Neue Entwicklungen: Österreichische Nationalbibliothek

Christa Müller*

ANNO – Der digitale Zeitungslesesaal der Österreichischen Nationalbibliothek

Aktuelle und zukünftige Entwicklungen im Überblick

DOI 10.1515/bfp-2016-0012

Zusammenfassung: Die Österreichische Nationalbibliothek (ÖNB) kann mit ihrem Digitalen Lesesaal für historische Zeitungen auf mehr als 12 Jahre Erfahrung zurückblicken. Die Integration der Volltextsuche wurde 2013 begonnen und wird bald vollständig implementiert sein. Die Autorin erklärt kurz den aktuellen Funktionsumfang von ANNO und widmet sich dann dem Nutzerverhalten. Die vorliegende Analyse wird für die Weiterentwicklung des Angebots und der Applikation verwertet.

Schlüsselwörter: Digitalisierung; Zeitungen; Volltextsuche

ANNO – The Digital Newspaper-Reading Room of the Austrian National Library. An Overview of Current and Future Developments

Abstract: The Austrian National Library can look back at 12 years of experience with its digital newspaper reading room. 2013, the integration of full text search started and will soon be finished. The author gives a short overview of the functionalities of ANNO and then concentrates on the user behaviour. This analysis is the base for further development of the application as well as the content of the ANNO project.

Keywords: Digitisation; newspapers; full text search

1 ANNO – Wie es dazu kam

An der ÖNB sollte Ende 2002 das Mikroverfilmen von Zeitungen neu ausgeschrieben werden und dabei wurde – damals ein innovativer Ansatz – ein Hybridverfahren gewählt, bei welchem nicht nur Mikrofilme, sondern im gleichen Schritt auch Digitalisate erstellt wurden. Das Scannen

von Zeitungen an einen externen Dienstleister auszulagern, hat sich als gute Entscheidung erwiesen und wird seitdem beibehalten. Eine Million und mehr Seiten pro Jahr in der Bibliothek zu scannen, würde sowohl personell als auch finanziell eine große Belastung bedeuten. Der externe Dienstleister besitzt mehrere Geräte für verschiedene Scanverfahren mit unterschiedlichen Größen. Da in der Ausschreibung verlangt wird, dass alle Nebenkosten (Transport, eventuell Schneiden der Bände, Datenträger und dergl.) in den Seitenpreis eingerechnet werden, kann mit diesem gut und verlässlich kalkuliert werden.

Um die Scans den Lesern zugänglich zu machen, wurde Anfang 2003 die Applikation „ANNO – AustriaN Newspapers Online“ – <http://anno.onb.ac.at/> – an der ÖNB durch eigene Mitarbeiter entwickelt.

Wenn auch pragmatische Überlegungen am Projektbeginn im Vordergrund standen, so gab es noch weitere Gründe, die gerade Zeitungen als geeignetes Material für ein erstes großes Bestandsdigitalisierungsprojekt erscheinen ließen. Zeitungen bestehen als allgemein bekanntes Medium noch heute, sie haben sich in Layout und Aufbau kaum verändert und sind daher für jeden historisch Interessierten einfach zu lesen. Zu erwarten stand also, dass historische Tageszeitungen online eine große Leserschaft ansprechen würden. Daneben sind sie für Wissenschaftler verschiedenster Fachrichtungen interessant, die sie als Quelle heranziehen können.

Ziel des Projekts ist es, Zeitungen besser verfügbar zu machen. Dies ist nicht nur den Nutzern geschuldet, sondern erfüllt zudem auch bibliothekarische Anliegen. Besonders Zeitungen, die zwischen 1850 und der Mitte des 20. Jahrhunderts erschienen, bestehen aus brüchigem, vom Zerfall bedrohten Papier. Indem digitale Surrogate erstellt werden, soll es – wie schon bei der früher üblichen systematischen Mikroverfilmung – zu einer Schonung der analogen Originale kommen. Auch wenn die Digitalisierung die schon bestehenden konservatorischen Probleme der Bibliotheken damit nicht lösen kann, ist zumindest die intensive Nutzung der Bestände deutlich eingeschränkt.

*Kontaktperson: Mag. Christa Müller, christa.mueller@onb.ac.at

Da die Zeitungen, sobald sie digitalisiert sind, nur mehr eingeschränkt ausgehoben, Lesern vorgelegt und reproduziert werden müssen, kann für die Zukunft der Bestand geschont werden. Dieser Vorteil gilt nicht nur für die digitalisierende Bibliothek, auch andere Sammlungen können die Zeitungen, die frei online zugänglich sind, aus der Benutzung nehmen.

Natürlich gibt es populäre Titel, die, wie z. B. die „Wiener Zeitung“ oder die „Reichspost“, allein innerhalb der Stadt Wien in mehrfacher Ausführung gebunden in Bibliotheksmagazinen vorhanden sind, doch sind zahlreiche ältere Ausgaben oft nur noch ein- oder zweimal nachgewiesen. Insbesondere in Kriegsjahren sind häufig große Verluste eingetreten, so dass nur noch ganz wenige Exemplare erhalten sind. Nachdem kaum eine Bibliothek lückenlos die Ausgaben aller Erscheinungstage besitzt, bietet sich durch die Digitalisierung erstmals die Möglichkeit einer echten Gesamtschau eines Titels. Durch die Kooperation von mehreren Institutionen kann virtuell eine Zeitungssammlung entstehen, die vollständiger ist als sie irgendeine Bibliothek physisch im Bestand hat. Der Zugriff auf verteilt vorhandene Ausgaben wird erleichtert, es entstehen integrierte Sammlungen, die in ihrer Zusammenführung größeren Nutzen haben als ihre Einzelteile.

2 ANNO – Wie es heute funktioniert

Zieht man Zeitungen als Quelle für die Recherche heran, dann gibt es zwei mögliche Einstiegspunkte. Man weiß entweder das Datum von einem Ereignis, zu welchem man die Rezeption in den Medien lesen möchte, oder man hat einen Begriff, einen Familien-, Firmen- oder Ortsnamen, zu dem man Näheres erfahren möchte. Dazu kann eine Einschränkung auf bestimmte Zeitungstitel, eine Periode von Jahren oder eine Region das Finden relevanter Treffer erleichtern. Beide Möglichkeiten bietet ANNO: Von Anfang an gab es den Einstieg über den Kalender, die Volltextsuche kam vor einigen Jahren dazu. Wie wir aus der Nutzung sehen, löst die Volltextsuche nicht den Einstieg über das Datum ab.

2.1 Die Blätterversion

Zu Beginn der Zeitungsdigitalisierung gab es eine sorgfältige Suche nach vorhandenen Lösungen. Diese waren entweder zu teuer und aufwändig in der Betreuung oder die Entwicklungen waren mit großem Kosten- und Zeitaufwand gleich zu Beginn verbunden. Es zeigte sich, dass bei einer sehr genauen und auf das Notwendigste be-

geschränkten Anforderung eine Eigenentwicklung der geringste Aufwand und die sicherste Investition war. Die so gegebenen funktionalen Erweiterungsmöglichkeiten und die fast beliebige Skalierbarkeit in der Zukunft bestärkten diese Entscheidung.

Hinter der Blätterversion von ANNO steht keine klassische Datenbank. Die Logik ist ausschließlich in der Verzeichnis- und Dateistruktur bzw. -benennung enthalten. Diese wird vom Dienstleister bereits beim Scannen erstellt. Beim Laden der Daten erfolgt nur die Kontrolle der logischen Strukturen und der technischen Lesbarkeit.

Gespeichert werden auf dem Server lediglich die Originalscans. In der Applikation werden aber verschiedene Formate und Größen angeboten. Die Umwandlung erfolgt on-the-fly.

2.2 Die Volltextsuche

Das große Desiderat für die Erforschung historischer Zeitungen war deren textliche Durchsuchbarkeit. Sie haben aber einige Eigenschaften, welche die elektronische Texterkennung sehr erschweren. Oft wurde Papier von minderer Qualität verwendet, was nach so vielen Jahrzehnten und Jahrhunderten zum Verbräunen geführt hat und damit einen geringeren Kontrast zwischen Schrift und Papier bewirkt. Die Druckerschwärze wurde oft mit weniger Sorgfalt aufgebracht als bei teuren Büchern. Für die Texterkennung ist es wichtig, dass die Schrift auf dem Scan nicht zu fett und nicht zu mager wiedergegeben wird, um ein gutes Resultat zu erreichen. Bei Büchern lässt sich das sehr gut für das ganze Dokument mit mehreren hundert Seiten einstellen. Bei Zeitungen ist es allerdings so, dass oft innerhalb einer Seite die Druckerschwärze unterschiedlich intensiv aufgetragen wurde. Links oben ist die Schrift z. B. mager, rechts unten sehr fett. Auch dies ist ein großes Problem für die Texterkennung. Bei Büchern läuft der Text in den meisten Fällen einspaltig von links nach rechts. Zeitungen sind mehrspaltig und oft sind pro Spalte nicht mehr als drei oder vier Wörter gesetzt. Um einen schönen Blocksatz zu erreichen wurde sehr oft abgeteilt. Ein Teil der Texterkennungssoftware ist die vorgeschaltete Layoutanalyse. Deren Aufgabe ist es, die Textblöcke korrekt zu identifizieren. Geschieht dies nicht ordentlich, so versucht das Texterkennungsprogramm über die Spalten hinweg den Text zu lesen und verliert damit jedes abgeteilte Wort, was bei Zeitungen einen sehr hohen Prozentsatz der Wörter ausmachen kann.

Texterkennungssoftware wurde ursprünglich für den Einsatz als Bürosoftware entwickelt und ist daher auf heutige Druckschriften optimiert. Im nicht-deutschsprachi-

gen Raum wurde auch früher schon die noch heute gebräuchliche Antiqua-Schrift verwendet. Im deutschsprachigen Raum war aber bis in die 30er-Jahre des 20. Jahrhunderts für deutschsprachige Texte die Frakturschrift in Verwendung. Diese Schrift ist aufgrund ihres optischen Charakters – dicke senkrechte Schattenstriche und schräge oder waagrechte, sehr dünne Haarstriche – per se für die elektronische Erfassung eine größere Herausforderung, da besonders die Haarstriche oft so zart sind, dass sie abreißen, und damit die Software die Buchstaben nicht mehr identifizieren kann.

Jede Texterkennung hat Lexika integriert, um erkannte Zeichenfolgen auf ihre Plausibilität zu prüfen. Dies funktioniert bei heutigen Bürotexten mit ihrem doch begrenzten Vokabular sehr gut. Wonach wird aber in Zeitungen gesucht? Der historisch Interessierte sucht häufig nach Personen- und Eigennamen sowie Ortsbezeichnungen, genau diese finden sich aber in den üblichen Lexika nicht. Genau mit diesen Herausforderungen hat sich vor einigen Jahren das EU-Projekt IMPACT¹ beschäftigt. Bei der Volltextsuche in Zeitungen stellt sich die Frage, welches die Einheit ist, in der gesucht wird: die Seite oder der Artikel. Dies hängt von der Periode ab, aus welcher das Material stammt. Erst um 1900 beginnt sich die heute übliche, optisch erkennbare Abgrenzung von Artikeln durchzusetzen. Bis dahin ist der Text meist ein fortlaufender Fließtext, der von Thema zu Thema überleitet. Die fett gedruckten Zeilen dazwischen sind keine Überschriften, sondern stehen mitten im Bericht zu diesem Thema. Daher ist die Segmentierung in Artikel als Vorarbeit für eine Suche innerhalb eines Artikels erst für spätere Perioden sinnvoll. Um aber relevante Treffer beim Suchen mit mehreren Begriffen zu finden – was nicht der Fall ist, wenn die Treffer auf der Seite weit auseinander liegen – wurde die Abstandssuche (dazu später) entwickelt.

ANNO war vom ersten Tag an bestrebt, eine Volltextsuche zu ermöglichen. Da die Entwicklung einer derartigen Suchfunktion zu Beginn des Projektes ein zu aufwändiges und kostenintensives Vorhaben darstellte, dauerte es fast zehn Jahre bis zu seiner Realisierung. Dabei stand man vor zwei großen Herausforderungen: Erstens musste die Texterkennung für Fraktur eine ausreichend gute Qualität haben und die Kosten dafür im Rahmen bleiben. Die zweite wichtige technische Komponente war die Suchmaschine.² Ersteres wurde durch das EU-Projekt „Europeana Newspa-

pers“³ gefördert. Zweiteres wurde möglich, als die Implementierung der volltextlichen Durchsuchbarkeit historischer Textbestände als Meilenstein in der Strategie der Bibliothek definiert wurde. Vor allem im Rahmen des Projektes „Austrian Books Online“ wurden seit 2012 Texte OCR-gelesen⁴, ein Volltextindex aufgebaut und die Suche in Volltexten implementiert. Im Frühjahr 2013 ging die ANNO-Suche online. Wie schon bei anderen Digitalisierungsprojekten an der ÖNB entschied man sich für einen stufenweisen Ausbau des Service. Da der Volltextkorpus mit automatischer Texterkennung erstellt wird, weisen die Texte eine unterschiedlich hohe Fehlerdichte auf. Nutzern wird daher empfohlen, ihre Suche auf mehrere Schlüsselbegriffe zu erweitern. Oft sind aber auch zu viele Treffer für die Suchenden ein Problem. Um relevante Ergebnisse zu bekommen, empfiehlt die ANNO-Hilfe beispielsweise, Familienname und Ortsname, Familien- und Firmenname, Firmen- und Ortsname, Orts- und Regionsbezeichnung zu kombinieren. Da diese Begriffe im Text aber oft nicht direkt nebeneinander stehen, führt eine Phrasensuche zu ungenügenden Ergebnissen. Da es keine verlässliche Methode gibt, auf Artelebene zu suchen, wurde als weiteres Feature eine „Abstandssuche“ integriert. Nach der Eingabe von zwei Begriffen im Suchfeld kann zusätzlich eine Zahl (z. B. ~10) eingegeben werden, um die beiden Suchbegriffe unabhängig von ihrer Reihenfolge in einem Abstand von maximal zehn Wörtern zu finden. Dies soll helfen, nicht nur möglichst viele, sondern möglichst relevante Treffer anzuzeigen.

ANNO hat entschieden, den OCR-gelesenen Text, mag er auch noch so fehlerhaft sein, immer anzuzeigen. Ist das OCR-Ergebnis einigermaßen gut, kann mit der Zugänglichmachung dieses Textes die sonst durch die Frakturschrift besonders bei einem jüngeren Publikum bestehende Barriere überwunden werden. Die Transparenz in Bezug auf das OCR-Ergebnis dient dazu, den Lesern zu zeigen, dass – anders als sie es von normalen Internetangeboten gewohnt sind – viele Texterkennungsfehler einen anderen Umgang mit den Ergebnissen nötig macht.

3 ANNO – Was es heute enthält

Der Zeitraum, welchen ANNO abdeckt, reicht mit gedruckten Zeitungen und Zeitschriften von 1700 bis 1944 (das Jahr 1945 wird im Januar 2016 zugänglich gemacht). Von

¹ Siehe: <http://www.impact-project.eu/>.

² Siehe dazu: Kann, Bettina; Hintersonleitner, Michael: Volltextsuche in historischen Texten. Erfahrungen aus den Projekten der Österreichischen Nationalbibliothek. In: BIBLIOTHEK – Forschung und Praxis. 39(1) (2015) S. 73–79.

³ Näheres zu dem Projekt auf der Projekt-Homepage: <http://www.europeana-newspapers.eu/>.

⁴ OCR = Optical Character Recognition, ein Verfahren zur Texterkennung.

Abb. 1: Anzahl der gescannten Seiten pro Kalenderjahr

1568 bis 1605 sind handgeschriebene Fugger-Zeitungen⁵ verfügbar. In den nächsten Monaten werden mehrere Neue Zeitungen aus dem 17. Jahrhundert hinzukommen. Unter Zuhilfenahme gedruckter Zeitungen kann derzeit das Geschehen aus 243 Jahren erforscht werden. In 23 Jahren davon stehen je mehr als 100 Zeitungstitel zur Verfügung, in über 120 Jahren je mehr als 30 Titel. Der Tag mit den meisten Zeitungsausgaben ist der 15. Juli 1916, an dem 73 verschiedene Zeitungen verfügbar sind. An über 800 Tagen gibt es 50 oder mehr Zeitungen. An mehr als 60000 Tagen ist mindestens eine Zeitungsausgabe verfügbar. Insgesamt sind es über 900000 Ausgaben. Diese verteilen sich auf 480 Zeitungs- und 360 Zeitschriftentitel.

Der umfangreichste Titel innerhalb von ANNO ist – nach Seiten und Ausgaben – die „Wiener Zeitung“ mit mehr als einer Mio. Seiten und 55000 Tagesausgaben. An zweiter Stelle folgt die „Neue Freie Presse“ mit über 600000 Seiten und 26000 Ausgaben. In der Rangliste der Titel mit den meisten Ausgaben folgen das Vorarlberger Volksblatt, die (Linzer) Tages-Post, das (Neuigkeits-)Welt-Blatt, das Linzer Volksblatt, das Prager Tagblatt, die Innsbrucker Nachrichten, Die Presse, die Vorarlberger Landes-Zeitung, das Bregenzer/Vorarlberger Tagblatt und das Prager Abendblatt.

Insgesamt sind 15 Mio. Seiten online zugänglich. Aufgrund des EU-Projektes „Europeana 1914–1918“⁶, für welches vermehrt Bestände aus der Zeit des Ersten Weltkriegs gescannt wurden, sind dies die materialreichsten Jahre.

4 ANNO – Wie es heute genutzt wird

4.1 Welche Interessen haben die Leser?

Die Zugriffszahlen von ANNO bestätigen das große Interesse am Online-Angebot historischer Zeitungen. Anhand der Anfragen zeigt sich, dass nicht ausschließlich Zeitungs- und Pressehistoriker dieses Material verwenden. Vielmehr ist es ein großer Kreis von Interessierten, die sich unterschiedlichen Fragestellungen zuwenden. Zu den Lesern gehören Genealogen und Prosopographen, die an Sterbe- und Tauflisten interessiert sind ebenso wie Heimatforscher, die ganze Jahrgänge nach lokalen Ereignissen durchsuchen. Musikjournalisten finden Konzert-, Opern- und Theaterberichte zu Uraufführungen. Frauenromane, Fortsetzungsromane und auch Werbung werden von Genderforschern als Quelle herangezogen. Kostümbildner sind an Modezeitschriften und Abbildungen in Frauenzeitschriften interessiert, Kunsthistoriker und Architekturhistoriker an Architektur- und Bauzeitschriften, viele Sportvereine und Sporthistoriker suchen Berichte zur Vereinsgeschichte und zu spezifischen Großereignissen, während Technikbegeisterte historische Fotos von Lokomotiven oder Lastkraftwagen suchen.

Immer wieder weisen uns Autoren darauf hin, dass sie ihre Forschung nur mit Hilfe von ANNO machen konnten. So erfahren wir auch von ungewöhnlichen Themen, die mit Hilfe der gescannten Zeitungen bearbeitet werden. Thomas Hofmann z.B. schrieb in der „Wiener Zeitung“ über das „Dauerthema Staubbelaftung“ im Wien des 19. Jahrhunderts⁷ und in der „Presse“ über „Klein-Kalifornien: Glücksritter wühlten im Donaukanal“.⁸ Frank Gebert

⁵ Siehe: <http://www.univie.ac.at/fuggerzeitungen>.

⁶ Siehe: <http://www.europeana1914-1918.eu/>.

⁷ Wiener Zeitung vom 13.8.2015: http://www.wienerzeitung.at/the-men_channel/wz_reflexionen/vermessungen/768574_Dauerthema-Staubbelaftung.html.

⁸ Die Presse vom 11.10.2015: http://diepresse.com/home/zeitgeschichte/4840494/KleinKalifornien_Glueksritter-wuehlten-im-Donaukanal.

hat uns informiert, dass er dank der Suchmöglichkeit in Zeitungen für die Biografie „Die Kriege der Viktoria Savs“⁹ viel zum Bild von Savs in der Öffentlichkeit gefunden hat.

4.2 Woher kommen die Leser?

Die bei weitem meisten Einstiege in die Applikation erfolgen durch direkten Aufruf der URL anno.onb.ac.at. Innerhalb der 300 häufigsten Treffer, d. h. Seiten, die mit einem Link auf ANNO verweisen, sind 48 Wikipedia-Beiträge. Bei vielen Einträgen zu historischen Ereignissen wird auf Berichte in den Zeitungen verlinkt, z. B. bei der Weltausstellung 1873 in Wien, dem Ringtheaterbrand in Wien am 8. Dezember 1881, beim Skandalkonzert am 31. März 1913 im Musikvereinssaal in Wien unter der Leitung von Arnold Schönberg, beim Amoklauf in Bremen im Jahr 1913 und dem Attentat von Sarajevo am 28. Juni 1914. Ebenso werden Zeitungsbeiträge bei Wikipedia-Artikeln zu historischen Persönlichkeiten als Quellen herangezogen: bei Elisabeth von Österreich, Kaiser Franz Joseph und weiteren Mitgliedern der Familie Habsburg-Lothringen, bei Alexander Roda Roda, Maria von Vetsera, Franz Schubert und vielen anderen, mehr oder weniger bekannten Personen. Aber auch bei Beschreibungen von Gebäuden werden passende Zeitungs- und auch Zeitschriftenartikel angeführt. Weiter finden sich noch Verweise bei Artikeln wie z. B. denen über die Österreichische Eishockey-Meisterschaft 1931/32, Milchproduktion, Volkszählung in Österreich-Ungarn 1900, Diamantidurum, Sexualethik und Straßenbahn Baden.

Aufgrund der zahlreichen Verweise in der freien Onlineenzyklopädie Wikipedia, die sich auf die Angebote von ANNO beziehen, hat die ÖNB inzwischen selbst damit begonnen, zu allen wichtigen Zeitungen in ANNO auch Artikel in Wikipedia einzustellen. Als Arbeitshilfe kann dabei dienen, dass in einem eigenen Eintrag zu ANNO alle Zeitungen, zu denen es bereits Wikipedia-Artikel gibt, aufgeführt werden und verlinkt sind.¹⁰

4.3 Wie viele Leser nutzen ANNO und wie lange bleibt ein Leser im Digitalen Zeitungslesesaal?

Täglich nutzen durchschnittlich 2500 Leser ANNO. 10 % bleiben länger als eine Stunde, weitere 10 % zwischen

einer halben und einer Stunde. Insgesamt 46 % der Besuche dauern länger als 15 Minuten. Der Durchschnitt über 660000 Besuche bisher im Jahr 2015 sind 13 Minuten pro Besuch.

21 % der ANNO-Leser nutzen auch die Volltextsuche, der Rest nutzt nur die Blätterversion. Dies mag derzeit noch damit erklärt werden können, dass die sehr wichtigen 20er- und 30er- Jahren des 20. Jahrhunderts noch nicht textlich durchsuchbar sind. Nutzer der Volltextsuche sind doppelt bis dreimal so ausdauernde Leser, d. h. sie bleiben länger im digitalen Lesesaal. 60 Seiten werden pro Besuch durchschnittlich angesehen.

4.4 Wie viele Seiten oder Ausgaben werden pro Besuch gelesen?

Es ist schwierig, für digitale Angebote Vergleichszahlen mit der analogen Welt zu liefern. Die Anzahl der Aushebungen in der klassischen Bibliothek kann mit Vollanzeigen einer Ausgabe pro IP-Adresse und Tag verglichen werden. Nach dieser Rechnung wurden im Jahr 2014 knapp 4,1 Mio. Zeitungsausgaben in ANNO ausgehoben, was 460 ausgegebenen Zeitungen pro Stunde entspricht. Damit kann gezeigt werden, dass der Digitale Lesesaal eine Intensität und einen Umfang an Nutzung ermöglicht, der in der analogen Bibliothek nicht zu bewältigen wäre.

Abb. 2: Anzahl der Aushebungen im jeweiligen Kalenderjahr

5 Die Volltextsuche – eine erste Analyse der Nutzung

Um die Nutzung der Volltextsuche besser zu verstehen und in Zukunft verbessern zu können, ist es notwendig zu wissen, wonach die Leser suchen. Dafür wurden die häufigsten 300 Suchwörter kategorisiert. Dies ergab, dass 30–

⁹ Gebert, Frank: Die Kriege der Viktoria Savs: Von der Frontsoldatin 1917 zu Hitlers Gehilfin. Wien: Kremayr & Scheriau 2015.

¹⁰ [http://de.wikipedia.org/wiki/Anno_\(Austrian_Newspapers_Online\)](http://de.wikipedia.org/wiki/Anno_(Austrian_Newspapers_Online)).

40 % der Suchen (je nach Referenzgröße) Personennamen und weitere 30 % Ortsnamen sind. Neben den Orts- und Personennamen sind die häufigsten fünfzehn Suchwörter Taufe, Gendarmerie, Juden, Zahnarzt, Orgel, Feuerwehr, Fußball, Erdbeben, Selbstmord, Schach, Kaiser, Zigeuner, Prostitution, Mord und Tramway.

Indem 70 % der Suchwörter Eigennamen sind, zeigt sich eine große Herausforderung. Genau diese Wörter sind in den Lexika, welche die Texterkennung nutzt, nicht enthalten. Damit besteht gerade bei so stark nachgefragten Begriffen das Problem, dass bei diesen die OCR-Ergebnisse fehlerhafter sein werden und als Konsequenz nur ein Teil der relevanten Stellen gefunden wird. Es besteht aber gleichzeitig auch die Hoffnung, dass relevante Wörter in einem Artikel mehrmals vorkommen und zumindest einmal exakt erkannt werden. Ob vorhandene Schlagwortlisten zur Unterstützung der Texterkennung eingesetzt werden können oder ob erst bei der Suche damit eine Unterstützung möglich ist, muss erst getestet werden. Da die Volltextsuche und der dazugehörige Index noch recht neue, aber doch sehr umfangreiche Bereiche sind, können noch weitere Optimierungen erhofft werden.

6 ANNO – Wie es weitergeht

Anfang 2016 soll das ganze ANNO – von 1700 bis 1944 – textlich durchsuchbar sein. Ab dann ist geplant, dass alle neuen Titel immer auch gleich textlich durchsuchbar eingebunden werden.

Da wir von vielen Lesern wissen, dass sie nach Abbildungen, Karikaturen und Fotos in den Zeitungen suchen, laufen erste Tests dazu, ob es möglich ist, diese Bilder automatisch zu identifizieren und dann suchbar zu machen. Es könnten z. B. alle Abbildungen aus einem Zeitschriftentitel oder aus einem Erscheinungsjahr angezeigt werden.

Die ÖNB hat mit dem Erfassen von Aufsätzen im Katalog schon gute Erfahrungen gemacht. Aufsatzsammlungen, die neu in die Bibliothek kommen, werden ausgewertet. Bei diesen wird das Inhaltsverzeichnis im Zuge der Buchbearbeitung gescannt, durch einen Dienstleister abgetippt und in den Katalog importiert. Statistische Auswertungen haben gezeigt, dass diese Bücher fünf Mal öfter bestellt werden als andere. Da der Nutzer dann bereits recht genau weiß, was ihn erwartet, kann davon ausgegangen werden, dass auch die Chance für die Leser größer ist, vorwiegend relevante Bücher zu bestellen. ANNO hat für drei Zeitschriftentitel die Aufsatzsuche schon realisiert – bisher allerdings noch in einer eigenen Datenbank und noch nicht im Österreichischen Verbundkatalog, dies soll

aber bald folgen. Indem man durch die Inhaltsverzeichnisinformationen weiß, wo ein Artikel beginnt und endet, kann der Artikel die logische Einheit sein, die in der Volltextsuche indiziert wird. Hat man diese Informationen nicht, müsste das Heft (mit mehreren logischen Einheiten, den Aufsätzen) oder der Jahrgang indiziert werden. In der Volltextsuche ist es aber vorzuziehen, dass eine logische/inhaltliche Einheit indiziert wird, da relevante Wörter in dieser Einheit prozentuell häufiger vorkommen und damit im Ergebnisranking der Artikel weiter vorne gereiht wird. Folgendes Beispiel kann das verdeutlichen. Ist in einem dreihundertseitigen Jahrgangsband einer Architekturzeitschrift ein zwanzigseitiger Artikel zum Thema Kettenbrücken erschienen, dann kommt das Wort Kettenbrücke innerhalb der zwanzig Seiten z. B. dreißigmal vor. Wäre der Jahrgangsband die indizierte Einheit, wäre die Dichte der Treffer bezogen auf die Gesamtmenge der Wörter geringer als im Artikel und damit wäre der Band weiter hinten in der Ergebnisliste der Suche gerankt als der Artikel. Außerdem können die Metadaten der Aufsätze in den Katalog der Bibliothek importiert werden. Es ist also wünschenswert, die Volltextsuche der Zeitschriften mit den Metadaten zu kombinieren.

Die ÖNB digitalisiert in einer Public Private Partnership mit dem Internetunternehmen Google ihren gesamten historischen, urheberrechtsfreien Buchbestand vom frühen 16. bis in die zweite Hälfte des 19. Jahrhunderts. Es ist geplant, rund 600000 Werke mit insgesamt ca. 200 Mio. Seiten zu digitalisieren, wobei die Hälfte bereits digitalisiert wurde. Die im Rahmen von „Austrian Books Online“ digitalisierten Bände sind sowohl über Google Bücher als auch über die Digitale Bibliothek der ÖNB nutzbar. Da sich unter diesen Bänden auch viele Zeitungen und Zeitschriften befinden, hat das ANNO-Team schon über 130 Periodika und mehr als 3 Mio. Seiten, die von Google gescannt worden sind, nachstrukturiert und über ANNO zugänglich gemacht. Üblicherweise ist in ANNO pro Titel ein Verzeichnis vorhanden, in diesem steht zu jedem Erscheinungsjahr ein Verzeichnis und darin wieder eines zu jedem Erscheinungstag, worin die Images dann liegen. Die ABO-Daten liegen bandweise in Verzeichnissen, die nach dem Barcode benannt sind, auf einem anderen Speichersystem. In ANNO befindet sich bei solchen Titeln im Tagesverzeichnis eine Indexdatei, welche die Links zu den Scans beinhaltet. Die unterschiedliche Datenstruktur und die verteilte Datenhaltung sind für den ANNO-Leser nicht relevant.

Sobald alle Jahrgänge textlich durchsuchbar sind, wollen wir die Trefferliste auch grafisch gestaltet anbieten. Dazu soll die Verteilung der Treffer über die einzelnen Jahrzehnte mit Hilfe eines Balkendiagramms veranschaulicht werden. Es kann den Suchenden helfen, zu sehen, in

welchen Perioden viele oder wenige bis keine Treffer vorhanden sind. Letzteres könnte auch darauf hindeuten, dass dieses Wort in dieser Zeit nicht verwendet wurde. Letztlich sollte es möglich sein, ein Jahrzehnt anzuklicken und sofort die Trefferliste oder – bei vielen Treffern – wieder eine Verteilung über die Einzeljahrgänge einzusehen.

Nachdem die ÖNB dieses Jahr großen Erfolg mit dem ersten Projekt hatte, das sich georeferenzierter Objekte bediente – dem Projekt „AKON – Ansichtskarten Online“ (<http://akon.onb.ac.at/>) – wird auch der Einsatz der Georeferenzierung in der Textsuche überlegt. Da die OCR-Ergebnisse dafür noch zu ungenau sind, wären möglicherweise z.B. die korrekten Aufsatztitel dafür geeignet. Zeitschriftenartikel, die einen georeferenzierbaren Ort in der Überschrift haben, könnten als Punkt auf einer Landkarte dargestellt werden. Dies wäre eine neue Form der Suche nach lokalen Informationen.

Die intensive Nutzung durch die Leser und die vielen Rückmeldungen zeigen, dass ein Ausbau von ANNO in

zweierlei Hinsicht gewünscht wird. Einerseits ist der Durst der Zeitungsleser auch nach 15 Mio. Seiten nicht gestillt. Immer wieder kommen Anfragen, wann welcher Titel wohl an der Reihe wäre. Außerdem sind beim von Anfang an praktizierten stufenweisen Ausbau lange noch nicht die letzten Absätze erreicht. Inhalt und Funktionalität können noch weiter wachsen.

Mag. Christa Müller
Österreichische Nationalbibliothek
Josefsplatz 1
A-1015 Wien
Österreich
christa.mueller@onb.ac.at