40.000 Bavarica-Volltexte Online

Seit 29. Juni 2011 ist im Rahmen der Bayerischen Landesbibliothek Online (BLO) eine neue, vom Münchener Digitalisierungszentrum (MDZ) entwickelte Volltextsuche online. Sie ermöglicht es, rund 40.000 retrodigitalisierte Werke mit Bayernbezug aus der Zeit vor 1870 gezielt nach Begriffen zu durchsuchen.

Von Florian Sepp und Sebastian Lutze

1. Ausgangsmaterial

Im Jahr 2002 startete das Münchener Digitalisierungszentrum (MDZ) zusammen mit anderen bibliothekarischen Kooperationspartnern die Bayerische Landesbibliothek Online (BLO), die sich seitdem unter einer stetig wachsenden Beteiligung verschiedenster Einrichtungen und Institutionen zu dem zentralen kulturwissenschaftlichen Informationsportal für Bayern entwickelte. Heute wird die BLO arbeitsteilig zwischen BLO und MDZ betreut. Als ein neuer Meilenstein ist nun dieses Volltextangebot aus den reichhaltigen Bavarica-Beständen der Bayerischen Staatsbibliothek entstanden.

Das Angebot enthält Volltexte aus vier Signaturenfächern, die getrennt durchsuchbar sind. Einschlägig ist natürlich das Fach Bavarica (Bavar.), in dem alle bayernrelevanten Titel mit Ausnahme derer zu Musik und Kunst sowie Karten und kartographischer Literatur aufgestellt wurden. Einbezogen ist ferner die Bibliothek der Bayerischen Berg-, Hütten- und Salzwerke AG (BHS). Die 1997 erworbene Spezialbibliothek des mittlerweile aufgelösten bayerischen Bergbauunternehmens umfasst Literatur zum Montan- und Salinenwesen mit dem regionalen Focus Bayern. Als dritter Bestand tritt das Fach Staatstheater (St.th.) mit historischem Aufführungsmaterial der Bayerischen Staatsoper hinzu. Den Abschluss bilden die „Ephemerides politicae“ (Eph.pol.), die politischen Zeitschriften und Zeitungen, ein eigentlich überregional ausgerichtetes Fach, das aber viele bayerische Periodika enthält.

Ebenfalls durchsucht werden können Titel mit den Erscheinungsjahren bis 1870, die mit dem Schlüssel „by“ als bayernrelevante Literatur ausgewiesen sind. Der Schlüssel wurde 1982 eingeführt und nur teilweise nachträglich vergeben. Da er von der Signatur unabhängig ist, erscheinen neben Bavar., BHS, St.th. und Eph.pol. auch weitere Signaturenfächer unter den Treffern.

Die Volltexte sind unkorrigiert, weshalb ihre Qualität unterschiedlich ist. Bekanntlich bereitet die Texterkennung vor allem bei Fraktur oft noch erhebliche Schwierigkeiten, die großenteils auf der Qualität der Vorlagen mit unterschiedlichen Drucktypen und Papierqualitäten sowie engen Bindungen beruhen, so dass diese Texte teilweise fehlerhaft erkannt werden. Damit die Nutzer einschätzen können, in welcher Qualität der Text erkannt wurde und welche Fehler eventuell systematisch auftreten, wird stets neben dem digitalen Bild der Seite auch der OCR-Text angeboten.

2. Technik

Das Angebot verwendet den Such-Server Solr[1], welcher vom MDZ mittlerweile in fünf Instanzen für zwölf Projekte (unter anderem Digi20, dMGH, Verkündungsplattform, ADB/NDB) erfolgreich eingesetzt wird. Solr basiert auf der ausgereiften Java-Softwarebibliothek Lucene[2] und ist durch die Open-Source-Lizenz und diverse Programmierschnittstellen[3] flexibel und vor allem transparent ausgelegt und somit die ideale Plattform für Forschung im Bereich Information-(Text-)Retrieval sowie Computerlinguistik. Die enorme Leistungsfähigkeit und ausgezeichnete Skalierbarkeit von Solr wird unter anderem eindrucksvoll durch das Projekt HathiTrust[4] veranschaulicht, deren Volltextsuche in über 9.000.000 Google-Digitalisaten ebenfalls mit Solr

realisiert wurde. HathiTrust gilt im Bereich „Large-Scale-Text-Retrieval“ als einzigartig und hat dank der gut dokumentierten Arbeit[5] die Entwicklungen des MDZ bezüglich der Konzeption und Implementierung entscheidend beeinflusst.

3. Funktionalitäten

Über Solr sind die Bavarica-Digitalisate direkt in die Schnellsuche der Bayerischen Landesbibliothek Online eingebunden und werden so zusammen mit der Orts- und Personendatenbank, den Projektbeschreibungen und den Karten der BLO durchsucht.

Die Suche in den rund 40.000 Volltexten erleichtert die von Google bekannte Autovervollständigung. Jedes automatisch vervollständigte Wort kommt auch tatsächlich in einem der Volltexte vor, womit Ergebnisse ohne Treffer vermieden werden. Bei der Näherungssuche werden solche Titel höher eingeordnet, in denen die Suchwörter in kurzen räumlichen Abständen zueinander vorkommen. Der Nutzer will meistens ja nicht nur wissen, ob die Wörter in einem Buch vorkommen, sondern interessiert sich vor allem für die Werke, in denen die Suchbegriffe in einem Satz oder Absatz enthalten sind. Möglich sind auch eine Phrasensuche (mit doppelten Anführungszeiten) und die Einstellung, welchen Abstand die gesuchten Wörter höchstens voneinander haben dürfen.

Die KWI[N]C-Ansicht (Keywords in Native Context) der gefundenen Suchwörter zeigt einen Ausschnitt aus dem gescannten Bild der Buchseite, womit der Nutzer schon in der Trefferliste die Ergebnisse bewerten kann. So sind Suchwörter in einer Fußnote oder Tabelle meist nicht so interessant wie die im Fließtext vorkommenden. Auch eine Ansicht des OCR-Textes ist möglich.

Orts- und Personennamen in den Volltexten wurden mit dem automatisierten Verfahren „Named Entity Recognition“[6] erkannt. Zum Einsatz kam dabei die Open-Source-Software GATE (General Architecture for Text Engineering)[7], welche durch die Java-Bibliothek GATE-Embedded[8] direkt in das PlugIn-System[9] von Solr integriert werden konnte. GATE bietet mit GATE-Developer[10] eine computerlinguistische Entwicklungsumgebung, in welcher Regeln für die Erkennung der Entitäten (Personen und Orte) entwickelt und getestet werden können. Die Trefferliste kann damit auf die Titel eingeschränkt werden, in denen eine bestimmte Person oder ein bestimmter Ort erwähnt wird.

Da eine komfortable und übersichtliche Navigation für die weitere Erschließung und Bewertung eines Einzeltitels durch den Benutzer eine enorme Rolle spielt, wurden die schon in der Kurztrefferliste verwendeten Key-Features KWI[N]C, die „Named Entity Recognition“ sowie die Anzeige der OCR auch bei der Suche im Band angewendet. Darüber hinaus erleichtern das automatisch erschlossene Inhaltsverzeichnis sowie Blätter-, Zoom- und Drehfunktion dem Benutzer die Navigation.

4. Perspektiven

Geplant ist, in diese High-End-Volltextsuche für Retrodigitalisate weitere Angebote zu integrieren, wie verschiedene Zeitschriften (ZBLG etc.), den Historischen Atlas von Bayern oder die bayernrelevanten Titel aus dem Projekt Digi20.

Die Volltexte finden Sie unter folgendem Link: www.bayerische-landesbibliothek-online.de/bavarica-volltexte

ANMERKUNGEN

1) http://en.wikipedia.org/wiki/Apache_Solr
2) http://en.wikipedia.org/wiki/Lucene
3) http://wiki.apache.org/solr/SolrPlugins
4) www.hathitrust.org
5) www.hathitrust.org/blogs/large-scale-search
6) http://en.wikipedia.org/wiki/Named_entity_recognition
7) http://en.wikipedia.org/wiki/General_Architecture_for_Text_Engineering
8) http://gate.ac.uk/family/embedded.html
9) http://wiki.apache.org/solr/SolrPlugins#Fields
10) http://gate.ac.uk/family/developer.html

DIE AUTOREN

Florian Sepp ist Mitarbeiter des Bavarica-Referenten in der Abteilung Bestandsaufbau und Erschließung III der Bayerischen Staatsbibliothek.

Sebastian Lutze ist Mitarbeiter des Referats Digitale Bibliothek in der Abteilung Bestandsaufbau und Erschließung III der Bayerischen Staatsbibliothek.