbank werden alle Metadaten gespeichert, die zu den jeweiligen Seiten vorhanden sind. Dies sind die Abmessungen der Seite, eine eventuell erkannte Seitenzahl oder ein Seitentitel sowie ein Flag, das gesetzt wird, wenn eine Titelseite erkannt worden ist. In diesem Fall wird in einer anderen Tabelle ein Zeitungseintrag vorgenommen, der die Daten enthält, die diese Zeitung beschreiben: Das Datum, die Ausgabennummer und der Ausgabentyp (z.B. Morgen-, Mittag-, Abendausgabe), da die NZZ zeitweise bis zu fünfmal am Tag erschienen ist. Diese Daten werden alle aus den XML-Dateien ausgelesen. Daneben gibt es aber noch eine andere Datenquelle: Die Bilder der Zeitungsseiten werden auf einem Centera-System der Firma EMC gesichert. Dabei handelt es sich im Wesentlichen um ein RAID-System, das auf transparente Speichererweiterung im Hintergrund ausgelegt ist und deswegen keinen direkten Zugriff auf enthaltene Daten erlaubt. Vielmehr werden Daten unter Verwendung eines eindeutigen Hashwerts abgerufen, der bei Speicherung von dem System erzeugt und geliefert wird. Diese sog. Centera-ID wird für jede Seite in der Datenbank hinterlegt, um einen Zugriff auf das Digitalisat zu ermöglichen.

### **3.2 Indexerstellung**

Die NZZ feierte im Januar des Jahres 2005 ihren 225. Geburtstag. Im Verlauf der Jahrhunderte sind ca. zwei Millionen Zeitungsseiten veröffentlicht worden. Dafür einen effizienten Suchindex zu erstellen, ist eine sehr große Aufgabe. In diesem Fall wurde sie allerdings dadurch erschwert, daß eine Stoppwortentfernung nicht möglich war, sollte nach Phrasen gesucht werden können, die durchaus nur aus Stoppwörtern bestehen können (z.B. "Sein oder nicht sein").

Als Indizierungswerkzeug wurde Apache Lucene ausgewählt, eine Java-Bibliothek, mit der sich Volltextsuchmaschinen entwickeln lassen (eine ausführliche Beschreibung dieser Bibliothek findet sich in [GH05]). Entscheidend für den Einsatz innerhalb des Projekts waren drei Punkte: Die Bibliothek passt gut in die Softwareum-

gebung der NZZ, sie ist open source, so daß einer speziellen Anpassung an das Projekt nichts im Weg stand und sie enthält bereits alle Suchfunktionalitäten, die im Projekt benötigt werden.

Neben der reinen Textindizierung erlaubt Lucene die Speicherung zusätzlicher Informationen über die Inhalte. So werden verschiedene Verweise auf die Datenbank hinterlegt, die der Webapplikation zeitaufwendige Datenbankabfragen ersparen. Die wichtigste Zusatzinformation zu den einzelnen hinterlegten Wörtern ist die Position jedes Worts innerhalb der Seite. Die Positionen der Wörter werden im Index abgelegt, aber nicht indiziert, so daß sie als Ergebnis zur Verfügung stehen, wenn ein Wort der Seite in eine Suchanfrage passt. Damit lässt sich das Highlighting auf den Ergebnissen ohne weitere Datenbankabfragen realisieren.

### 3.3 Weboberfläche

Die Weboberfläche ist im Design der anderen Bestandteile der Website der NZZ gehalten und bewusst einfach strukturiert. Der Ablauf entspricht weitestgehend klassischen Suchmaschinen. In das Sucheingabefeld können Wörter und/oder in Anführungszeichen eingeschlossene Phrasen eingegeben werden. Optional kann ein zu durchsuchender Datumsbereich angegeben werden. Diese Seite führt auf eine Ergebnisliste, chronologisch aufsteigend sortiert. Die Auswahl eines der Listeneinträge führt auf eine Seite, die ein PDF der Zeitungsseite enthält, auf der die Suchterme farblich hervor gehoben sind. Daneben erlaubt eine Navigationsleiste das Blättern in der betreffenden Ausgabe sowie der Ergebnisliste der Suche. Falls es sich um eine in Fraktur gesetzte Zeitung handelt, wird ein automatisch erzeugtes Transkript angeboten - ebenfalls ein PDF.

## 4 Zusammenfassung und Ausblick

In diesem Beitrag wurden die Verarbeitungsschritte zur Realisierung des digitalen Archivs der Neuen Zürcher Zeitung vom Mikrofilm bis zur Webapplikation beschrieben.

Weiterentwicklungen und Verbesserungen des digitalen Archivs sind eine Fuzzy-Suche und eine automatische Suchanfragenerweiterung für die altdeutsche Rechtschreibung. Für die Verarbeitung befindet sich eine automatische Artikelseparierung in der Entwicklung.

## Literatur

[GH05] Otis Gospodnetic und Erik Hatcher. *Lucene in Action*. Manning Publications Co., 2005.