DIGITALE REPRODUKTION UND INHALTSERSCHLIEßUNG VON HISTORISCHEN TAGESZEITUNGEN IM STEIERMÄRKISCHEN LANDESARCHIV

Josef Riegler[*]

UDK: 004.3:070:930.25(436.4)

Josef Riegler: Digitale Reproduktion und Inhaltserschließung von historischen Tageszeitungen im Steiermärkischen Landesarchiv. *Tehnični in vsebinski problemi klasičnega in elektronskega arhiviranja, Zbornik referatov z dopolnilnega izobraževanja, Maribor 7/2008, str. 431-436.*

Izvirnik v nemščini, izvleček v nemščini in angleščini, povzetek v slovenščini.

Im Beitrag wird über rechtliche, technische und organisatorische Aspekte des großen Projektes der Digitalisierung und Indexierung von cca. 2,5 Mio. Seiten der steirischen Tageszeitungen im Steiermärkischen Landesarchiv diskutiert. Ebenfalls werden die für die Realisierung und den Status des Projektes ausgewählten Parameter erklärt.

UDC: 004.3:070:930.25(436.4)

Josef Riegler: Digital Reproductions and Content Capture of Historic Newspapers in the Provincial Archives of Styria. *Technical and Field Related Problems of Traditional and Electronic Archiving. Conference Proceedings, Maribor 7/2008, pp. 431-436.*

Original in German, abstract in German and English, summary in Slovenian.

The article discusses legal, technical and organisational aspects of a large project for the digitalisation and indexation of approximately 2.5 million pages of historic newspapers in the Provincial Archives of Styria. The author also explains the chosen parameters for the realisation and the status of the project.

Ključne besede: digitalna reprodukcija, digitalni arhiv, Štajerski deželni arhiv, štajersko časopisje.

Die zeitgeschichtliche Forschung und viele Bürger brauchen einen einfachen Zugang zu den historischen Tageszeitungen. Dieser Zugang war in der Steiermark bisher nur über Mikrofilmbestände möglich. Für die umfassende Erschließung des Zeitungsinhaltes fehlten bisher die erforderlichen Personalkapazitäten. Die einzige machbare Möglichkeit, einen einfachen, von den Kosten her zu bewältigenden Zugang zu schaffen besteht in der Digitalisierung der Zeitungsseiten und der anschließenden Umwandlung der Inhalte in maschinenlesbaren Text. Dafür wurde vom Steiermärkischen Landesarchiv unter der Be zeichnung „Digitales Steirisches Zeitungsarchiv“ in Kooperation mit der Steiermärkischen Landesbibliothek und der Kleinen Zeitung ein Projekt definiert. Ziel des Projektes ist es, alle verfügbaren steirischen Tageszeitungen in durchsuchbarer Form in den Lesesälen des Archivs und der Bibliothek zur Verfügung zu stellen. Alle Ausgaben, die keinen Schutzfristen unterliegen, stehen nach Projektabschluss via Internet zur Nutzung bereit. Für die Digitalisierung stehen Zeitungsbestände des Steiermärkischen Landesarchivs, der Steiermärkischen Landesbibliothek, der Kleinen Zeitung in Graz und einiger privater Leihgeber zur Verfügung.

[*] ***Dr. Josef Riegler, Direktor des Steiermärkischen Landesarchivs, Karmeliterplatz 3, A-8010 Graz, Austria.***

I. MACHBARKEITSSTUDIE

Die Machbarkeit des Projekts in technischer, rechtlicher, organisatorischer und finanzieller Hinsicht wurde 2005 geprüft. Ihm Rahmen eines Vorprojekts wurden umfangreiche Tests durchgeführt und die kritischen Parameter erhoben. Bei Scandienstleistern wurden national und international die Kosten für die Abwicklung des Projektes von der Abholung der Originale bis zur Lieferung der fertig in durchsuchbaren Text umgewandelten Zeitungsseiten erhoben. Parallel dazu wurden die rechtlichen Rahmenbedingungen geprüft und die Risikobereiche ermittelt. In einem dritten Bereich erfolgte die Evaluierung der technischen Möglichkeiten, die digitalisierten Zeitungen optimal zu nutzen. Das Gesamtvolumen der zu digitalisierenden Zeitungen liegt bei rund 2,5 Millionen Seiten, davon rund ein Drittel im Großformat. In das digitale Zeitungsarchiv werden sieben wichtige steirische Tageszeitungen ab Erscheinungsbeginn bis zur Einstellung des Mediums respektive bis zu jenem Zeitpunkt aufgenommen, an dem die Verlage elektronisch archivierte Versionen zur Verfügung stellen. Die folgenden Ausführungen sind einzelnen Bereichen gewidmet, die für eine erfolgreiche Digitalisierung von Tageszeitungen Bedeutung haben.

II. COPYRIGHT UND VERWERTUNGSRECHTE

Copyright ist (leider) nicht gleichbedeutend mit „right to copy“. Für die Nutzung der Werke fremder Personen gelten nationale und internationale, gesetzlich festgeschriebene Rahmenbedingungen. Sie schützen in erster Linie die Rechte der Urheber, geben aber auch jenen, die Schöpfungen anderer nutzen wollen, klare Vorgaben. Nach derzeitigem Stand des österreichischen Urheberrechts erlischt dieses 70 Jahre nach dem Tod des Urhebers. Dann sind diese Werke gemeinfrei und können von jedermann für jeden beliebigen Zweck genutzt werden. Für urheberrechtlich geschütztes Gut in Archiven und Bibliotheken gelten teilweise Sonderbestimmungen. Das österreichische Urheberrecht kennt z. B. das Recht auf die Reproduktion zum persönlichen Gebrauch, die ohne Zustimmung des Urhebers angefertigt werden kann. Solche Kopien können auch zum persönlichen Gebrauch einer juristischen Person bestimmt sein. Die technische Form dieser erlaubten Kopie ist gesetzlich nicht geregelt. Das bietet die Möglichkeit, Zeitungen auch in digitaler Form zur Nutzung innerhalb eines Archivs oder einer Bibliothek anzubieten. Wie komfortabel die Benutzung der digitalen Zeitungsseiten gestaltet wird und wie gut die digitale Reproduktion sein kann, unterliegt keinen Beschränkungen. Schranken setzt das Urheberrecht allerdings dort, wo es um die Herstellung von Reproduktionen von Zeitungsseiten für die Benützer geht.

Der vielfach von Benützern gewünschte Idealzustand, alte Ausgaben von Tageszeitungen über das Internet benutzen und im Text recherchieren zu können, liegt vorwiegend aus rechtlichen Gründen, noch in weiter Ferne. Beschränkungen des Urheberrechts sind auch der Grund, dass keine Tageszeitung des deutschen Sprachraumes größere Mengen historischer Ausgaben im Volltext oder im originären Erscheinungsbild zur Nutzung anbietet. Was sind die weiteren großen Hindernisse für die Online-Nutzung digitalisierter Zeitungen? Da sind zunächst die Rechte der Zeitungsverlage an den von ihnen herausgegebenen Medien zu nennen. Mit den Verlagen können aber, falls ein Archiv oder eine Bibliothek digitalisierte Zeitungen zur Nutzung anbieten will, entsprechende Verträge abgeschlossen werden.

Schwerer wiegen die Rechte der Autoren an den von ihnen verfassten Texten. Die Verfasser der von den Zeitungen veröffentlichen Texte standen häufig in einem Vertragsverhältnis zum Zeitungsverlag. Hier kann man annehmen, dass die Veröffent-

lichungsrechte an den Texten an die Verlage übergegangen sind. Bei der Veröffentlichung von Texten anderer Personen ist die Rechtslage nicht so eindeutig. Ob der Text dabei dem ursprünglichen Erscheinungsbild entspricht oder in neuem Layout erscheint, ist dabei nicht von Belang. Die größte Hürde bei der Nutzung von digitalisierten Zeitungen sind Rechte der Fotografen sowie der Bildagenturen an den in der Zeitung gedruckten Bildern. Selbst wenn ein Fotograf einem Zeitungsverlag vor vielen Jahren ein umfassendes Verwertungsrecht an seinen Bildern eingeräumt hat, gestatten diese Verträge nach Ansicht vieler Juristen die Verbreitung dieser Bilder im Internet nicht. Das wird damit begründet, dass zum Zeitpunkt des Vertragsabschlusses das Internet noch nicht erfunden war und daher das Verwertungsrecht an den Bildern über dieses Medium nicht eingeräumt werden konnte. Daher stünde den Fotografen in diesem Fall eine Entgeltanspruch zu. Es ist aber unmöglich, alle Urheber und Rechteinhaber der in den Zeitungen über die Jahrzehnte hinweg veröffentlichten Bilder ausfindig zu machen und von ihnen das Veröffentlichungsrecht für das Internet zu erwerben.

Werden ganze Zeitungsseiten in ihrem ursprünglichen Erscheinungsbild dennoch im Internet zur Nutzung bereitgestellt, besteht das große Risiko, von Fotografen, deren Rechtsnachfolgern oder Berufsverbänden geklagt zu werden. Das finanzielle Risiko ist hier extrem groß. Das ist der Grund für die faktische Nichtexistenz jüngerer historischer Zeitungen im Internet. Kein Zeitungsverlag, kein Archiv und keine Bibliothek kann dieses finanzielle Risiko eingehen.

III. ZUM STAND DER DIGITALISIERUNGSVERFAHREN BEI ZEITUNGEN

1. DIGITALISIERUNG VOM MIKROFILM

Beim Digitalisieren vom Mikrofilm stand und steht noch immer der ökonomische Aspekt im Vordergrund. Sofern ausreichend gute Mikroformen vorhanden sind, kann die Digitalisierung an Dienstleister außer Haus vergeben werden, die in kurzer Zeit relativ große Seitenmengen in ein digitales Abbild umwandeln können. Allerdings ist dabei die Qualität der Scans an die Qualität des Mikrofilms gebunden. Hier gibt es eine große Schwankungsbreite. Weiße Schrift auf schwarzem Grund ist für die Augen nicht gut lesbar – dieses Problem kann durch entsprechende Software sehr leicht gelöst werden.

Beim kostengünstigen Mikrofilm können nur bitonale Abbilder der Vorlage hergestellt werden. Das bedeutet, dass die Wiedergabequalität der Bilder in den Zeitungen nur in beschränktem Maß möglich ist. Bei blassen Druckvorlagen kann es außerdem vorkommen, dass einzelne Textteile bei der Tontrennung in Schwarz und Weiß ausfallen. Dieser Textverlust ist auch durch die beste Software nicht auszugleichen. Der Silberfilm hoher Qualität war und ist ziemlich teuer, erfordert viel Sachkenntnis auf allen Verarbeitungsstufen und wurde daher nur selten für die Verfilmung von Zeitungen eingesetzt.

Die Filmqualität reicht oft nicht aus, um ein gutes digitales Bild der Zeitungsseiten zu erzeugen. Für den nächsten, aus Sicht des Inhaltszuganges entscheidenden Schritt, nämlich durch automatische Texterkennung einen maschinenlesbaren und so leicht durchsuchbaren Text zu erzeugen, ist die Filmqualität oft zu gering. Die Steiermärkische Landesbibliothek verfügt z. B. über einen umfangreichen Bestand an verfilmten Tageszeitungen. Für die Digitalisierung sind diese Filme wegen der nicht ausreichend guten Qualität allerdings nicht geeignet.

2. DIGITALISIERUNG DER ORIGINALAUSGABEN

Sind von Zeitungsbeständen keine brauchbaren Mikroformen vorhanden, müssen die Originalausgaben zur Digitalisierung herangezogen werden. Kaum jemand würde auf die Idee kommen, Zeitungsbestände zuerst zu verfilmen und dann, ausgehend vom Mikrofilm, eine digitale Version herzustellen und dies mit Kostenargumenten begründen. Dennoch gibt es zwischen Digitalisierung und Mikrofilm einen Zusammenhang. Um eine kostengünstige Langzeitsicherung der Digitalisate zu erreichen, können heute die digitalen Daten im COM-Verfahren (computer output on microform) auf Mikroformen geschrieben werden. Bei strikter Einhaltung aller Qualitätskriterien sind diese Mikrofilme meist so gut, dass sie bei Bedarf wieder digitalisiert und durch Texterkennung neuerlich maschinenlesbar gemacht werden können. Die Digitalisierung der Originalseiten ist technisch sowohl im Aufsichtverfahren als auch im Durchlaufverfahren möglich. Während im Aufsichtverfahren auch sehr große Zeitungsbände gescannt werden können, weisen die Durchlaufscanner in der maximal möglichen Breite Beschränkungen auf rund 30,0 bis 30,5 cm auf. Die Seitenlänge ist bei modernen Geräten ohne Probleme zu bewältigen. Da Projekte für die Digitalisierung von Zeitungen meist ein hohes Seitenvolumen aufweisen spielt die Zahl der in einer Zeiteinheit möglichen Scans auf der Kostenseite eine sehr große Rolle. Während von einem sehr guten Scanneroperator an einem Arbeitstag vielleicht 1.500 Seiten gescannt werden können, schaffen Durchzugscanner bei gleichen Personalkosten und Seitengrößen 8.000 und mehr Seiten pro Tag. Der Gedanke, das beim Aufsichtscannen so zeitaufwendige Umblättern von einer Maschine machen zu lassen, wurde bereits in die Tat umgesetzt. Die ersten Scanroboter sind schon im Einsatz. Allerdings können die meisten Roboter nur Bücher bis ca. 30 cm Höhe verarbeiten, für großformatige Zeitungsbände ist das Angebot noch sehr gering.

3. MANUELLE HERSTELLUNG MASCHINENLESBARER ZEITUNGSARTIKEL

Ein dritter Weg, die Inhalte der Zeitungen zu digitalisieren, wurde schon vor Jahren von einigen amerikanischen Zeitungsverlagen beschritten. Die Zeitungsseiten wurden im Sinne einer globalen Arbeitsteilung in Niedriglohnländern wie China oder Indien manuell Seite für Seite und Artikel für Artikel abgeschrieben. Über die Fehlerquote, die beim Abschreiben eines Textes in einer Sprache, die von den Schreibenden oft nicht verstanden wurde, liegen keine verlässlichen Angaben vor. Erreicht wurde mit diesem Verfahren jedoch, dass die Inhalte der Zeitungen maschinenlesbar sind und somit leicht genutzt werden können. Damit ist der wichtigste Zweck der Zeitungsdigitalisierung angesprochen. Der Benutzer möchte einen möglichst einfachen Weg zu den ihn interessierenden Zeitungsartikeln vorfinden.

IV. ZUGANG ZU DEN INHALTEN DER ZEITUNGEN

Die einfache Lösung ist die Umwandlung der originalen Zeitungsseiten in ein digitales Bild und dessen Speicherung im Filesystem. Für die Benutzung wird nur eine klar strukturierte Oberfläche benötigt, die möglichst einfach zu den gewünschten digitalen Zeitungsausgaben mit ihren logisch angeordneten Seiten führt. Dass die Navigation von einer Seite zur darauffolgenden mit nur einem Mausklick möglich sein sollte, versteht sich von selbst. Diese Form der Benützung von Zeitungen auf dem Computerbildschirm ist streng genommen nichts anderes als die Benützung einer Tageszeitungsausgabe auf Mikroform mit anderen technischen Hilfsmitteln. In manchen Projekten werden die Artikel einer gescannten Zeitungsausgabe mit OCR-Verfahren in maschinenlesbaren Klartext umgewandelt. Dem Benutzer wird dabei ein Artikel in

neuem Layout auf dem Bildschirm präsentiert – aber auch nur dieser. Benutzer erhalten keine Information über den Zusammenhang dieses Artikels mit anderen zum gleichen Thema, sehen kein zum Artikel gehörendes Bild und erhalten auch keinen Eindruck vom Erscheinungsbild der ursprünglichen Zeitungsseite.

Diese maschinenlesbaren Artikel können leicht in eine Indexdatenbank eingebracht und damit abrufbar gemacht werden. Dieses Verfahren ist nur dann wirklich gut, wenn jeder Artikel auf richtige Texterkennung geprüft wurde. Selbst die beste Texterkennungssoftware ist nicht in der Lage, in Normaltype gedruckte Vorlagen zu 100 % richtig zu erkennen. Dass das Korrekturlesen einer extrem großen Textmenge nicht finanzierbar ist, braucht nicht besonders betont zu werden. Dazu kommt, dass z. B. in Österreich bis in die ersten Jahre nach dem Ende des Zweiten Weltkriegs die Zeitungen in Frakturschrift gedruckt worden sind. Erst seit Kurzem ist eine OCR-Software verfügbar, die auch diese Schrift einigermaßen richtig erkennen kann. In diesem Bereich sind noch einige technische Fortschritte zu erwarten. Dennoch ist es nicht sinnvoll, einen unkorrigierten OCR-Text mit all seinen Erkennungsfehlern und nicht erkannten Zeichen als reine Textseite verfügbar zu machen – das menschliche Auge bleibt zu stark an den Fehlerstellen hängen. Der Ausweg aus diesem Dilemma kann über PDF-Dateien genommen werden. Dabei werden für jede Seite zwei Schichten angeboten. Die obere Schicht ist das Abbild der Zeitungsseite, darunter liegt, unsichtbar, für die Maschine aber lesbar, der Klartext. Fehler in der Texterkennung sind zunächst nicht sichtbar, sondern nur durch gezieltes Abrufen eines auf dem Image angezeigten, meist nicht mit klaren Konturen gedruckten Wortes zu eruieren. Wird kein Treffer angezeigt, wurden die entsprechenden Zeichen nicht richtig erkannt. Erkennungsquoten von über 99 % sind als sehr gut einzustufen. Gegen diese Methode könnte eingewandt werden, dass bei der Suche nur die eindeutig erkannten Wörter gefunden werden können, nicht oder fehlerhaft erkannte Wörter werden nicht gefunden. Dem ist entgegenzuhalten, dass durch den beschriebenen Weg aber der Großteil der Artikel gefunden werden kann – bei extrem großer Zeitersparnis und dennoch voller Lesbarkeit des gefundenen Artikels.

V. DAS PROJEKT „DIGITALES STEIRISCHES ZEITUNGSARCHIV DES STEIERMÄRKISCHEN LANDESARCHIVS

Das Ergebnis des eingangs kurz vorgestellten Vorprojektes war: Die Digitalisierung ist technisch und finanziell machbar. Da das Verbreiten im Internet aus rechtlichen Gründen nicht möglich ist, ist die Benützung aller urheberrechtlich noch geschützten Zeitungsausgaben nur innerhalb des Landesarchivs und der Landesbibliothek möglich. Urheberrechtlich nicht mehr geschützte Zeitungsausgaben werden im Internet zugänglich gemacht. Nach einer eingehenden Prüfung der Kosten für die Digitalisierung durch einen Dienstleister und der Kosten für die Eigendigitalisierung fiel die Entscheidung, die Digitalisierung im Steiermärkischen Landesarchiv selbst durchzuführen. Folgende Eckpunkte wurden festgelegt:

- Ein leistungsfähiger farbtauglicher Durchzugscanner wird beschafft
- Aufsichtscans werden mit der vorhandenen Scannerausrüstung des Landesarchivs durchgeführt
- Nach Möglichkeit werden aus Kostengründen die Zeitungsbände für die Bearbeitung im Durchzugsscanner aufgelöst und nicht mehr neu gebunden.
- Die bearbeiteten Zeitungsseiten werden in säurefreien Umschlägen und Archivboxen gelagert.

- Zusätzlich benötigtes Personal wird mit freien Dienstverträgen angestellt.
- Die Speicherung der großen Datenmengen (ca. 8 Terabyte) erfolgt redundant auf kostengünstigen Massenspeichern innerhalb des Landesarchivs. Sicherungskopien werden ausgelagert.
- Die Scans werden vom Scanner als TIFF-Dateien im Farbraum RGB mit einer Auflösung von 400 dpi ausgegeben, Komprimierung auf ca. 5-6 MB Größe.
- Diese Daten werden automatisiert in das Format JPEG2000 umgewandelt, Dateigröße ca. 1,5 MB je Seite.
- Aus den TIFF-Dateien werden durch OCR durchsuchbare PDF-Dateien als finale Nutzungsform im ursprünglichen Erscheinungsbild der Seite hergestellt. Je Originalseite wird eine einzelne PDF-Datei erzeugt.
- Eine durchgehende Erkennungsquote bei Normalschrift von über 99 % soll erreicht werden.
- Der unter dem PDF-Bild liegende Volltext wird in eine Indexdatenbank aufgenommen.
- Die Anzeige der Fundstelle des gesuchten Wortes erfolgt durch farbige Unterlegung im originalen Layout der Zeitungsseite.
- Die Daten werden in ein stabiles, hierarchisch strukturiertes Filesystem eingebracht. Sie sind sowohl durch ein Benutzerinterface als auch durch einfache Navigation im Filesystem abrufbar.
- Die digitalisierten Zeitungen werden in der ersten Phase sukzessive im Landesarchiv zur Nutzung bereitgestellt.

Das Projekt wurde durch die Steiermärkische Landesregierung Ende 2007 mit einem Finanzrahmen von rund 400.000,- € bewilligt und im Februar 2007 gestartet. Hard- und Software für drei Workstations und Speicherplatz wurden beschafft sowie drei Arbeitskräfte für dieses Projekt eingestellt. Erforderliche Zusatzleistungen im Programmierbereich sind budgetiert und werden zugekauft. Das Projekt soll 2009 abgeschlossen sein. Nach der Durchführung von Feinabstimmungen an Logistik, Hierarchie der Filestruktur, Scannern und Software ist der Produktionsbetrieb Mitte Februar 2007 angelaufen. Innerhalb eines Jahres wurden rund 900.000 Seiten digitalisiert. In der ersten Phase werden die Tageszeitungen ab 1945 verarbeitet, in der zweiten Phase alle älteren Zeitungen.

POVZETEK

DIGITALNE REPRODUKCIJE IN VSEBINSKO ZAJEMANJE ZGODOVINSKEGA ČASOPISJA V ŠTAJERSKEM DEŽELNEM ARHIVU

Štajerski deželni arhiv v sodelovanju z Deželno knjižnico Štajerske, časnikom Kleine Zeitung in nekaterimi drugimi partnerji oblikuje Digitalni arhiv štajerskih časnikov. Le-ta obsega zgodovinske izdaje sedmih štajerskih časnikov z okrog 2,5 mil. stranmi. Tiste izdaje, ki niso zaščitene z zakonom, bodo postale dostopne prek interneta. Vse ostale bodo za raziskovanje dostopne v čitalnicah pristojnih ustanov. Na voljo bodo originalni izgledi strani z vsemi slikami in ilustracijami. Vse indeksirane besede bodo shranjene v bazi podatkov. Zadetki iskanih besed bodo poudarjeni. Digitalizacija in proces skeniranja strani se izvajata v Štajerskem deželnem arhivu.