Die ULB Sachsen-Anhalt beherbergt eine der größten Zeitungsbestände Deutschlands. Foto: Markus Scholz

Zeitungsquellen für moderne Forschung nutzbar machen

ULB Sachsen-Anhalt digitalisiert historische Zeitungen

Der Anspruch an Digitalisierungsprojekte in Bibliotheken steigt. Über die Bestandserhaltung hinaus soll Nutzerinnen und Nutzern die intensive digitale Beschäftigung mit den Büchern, Handschriften, Drucken oder Zeitschriften ermöglicht werden. Anke Berghaus-Sprengel und Benjamin Auberer von der Universitäts- und Landesbibliothek (ULB) Sachsen-Anhalt in Halle (Saale) berichten anhand eines Zeitungsdigitalisierungsprojekts, wie sich diese Ansprüche in der Praxis umsetzen lassen.

Zeitungen sind eine zentrale Quelle für alle historisch arbeitenden Disziplinen. Sie liefern wichtige Einblicke in die Politik, Wirtschaft, Kultur und Gesellschaft einer Epoche. Als Pflichtexemplarsbibliothek besitzt die Universitäts- und Landesbibliothek Sachsen-Anhalt (ULB) einen der größten Zeitungsbestände Deutschlands: über 1 300 Zeitungstitel vor dem Erscheinungsjahr 1945, davon etwa 800 regionale Titel aus Mitteldeutschland.

Als zentrale Exponenten dieser Regionalzeitungen besitzen insbesondere der »General-Anzeiger für Halle und den Saalkreis« sowie die »Saale-Zeitung« große Bedeutung für die Wirtschafts- und Sozialgeschichte Mitteldeutschlands. Sie erschienen erstmals in der zweiten Hälfte des 19. Jahrhunderts und erreichten um 1900 mehr als die Hälfte der Bevölkerung in Halle und des damaligen Saalkreises.

Zeitungen sind eine zentrale Quelle für alle historisch arbeitenden Disziplinen.

Zahlreiche Anfragen auf die Zeitungsbestände aus Wissenschaft und Gesellschaft demonstrieren das nachhaltige Interesse an den Medien. Bereits in den 1990er-Jahren erfolgte aus Bestandserhaltungsgründen eine Verfilmung auf Mikrofilm, um der Zersetzung des

Papiers entgegenzuwirken. Der Zugriff auf die Zeitungen per Mikrofilmlesegerät erweist sich jedoch als zeitaufwendig und mühselig und entspricht nicht mehr den Anforderungen moderner wissenschaftlicher Praxis.

Texterkennung durch maschinelles Lernen weiterentwickeln

An der ULB startete deshalb Mitte 2019 ein Projekt zur Digitalisierung der beiden Regionalzeitungen. Gefördert von der Deutschen Forschungsgemeinschaft (DFG) werden innerhalb von zwei Jahren knapp eine Million Zeitungsseiten digitalisiert und online im Open Access (Unter der Lizenz CC-BY-SA 3.0 DE) für jeden frei verfügbar sein. Damit ermöglicht die ULB Sachsen-Anhalt die historische Auseinandersetzung mit diesem einzigartigen Forschungsmaterial. Wichtig ist dabei, dass die Inhalte einfach und schnell im Volltext gefunden werden können – so wie bei einer Google-Suche.

Zahlreiche Anfragen auf die Zeitungsbestände aus Wissenschaft und Gesellschaft demonstrieren das nachhaltige Interesse an den Medien.

Um dies zu erreichen, wird die Texterkennungssoftware (OCR) Tesseract eingesetzt. Mithilfe maschinellen Lernens und eines an der ULB entwickelten Trainingsworkflows ist diese Software in der Lage, Buchstaben, die auf den Zeitungsseiten damals ähnlich aussahen, zu unterscheiden und zu lernen, die unterschiedlichen Schrifttypen korrekt zu erkennen.

Auf diese Weise werden die gescannten Seiten maschinell lesbar und die Inhalte lassen sich mit jedem beliebigen Stichwort durchforsten. Dieser Volltext ermöglicht der Forschung darüber hinaus das Verfolgen von innovativen und datengetriebenen Fragestellungen aus dem Bereich der Digital Humanities. Dadurch werden beispielsweise empirisch fundierte, korpusbasierte Untersuchungen möglich.

Gute OCR-Ergebnisse benötigen eine hohe Bildqualität

Die Ergebnisse der Texterkennung sind jedoch nur so gut wie die Qualität der zugrundeliegenden Bilder. Im Projekt wird daher großer Wert auf eine hohe Bildqualität gelegt. Erst dann können die Inhalte durch Volltextanalysen sehr gut aufbereitet werden.

Bereits jetzt liegt die Trefferquote bei circa 95 Prozent. Das Ziel ist es jedoch, diese noch weiter zu steigern. Das macht für die Nutzung der Zeitungen sehr viel aus.

Bei der Digitalisierung der Mikrofilme werden die vorliegenden Mikroverfilmungen genutzt, die mithilfe eines Rollfilmscanners der Firma Zeutschel (OM 1800) digitalisiert werden. Hierbei wird immer ein Film komplett eingescannt und anschließend durch die Software Quantum Pro automatisch in Doppel- beziehungsweise Einzelseiten geteilt. Die Digitalisierung erfolgt nach den Praxisregeln Digitalisierung der DFG. Der eingesetzte Rollfilmscanner erzielt mindestens 470 dpi echte optische Auflösung (bezogen auf ein Original im A1-Format) und mindestens 12 bit Graustufen. Damit lassen sich die Filme effizient und qualitativ möglichst hochwertig digitalisieren. Die Einzelseiten werden gerade ausgerichtet und die Ränder soweit möglich minimiert, um die Speicherplatznutzung zu optimieren. Bei den Zeitungen, für die keine Verfilmung vorliegt, erfolgt die Digitalisierung direkt vom Original.

Digitalisierung soll neue Forschungen stimulieren

Die Kombination aus Rollfilmscanner und Texterkennungssoftware liefert sehr gute Ergebnisse. Bereits jetzt liegt die Trefferquote bei circa 95 Prozent. Das Ziel ist es jedoch, diese noch weiter zu steigern. Das macht für die Nutzung der Zeitungen sehr viel aus.

Mit dieser Texterkennungsrate ist es dann möglich, in Zeitungsartikeln oder Werbeanzeigen zuverlässig nach beliebigen Begriffen zu suchen. Zum Beispiel erhalten Forschende mit der Eingabe »Kapp-Putsch« detaillierte Einzelheiten über die »Schreckenstage in Halle« vom 13. bis 23. März 1920 und erfahren, dass die Stadt eine der großen Zentren der politischen Gewalt in den Zwanzigerjahren war. Auch die Geschichte einzelner Firmen und Personen aus der Geschichte Halles lässt sich so bequem über den Erscheinungszeitraum der Zeitungen verfolgen. Für Forschung und Lehre eine echte Bereicherung. Mit der Digitalisierung möchte die ULB über die pure Bereitstellung von Material hinaus zum Bearbeiten innovativer und datengetriebener Fragestellungen zur Erforschung der Geschichte Halles und seiner Region anregen.

Anke Berghaus-Sprengel,
Direktorin der Universitäts- und Landesbibliothek Sachsen-Anhalt;
Benjamin Auberer,
Fachreferent der Universitäts- und Landesbibliothek Sachsen-Anhalt

Die Universitäts- und Landesbibliothek Sachsen-Anhalt

Die Universitäts- und Landesbibliothek (ULB) Sachsen-Anhalt wurde bereits 1696 zwei Jahre nach der Entstehung der hallischen Universität gegründet. Mit einem Bestand von 5,5 Millionen Medien ist die ULB die größte wissenschaftliche Allgemeinbibliothek in Sachsen-Anhalt. Neben einer großen Zeitungs- und Zeitschriftenkollektion beherbergt die ULB in Halle über 115 000 Handschriften und eine wertvolle Sammlung an Drucken vom 15. bis zum 18. Jahrhundert.
https://bibliothek.uni-halle.de/