TEXTE IM DIGITALEN UMFELD

VOLLSTÄNDIGE DIGITALISIERUNG DES INNSBRUCKER ZEITUNGSARCHIVS

ALBERT GREINÖCKER, GÜNTER MÜHLBERGER

ABSTRACT

Die Zeitungsausschnittssammlung des Innsbrucker Zeitungsarchivs gehört zu den größten im deutschen Sprachraum und bietet umfassende Informationen zu zehntausenden von Autoren, Schauspielern und Regisseuren aus der ganzen Welt. Seit Oktober 2004 wird der gesamte Altbestand – immerhin rd. 800.000 Seiten – an der UB Innsbruck gescannt, automatisch texterkannt und im Internet der Öffentlichkeit zur Verfügung gestellt. Dieser Beitrag erläutert die technischen Hintergründe des Projekts. Das Online-Archiv ist im Testbetrieb bereits unter der Adresse: http://webapp.uibk.ac.at/iza abrufbar.

EINLEITUNG

Dieser Beitrag bezieht sich auf die konkrete Umsetzung der „Digitalisierung und des Onlineauftritts der Zeitungsausschnittssammlungs-Anwendung des Innsbrucker Zeitungsarchivs". Es sollen alle Schritte, die notwendig waren, um aus der bestehenden Zeitungsausschnittssammlung eine Online-Lösung zu implementieren, die völlig neue Möglichkeiten der Archivbenutzung (wie z.B. Volltextsuche) schafft, zu beschrieben werden. Ziel dieses Artikels soll auch ein Blick aus Sicht der Informationstechnologie auf diese Aufgabe sein. Das Projekt hat insgesamt eine Laufzeit von 2 Jahren und wird im Sommer 2006 abgeschlossen sein.

BESCHREIBUNG DES ARCHIVS

Die Inhalte des Archivs sind sehr weitschichtig: Seit 1960 sammelt das Archiv Artikel auf der Basis ausgewählter deutschsprachiger Tages- und Wochenzeitungen. Neben der klassischen Buchkritik werden auch Artikel zu allen Bereichen des literarischen Lebens aller Zeiten, Kulturen und Sprachen [1] berücksichtigt. Die einzelnen Artikel wurden bis zum Oktober 2000 aus den Zeitungen ausgeschnitten, auf A4-Blätter geklebt und eingeordnet.

Auch der Umfang des Archivs kann sich sehen lassen. Der gesamte Bestand ist in 2306 Mappen systematisch abgelegt und besteht aus etwa 600.000 alphabetisch geordneten Einzelartikeln (dies entspricht etwa 800.000 Seiten). 8 Hauptkategorien (z.B. deutschsprachige Autoren, fremdsprachige Autoren, Theaterhäuser,…) und etwa 40.000 Hauptdossiers bilden die Struktur des Archivs. Hauptdossiers können sein: Autorennamen, Künstlergruppen, Schlagworte, usw. Ein beträchtlicher Teil (etwa 40%) der Dossiers beinhalten nur 1 Artikel, manche jedoch sehr viele, z.B. Johann Wolfgang von Goethe 3376 und Günter Grass 2502. Um dieser Komplexität Herr zu werden, wurden für diese umfangreichen Dossiers noch Unterdossiers eingeführt (insgesamt etwa 9000), die eine zielgerichtetere Suche nach einem bestimmten Artikel ermöglichen.

Alle Artikel ab September 2000 werden bereits mit der Clipping-Software „LibClip" bearbeitet und in einem von dem hier beschriebenen Projekt unabhängigen online-System zur Verfügung gestellt. Es sind in diesem System zusätzliche Features wie ein Thesaurus enthalten, es bietet aber z.B. nur eine Metadatensuche und keine Volltextsuche. Der jährliche Zuwachs beträgt ca. 25.000 Zeitungsausschnitte. Das Innsbrucker Zeitungsarchiv ist somit die größte universitäre Dokumentationsstelle für journalistische Literaturkritik im deutschen Sprachraum [2].

SCANPROZESS

Am Anfang des Verarbeitungsprozesses stehen das Scannen und die damit verbundenen Aufgaben. Die Seiten werden sowohl farbig (8 bit Farbtiefe) als auch schwarzweiß mit 300 dpi gescannt. Die farbigen Scans stellen die Basis für die Anzeige der Artikel dar und die schwarzweiß-Seiten werden für die OCR-Erkennung (aufgrund der geringeren Dateigröße ist ein größerer Durchsatz möglich), für die Auslieferung als PDF (mit hinterlegtem Volltext) und als Druckvorlage verwendet.

Beim Scannen müssen auch zusätzliche Schritte erledigt werden, so etwa
- das Drehen von quergestellten Artikeln (die Überschrift muss immer in der richtigen Orientierung sein) [3] ,
- das Scannen der Rückseite, wenn diese ebenfalls mit einer Artikelseite oder einem dazugehörigen Bild beklebt wurde,
- die Eingabe der Dossiernamen.

Wenn eine Mappe mit dem Scannen abgeschlossen ist, dann wird diese auf ein Netzwerklaufwerk kopiert, wo sie vom Verarbeitungsprogramm später zur weiteren Verarbeitung abgeholt wird.

Diese zusätzlichen Schritte kosten sehr viel Zeit: Die Verarbeitung einer Mappe (~360 Seiten) dauert insgesamt 1 Stunde, davon fallen ca. 20 min. auf Scannen und 10 min. auf spezifische Tätigkeiten (z.B. Drehen der Artikel, etc.), der Rest ist Metadateneingabe (Indexeinträge) und das Holen der Mappen aus dem Archiv.

AUTOMATISIERTE WEITERVERARBEITUNG

Bei allen in weiterer Folge beschriebenen Schritten ist keinerlei Benutzerinteraktion mehr notwendig. Die Applikation überprüft selbständig, ob ein neuer Ordner für die Verarbeitung verfügbar ist, holt diesen vom Netzordner und startet die Verarbeitung.

Alle Schritte werden von einem eigens dafür entwickelten generischen Batchprozessor abgearbeitet. Diese Verarbeitungs-Engine ist mit .NET Technologie realisiert. Die einzelnen Verarbeitungsschritte sind in Prozesse aufgeteilt, die sequentiell und parallel ablaufen können. Alle Prozesse sind mittels XML-Files konfigurierbar. Mehrere Unterprozesse können zu Prozessen zusammengefasst werden. Es können sehr einfach neue Prozesse mit zusätzlicher Funktionalität hinzugefügt werden. Ein Scheduler ist für die Ablaufsteuerung verantwortlich. Es können einzelne Prozesse periodisch zu bestimmten Zeiten (also z.B. jeden Tag um Mitternacht) oder als Reaktion auf ein bestimmtes Ereignis (wie z.B. eine Veränderung in einem zur Beobachtung registrierten Verzeichnis) gestartet werden.

Das Verarbeitungsprogramm läuft völlig automatisch ab, Fehler treten sehr selten auf und werden durch entsprechende Prüfroutinen (d.h. jeder Prozess hat eine Menge von Vor- und Nachbedingungen, die erfüllt sein müssen) bereits im Vorfeld abgefangen.

Am Anfang der automatisierten Verarbeitung steht die Volltexterkennung. Verwendet wurde ABBYY Finereader Version 6.0.Scripting Edition [4]. Angesprochen wird Finereader über das COM-Interface [5]. Da das Ergebnis der Volltexterkennung/ Strukturerkennung die Basis für beinahe alle weiteren Bearbeitungsschritte darstellt, wird alles in ein selbst definiertes XML-Zwischenformat [6] geschrieben. Folgende Information ist darin enthalten: Für jedes Zeichen wird Schriftart, Schriftgröße, Formatierungseigenschaften (wie z.B. bold, italic,…), ein Unsicherheitsfaktor der Erkennung sowie die Position des Zeichens auf der gescannten Seite gespeichert. Einzelne Zeichen werden zu Absätzen und diese wiederum zu Blöcken zusammengefasst. Auch wo sich Bilder auf der erkannten Seite befinden, wird abgespeichert.

Folgende projektspezifische Probleme sind bei der Volltexterkennung aufgetaucht:

- Drehen: Beim Scannen wurde die Regel festgesetzt, dass die Überschrift immer in der richtigen Orientierung auf der Seite stehen muss. Dadurch kann es sein, dass der Flusstext in der falschen Orientierung liegt. Aus diesem Grund müssen Artikel mit quer geklebten Titeln in drei Richtungen (original, gedreht 90° nach links und 90° nach rechts) erkannt werden, da nicht vorhersagbar ist, wie die Textrichtung läuft. Nach der Texterkennung kann man die richtige Orientierung dann mit relativ geringem Aufwand bestimmen.
- Ein Stempel des IZA ist auf vielen der gescannten Seiten aufgedruckt, deshalb wird der darauf enthaltene Text in den Text des Artikels fälschlicherweise mit aufgenommen.
- Eine Regel im IZA lautet, dass Metadaten, wie die Tageszeitung, in der der Artikel erschienen ist, und das Erscheinungsdatum direkt in den Titelbereich geschrieben werden müssen, oft auch in den Titel des Artikels hinein, was das OCR-Ergebnis verschlechtert.

Durch die Abspeicherung der OCR-Information wird jeder weitere Schritt der Verarbeitung wiederholbar, ohne wieder eine OCR-Erkennung durchführen zu müssen. Auch die Transparenz bzw. die Überprüfbarkeit des Ergebnisses wird erhöht, da XML auch für den Menschen lesbar ist.

Folgende weitere Schritte wurden vom Verarbeitungsprogramm noch durchgeführt:

1. Automatische Entscheidung, ob ein Artikel ein Kurzartikel (etwa 25 % aller Artikel) ist oder nicht: Dies ist entscheidend für die Anzeige der Artikel im Web, da Kurzartikel dem Benutzer komplett gezeigt werden, von längeren Artikeln jedoch nur der Titelbereich und alle Seiten des Artikels als Thumbnails. Als kein Kurzartikel wurde ein Artikel klassifiziert, bei dem eine Überschrift vorhanden war und die Anzahl der Zeichen über einem bestimmten Schwellenwert lag [7].
2. Artikelseparierung: Da ein Artikel aus mehreren Seiten bestehen kann, aber keine Information darüber gegeben ist, wann ein Artikel anfängt bzw. endet, erfolgt der Prozess der Artikelseparierung ebenfalls automatisch. Es wurde ein einfaches Regelwerk erstellt, auf dessen Basis diese Separierung vorgenommen wird. Die Fehlerrate beträgt etwa 10%. Die Ursache für die Fehler liegt großteils im sehr unterschiedlichen Ausgangsmaterial begründet (viele unterschiedliche Zeitungen machen die Erstellung eines allgemein gültigen Regelwerkes sehr schwierig). Auch werden manche der Überschriften als Bilder erkannt (was ein häufiger Fehler bei sehr großen Schriften ist), was die automatisierte Weiterverarbeitung erschwert.
3. Erkennen und automatisches Ausschneiden der Titelregionen pro Artikel (wenn kein Kurzartikel): Die Titelregionen werden für die Anzeige eines Artikels verwendet und zeigen dem Benutzer den Titel sowie die Metadaten, die meist

in handschriftlicher Form beim Titel stehen. Die Titelregion wird über die Schriftgröße gewonnen. Textblöcke mit großen Schriften werden zu einem Titelblock zusammengefasst.

4. Volltextgenerierung: Für das Abspeichern des Volltextes wurde das Dokumentenformat TEI („Text Encoding Initiative“ [8]) verwendet. TEI ist ein in XML formulierter Standard für die elektronische Textcodierung und hat sich zu einem de-facto Standard für das wissenschaftliche Bearbeiten von digitalen Texten entwickelt. Dieser Text dient auch als Basis für die Volltextsuche im System.
5. Ausschneiden der redaktionellen Bilder: Die OCR- Software gibt Information über Bildblöcke auf einer Seite zurück. Diese einzelnen Bilder werden ausgeschnitten.
6. Datenbank-Upload: Die Dossierstruktur wird beim Einspielen in die Datenbank aufgebaut. Der Volltext wird als natives XML in der Datenbank abgelegt, um XML-Features z.B. bei der Suche nutzen zu können.

Die Dauer für die vollständige Verarbeitung, also OCR und alle damit verbundenen Arbeitsschritte, sowie das Einspielen in die Datenbank, beträgt pro Mappe (~360 Seiten) ca. 3h. Der größte „Zeitfresser“ ist dabei das dreifache OCR-Erkennen der Seiten mit unterschiedlicher Textorientierung.

WEBPRÄSENZ

Der Benutzer kann grundsätzlich auf zwei verschiedene Arten das digitale Archiv benutzen. Einerseits über einen blätternden Zugang, wo man sich entsprechend der Hierarchie (Hauptkategorie, Indexseite der Dossiers pro Buchstabe, Dossier und ggf. Unterdossier) bis zu den Artikeln selbst bewegen kann. Dieser Zugang ist für Benutzer interessant, die sich einen Überblick über den Bestand machen möchten. So wird auch die Abrufbarkeit von Artikeln ermöglicht, die über die Suche wegen schlechter OCR-Ergebnisse nicht gefunden wurden. Für Suchmaschinen wie Google bietet sich dadurch die Möglichkeit, der gesamten Hierarchie zu folgen und diese in ihrem Index aufzunehmen. Dies erweitert das Spektrum an zusätzlichen Benutzern beträchtlich [9].

Andererseits besteht die Möglichkeit, mittels einer kombinierten Metadaten- und Volltextsuche die gewünschten Artikel zu finden. Um zu gewährleisten, dass die am besten der Suchanfrage entsprechenden Artikeln weiter vorne gereiht werden, wurde ein besonderes Ranking implementiert: Ein Artikel, wo der Suchbegriff im Dossiernamen bzw. im Titel des Artikels [10] gefunden wurde, wird weiter nach

vorne gereiht als jene Artikel, bei denen der Suchbegriff nur im Flusstext enthalten ist. Wenn man sich nur für Treffer in einem bestimmten Dossier interessiert, kann man die Suche auf dieses einschränken.

Als besonderes Feature wurde auch eine Bildsuche integriert. Diese listet redaktionelle Bilder (die ebenfalls automatisch ausgeschnitten werden) auf und erlaubt so einen unkonventionellen Zugang zu den Beständen des Archivs. Insgesamt werden etwa 250.000 redaktionelle Bilder abrufbar sein.

Die Volltextindexierung basiert auf den automatisch pro Artikel generierten TEI-Dokumenten. Im Moment ist das verwendete TEI-Dokument sehr einfach gehalten, es werden nur title, p (Absatz), pb (page break) und figure (Bild) verwendet.

Diese TEI-Dateien werden direkt in der Datenbank als XmlType abgelegt, was eine Reihe von Vorteilen mit sich bringt:

- Man kann eine Gewichtung vornehmen, basierend auf den Tags (d.h. man kann einen Titel höher gewichten als einen Treffer im Flusstext (Oracle bietet sehr viele nützliche Features für die XML-Bearbeitung an).
- Man könnte z.B. nur bestimmte Bereiche des Textes dem Benutzer zur Verfügung stellen (z.B. nur den/die Titel oder nur den ersten Absatz).
- Ein einfaches XSL-Stylesheet reicht aus, um den gesamten Text in unterschiedlichen Skinnings und Formaten darzustellen (html, pdf, …).

Die einzelnen Artikel werden folgendermaßen präsentiert: Links werden alle Seiten, die zu einem Artikel gehören, als Thumbnails angezeigt, um einen Überblick zu verschaffen. Die Titelregion des Artikels wird gesondert ausgeschnitten und rechts der Thumbnails angezeigt.

Um wiederum die Trefferquote der Suchmaschinenabfragen zu erhöhen, wurden Metadaten über den Artikel als Dublin Core Einträge in den HTML-Code integriert, z.B.:

```
<meta name="DC.creator" content="Innsbruck Zeitungsarchiv;IZA">
<meta name="DC.title" content="Kafka lebt"/>
<meta name="DC.title" content=" Drei zeitgenössische Autoren suchen nach neuen
  Zugängen zum Prager Dichter Franz K."/>
<meta name="DC.language" content="de"/>
<meta name="DC.subject" content="deutschsprachige literatur"/>
<meta name="dc.relation.ispartof" content="http://www.iza.uibk.ac.at">
<meta name="dc.generator" content="Abteilung für Digitalisierung;DEA">
```

Weiters ist ein Bestellsystem mit Warenkorb-Funktionalität integriert. Einzelne Artikel oder Seiten können markiert und ausgewählt werden. Nach dem Eintragen persönlicher Daten werden die Artikel als Papierkopie (bzw. Ausdruck des PDFs) zugeschickt. Es besteht auch die Möglichkeit, die Artikel eines gesamten Dossiers zu bestellen.

Bei der Implementierung wurde auf Sprachunabhängigkeit geachtet, es ist also sehr leicht möglich, eine weitere Sprache für die Benutzeroberfläche einzuführen. Man muss nur eine Sprachtabelle, die alle Texte der Benutzeroberfläche enthält, exportieren, einem Übersetzer übergeben und wieder importieren. In der Applikation kann man dann die gewünschten Sprachen einstellen. Das Einbinden redaktioneller Texte wurde mit dem hauseigenen XML-Content Managementsystem XIMS [11] realisiert. Das System lässt das Editieren von HTML-Inhalten in WYSIWYG-Manier zu. Diese Texte werden dann in das IZA-System als XHTML importiert und, den eigenen Bedürfnissen angepasst, angezeigt.

Um weitere Nachbearbeitungen der bereits in das System eingespielten Inhalte vornehmen zu können, Daten zu exportieren und das Bestellsystem zu bedienen, wurde ein Administratorbereich eingerichtet, der Folgendes zulässt:

- Bilder löschen: Da beim automatischen Ausschneiden u. U. auch Bilder ausgeschnitten werden, die sich für die Darstellung im Web nicht eignen, ist es notwendig, nachträglich diese Bilder löschen zu können.
- Edieren der Dossiernamen: Wenn bei der Eingabe der Dossiernamen ein Fehler passiert ist, kann dieser leicht geändert werden.
- PDF-generieren on the fly: Wenn man für den geschützten Bereich angemeldet ist, kann man Seiten auswählen, aus denen ein gesamt-PDF generiert werden soll und dieses downloaden.
- Bestellsystem:
  - o Alle Bestellungen werden in einer Liste dargestellt und mit einem Status versehen.
  - o Ein einfacher Klick reicht aus, um die bestellten Seiten als pdf zu erhalten, die dann für den Versand nur mehr ausgedruckt werden müssen.
  - o Bestellstatistiken werden bereitgestellt.
  - o Automatische Rechnungsgenerierung.
- Artikelseparierungs-Korrektur: Wenn bei der automatischen Artikelseparierung ein Fehler passiert ist, d.h. wenn z.B. eine Seite, die bereits eine Startseite eines neuen Artikels ist, noch dem Vorgängerartikel zugewiesen wird oder wenn ein eigenständiger Artikel als Teil eines anderen Artikels erkannt wurde, können diese Fehler behoben werden.

Die Webanwendung, der Filestore sowie auch die Datenbank werden vom zentralen Informatikdienst der Universität Innsbruck gehostet.

CHARAKTERISIERUNG DES DATENBESTANDES

Noch einmal als kurze Zusammenfassung, was alles gespeichert wird:

Pro Seite:
- Original Jpeg
- Original Tif (Group 4)
- Volltexthinterlegtes Pdf
- Finereader-Output als XML
- Thumbnails als Jpeg für die schnelle Anzeige

Pro Artikel:
- Ausgeschnittene redaktionelle Bilder, die im Artikel enthalten sind
- TEI-File (speichert den Volltext mit einer einfachen semantischen Auszeichnung (Titel, Bild, Absatz, Seitenumbruch)
- Titelregion als Jpeg

Die große Mehrheit (etwa 3/4) der Artikel besteht nur aus einer Seite, 17% aus 2 Seiten, und der Rest (etwa 7-8%) bestehen aus 3 oder mehr A4 Seiten.

Erwartetes Datenvolumen:
- Datenbank: mehrere hundert MB (der Volltextindex wird relativ groß)
- Filestore: ca. 1,5 TB

AUSBLICK

Ab Sommer 2006 wird das gesamte Archiv elektronisch verfügbar sein.

Die Erkenntnisse sowie auch die Software, die aus diesem Projekt entstanden ist, sollen auch für andere, ähnlich organisierte Archive eingesetzt werden. Zurzeit wird die Technologie dazu eingesetzt, eine Nachbearbeitung des Bestandes der Digitalen Bibliothek alo („Austrian Literature Online“ [12]) vorzunehmen. alo besteht im Moment aus etwa 700.000 gescannten Buchseiten, die alle texterkannt und auf ähnliche Weise verarbeitet werden wie die Einzelseiten des IZA.

Künftige Verbesserungen könnten umfassen:

- Unicode: die zurzeit verwendete Oracle Datenbank (V9.2) unterstützt Unicode noch nicht. Da der Volltext der einzelnen Artikel aber direkt in der Datenbank liegt, können darin einzelne Zeichen (z.B. Anführungszeichen unten) nicht als solche gespeichert werden und gehen deshalb verloren. Die XML-Dateien, die beim Volltexterkennen generiert wurden, enthalten unter anderem auch den gesamten Volltext (diese werden extern auf einem Filestore abgelegt).
- Erweiterung der Suchmöglichkeiten um die Suchmöglichkeiten, die Oracle Text ohnehin zur Verfügung stellt.
  o Wildcardsuche
  o Unscharfe Suche
- Integration einer elektronischen Bezahlungsmöglichkeit
- Ein Versuch könnte unternommen werden, auf Basis des generierten Volltextes Dubletten, die es im gesamten System gibt, zu identifizieren.

ANMERKUNGEN

1 Mehr Information befindet sich unter: http://iza.uibk.ac.at
2 http://iza.uibk.ac.at
3 Dadurch ist aber keinesfalls sicher gestellt, dass die Artikel in der richtigen Orientierung liegen. Dies wird mit dem Ergebnis der OCR-Engine erkannt.
4 Mit der Zeit hat sich die Qualität der Finereader-Produkte wesentlich verbessert, also hätten mit einer neueren Version bessere Ergebnisse erzielt werden können.
5 COM- Component Object Model. Für die Verwendung gemeinsam genutzter Komponenten unter Windows.
6 Finereader Version 7 bietet zwar die Möglichkeit, gegen zusätzliche Lizenzkosten die Ergebnisse im XML-Format zu exportieren, diese sind aber nicht detailliert genug.
7 Es wurden aber noch mehr Kriterien für die Entscheidung herangezogen.
8 http://www.tei-c.org
9 In einem ähnlichen Projekt, der Zettelkatalogsanwendung der UB Innsbruck (http://webapp.uibk.ac.at/alo/cat) mit rund 2 Mio. Katalogkarten, die OCR-erkannt wurden, konnten in dieser Hinsicht gute Ergebnisse erzielt werden.
10 Die Information über die Schriftgröße liefert die OCR-Software. Aufgrund dieser wurde entschieden, ob ein bestimmter Text eine Überschrift ist oder nicht.
11 http://xims.info
12 http://www.literature.at

ADRESSE DER AUTOREN

Albert Greinöcker, Günter Mühlberger
Universitätsbibliothek Innsbruck
Abteilung für Digitalisierung und elektronische Archivierung - DEA
A-6020 Innsbruck, Innrain 52A
E-Mail: albert.greinoecker@uibk.ac.at, guenter.muehlberger@uibk.ac.at
Web: http://www.uibk.ac.at/ub/dea/