API

OPEN ACCESS

Publiziert im Februar 2025
DOI: 10.15460/apimagazin.2025.6.1.221

Zeitschriftenartikel

Begutachtet

Begutachtet:

Prof. Christine Gläser
HAW Hamburg
Deutschland

Erhalten: *10. November 2024*
Akzeptiert: *14. November 2024*
Publiziert: *04. Februar 2025*

Empfohlene Zitierung:

HARTWIG, Uwe, 2025: Volltext für digitale Sammlungen separat erzeugen. In: *API Magazin* **6**(1) [Online] Verfügbar unter: DOI 10.15460/apimagazin.2025.6.1.221

Volltext für digitale Sammlungen separat erzeugen

Einblick in das OCR-D-Projekt ODEM an der Universitäts- und Landesbibliothek Sachsen-Anhalt

Uwe Hartwig[1]*

[1] Universitäts- und Landesbibliothek Sachsen-Anhalt
Softwareentwicklung

* Korrespondenz: redaktion-api@haw-hamburg.de

Zusammenfassung

Die Sammlung ohne Volltext, der Volltext Jahrzehnte alt: Wie können Volltextdaten eigenständig und unabhängig von Retrodigitalisierungsprojekten erzeugt oder verbessert werden? Der folgende Beitrag erklärt Hintergründe und Motivation für das separate OCR-D-Implementationsprojekt ODEM. Er stellt kurz dessen Spezifika und deren Umsetzung in der Praxis bibliothekarischer Digitalisierungssysteme vor.

Schlagwörter: Volltext, OCR, VD18, Retrodigitalisierung

Create full-text for digital collections separatly

Insight into the OCR-D project ODEM at the University and State Library of Saxony-Anhalt

Abstract

The collection still without full-text, the full-text far outdated: how can full-text data be generated or improved independently of retro-digitization projects? The following paper explains the background history and motivation, which lead to the separate OCR-D implementation project ODEM. It explains ODEM's features and shows its application in the real-life of librarian digitization systems.

Keywords: Full-text, OCR, VD18, Retro-digitization

DOI: 10.15460/apimagazin.2025.6.1.221

1 Bibliotheken und Sammlungen

Die moderne Bibliothek als Institution ist die Bewahrerin des kulturellen Erbes. Neben dem Erhalten und Zusammentragen verschiedenster Medien managen insbesondere Hochschulbibliotheken heutzutage auch den Zugriff auf Informationsträger mit Ressourcen. Der klassische Bestand einer Bibliothek wird seit den 1990er Jahren um digitale Sammlungen ergänzt. Solche Sammlungen bestehen aus Objekten, die nur in digitaler Form existieren, wie PDFs, Webseiten oder Datenbankeinträge.

Darüber hinaus gibt es auch Sammlungen, die einer analogen Vorlage folgen. Sie machen vorhandene Bestände über das Internet zugänglich und vereinfachen damit gleichzeitig deren Erhaltung (Harloff 2001, S.2). Diese Art der digitalen Sammlung besteht aus sogenannten „Retrodigitalisaten“. Ein Retrodigitalisat ist die zusätzliche Überführung eines bereits analog vorliegenden Dokuments in eine digitale Form, die mit Methoden der IT gehandhabt werden kann (Liesken 2000, S. 133). Oft wurden diese Drucke vor Jahrhunderten veröffentlicht. Während die Originale konserviert bleiben, stehen die digitalen Abbilder der Allgemeinheit zur Verfügung. Es handelt sich um ein nachträglich erzeugtes Objekt – dafür steht das „retro“ im Namen. Grundlegend besteht das Retrodigitalisat aus einer Sammlung von Bilddateien, verbunden mit bibliothekarischen Katalogdaten. Das bedeutet, dass es keine zusätzlichen Informationen zum textuellen Inhalt und zur Struktur gibt, wie sie in modernen PDF-Dateien integriert sind. Was und an welcher Stelle in einem Buch enthalten ist, kann nicht beantwortet werden. Zwar kann man mit entsprechender Visualisierungs-Software in diesen Objekten „blättern“, aber für eine Suchanfrage sind sie nicht geeignet.

Werden Inhalt und Struktur ebenfalls erfasst und liegen in maschinell auswertbarer Form vor, sind Anwendungen denkbar, die über die Möglichkeiten des gedruckten Buches hinausgehen, wie übergreifende Recherchen oder historische Sprachforschungen auf verschiedensten Endgeräten (Harloff 2001, S. 11; Kempf 2015, S. 271; Kann und Hintersonnleitner 2015, S. 75). Den „Rohstoff“ für solche Mehrwerte bilden Volltextdaten. Unter diesem Begriff versteht man Angaben über den Text einer gescannten Druckseite. Häufig wird der Begriff „Optical character recognition“ (OCR) synonym verwendet, obwohl sich dieser ursprünglich nur auf das Erkennen sichtbarer Zeichen einer Rastergrafik bezieht (Rice et al. 1996, S.1). Die visuelle Anordnung von Zeichen in größere, logisch zusammenhängende Wörter, Zeilen oder Textregion wird ausgeklammert. Unter dem Begriff „Volltextdaten“ wird im Rahmen dieses Beitrages die Einheit aus erkanntem Zeichen und seiner nächsthöheren Einheit, dem Wort, mit Angaben zu deren Position auf einem Bild verstanden.

Damit die Vorteile einer universellen Verfügbarkeit ausgespielt werden können, müssen die Inhalte in eine verknüpfbare Form gebracht werden.

DOI: 10.15460/apimagazin.2025.6.1.221

1.1 Das Verzeichnis der Druckerzeugnisse des deutschsprachigen Raums (VD)

2004 trafen sich an der Universitäts- und Landesbibliothek (ULB) in Halle (Saale) die Deutsche Forschungsgemeinschaft (DFG) und Bibliotheken mit großen Beständen an historischen, deutschen Drucken zu einem Rundgespräch, um über die weitere Entwicklung der retrospektiven Nationalbibliographie Deutschlands zu beraten. In diesem Rahmen wurde die Organisation für ein „Verzeichnis der im deutschen Sprachraum erschienen Drucke des 18. Jahrhunderts“ (VD 18) festgelegt (Haller 2007, S.14).

Nach einer Pilotphase wurden an der ULB ab 2009 über 45.000 historische Drucke bis 2021 in mehreren Projektphasen und Kooperationen bearbeitet. In verschiedenen Digitalisierungssystemen wurden insgesamt 6,7 Millionen Seiten digitalisiert. Zwar wurden die Bände im Katalog eingearbeitet, strukturell erschlossen und über das Internet bereitgestellt, aber es fehlte der Volltext.

1.2 Volltext in der Massendigitalisierung

Zu Beginn der Massendigitalisierung in den 1990er Jahren gab es Gründe, Volltext hintenanzustellen. Für diese Aufgabe spielte man mit dem Gedanken, die Texte per Hand abzutippen. Aber natürlich war man sich auch damals bewusst, dass das manuelle Abtippen aller Texte bei Millionen von Bänden aus Kostengründen undenkbar war (Harloff 2001, S. 7ff). Hilfe versprach man sich von computergestützten Erkennungssystemen. Im Bereich standardisierter, moderner Schriftarten wurden bereits Genauigkeiten im Bereich von 99 % erzielt (Rice et al. 1996, S.6). Allerdings ergaben Testläufe mit Frakturschriften in historischen Drucken 1997 lediglich Erkennungsraten von 60-70 % (Harloff 2001, S.8). Die Resultate waren unbrauchbar. Somit entstand die Einschätzung, dass eine derartige OCR-Analyse „mittlerer Genauigkeit“ (Liesken 2000, S. 145) höchstens als unstrukturierter Sucheinstieg einen gewissen Wert liefern können. Selbst bei der Erzeugung mit Rechenmaschinen und Dienstleistern blieben die Kosten astronomisch. Schätzungen an der Bayerischen Staatsbibliothek (BSB) gingen davon aus, dass 40 Mrd. D-Mark für den damaligen Bestand von 7,5 Mio. Bänden der BSB anzusetzen seien (Liesken 2000, S. 134).

2005 war in einer Evaluierung zur „Retrospektiven Digitalisierung von Bibliotheksbeständen“ zu lesen, dass nur 15 % der evaluierten Einrichtungen Volltext auf verschiedenen Wegen erstellten. Bei automatisch generierten Daten fand keine systematische Qualitätsprüfung statt. In den wenigen Fällen, in denen Kontrollen erfolgten, sei die Fehlerrate beträchtlich gewesen (Czmiel et al. 2005, S. 25ff.). Detailliertere Angaben fehlen in der Evaluierung, aber es ist davon auszugehen, dass die Genauigkeit noch nicht besser war als Ende der 1990er Jahre.

2015 wurde konstatiert, dass automatisch generierter Volltext für das wissenschaftliche Arbeiten häufig wertlos sei (Nölte et al. 2016, S. 34). Auf der anderen Seite wurde die Anreicherung von (Retro-)Digitalisaten mit Struktur- und Volltextda-

DOI: 10.15460/apimagazin.2025.6.1.221

ten, wenn diese fehlten, zum festen Teil der neuen „Sammlungsaufgabe“ der Bibliothek erklärt (Stäcker 2019, S.307).

In der Zwischenzeit verbesserten sich OCR-Systeme gravierend durch Fortschritte bei der Implementation von Verfahren der Künstlichen Intelligenz (KI). Selbstlernende Systeme brachten gerade bei den eher unregelmäßigen, historischen Schriftarten große Verbesserungen. Neben kommerziellen Angeboten entstanden Open Source KI-Systeme wie „Tesseract-OCR“ (Smith 2007), „Kraken“ (Kiessling et al. 2019) oder „Calamari“ (Wick et al. 2018). Darauf bauten offene Plattformen wie „Transkribus“ (Kahle et al. 2015), „OCR4all/LAREX“ (Reul et al. 2017) oder eScriptorium (Kiessling et al. 2019) auf, die das Handling der komplexen OCR-Systeme für technisch weniger versierte Anwender vereinfachen. 2014 berief die DFG die koordinierte Förderinitiative „OCR-D“ ein mit dem Ziel, die Situation der Volltexte für VD-Bestände des 16.-18. Jahrhunderts umfassend zu verbessern (Engl 2020, S.2).

Die Lage besserte sich langsam. 2020 war in einer Erhebung unter Bibliotheken mit VD-Beständen zu lesen, dass 61 % der Einrichtungen bereits Erfahrungen mit dem Thema Volltext gesammelt hatten. Ein Drittel plane dazu Projekte (Engl 2020, S. 19).

2 Volltextdaten für digitale Sammlungen

Volltextdaten sind stets Thema, wenn DFG-Mittel für ein Retrodigitalisierungsprojekt beantragt werden. In den aktuellen „Praxisregeln zur Digitalisierung“ wird für jedes Vorhaben zur Retrodigitalisierung die Auseinandersetzung mit dieser Thematik erwartet. Für Druckwerke ab 1850 ist die Herstellung von Volltextdaten verpflichtend (Altenhöner et al. 2023, S. 30). Wie diese Daten erzeugt werden, schreibt die DFG nicht vor und verweist auf die Projektanforderung und den Zustand der Originale (ebd.).

Bei Projekten zur Massendigitalisierung mit Millionen von Druckseiten können die entsprechenden Volltextdaten nicht auf wirtschaftliche Art und Weise per Hand transkribiert werden. Automatisierte Texterkennungssysteme wie die bereits angeführten kommen zum Einsatz oder externe Firmen bieten ihre Dienste an.

2.1 Die Weiterentwicklung des Retrodigitalisats

Das klassische Retrodigitalisat besteht nach obiger Definition mindestens aus Bildern und Katalogmetadaten und ist „fertig“, wenn der Digitalisierungsprozess abgeschlossen ist. Somit besteht immer eine Kopplung von Projekt und Ergebnis. Nach erfolgreichem Abschluss sind keine Ressourcen vorhanden, um in-time generierte Daten mittel- oder längerfristig zu pflegen. Außerhalb des Förderkreislaufs entsteht eine Art Stillstand, ein „Project Lock-in“. Das Retrodigitalisat ist in seinen Bestandteilen eingefroren. Was aus konservatorischer Perspektive womöglich begrüßenswert ist, führt mittelfristig zu Problemen. Ein Buchdruck ist von Natur aus ein Vorgang mit

DOI: 10.15460/apimagazin.2025.6.1.221

abgeschlossenem Ergebnis. Das Digitalisat kennt solche physikalischen Beschränkungen nicht.

Bilder werden in der Regel *einmal* zu Beginn der Digitalisierung erzeugt. Wenn im Ablauf Bilder für Webpräsentationen in gewünschten Auflösungen und Qualitätsstufen generiert sind, können sie im Archivierungssystem abgelegt und bei Bedarf entnommen werden. Eine hochauflösende, aber bereits komprimierte Vorlage verbleibt als Arbeitsgrundlage im System.

Anders stellt sich die Situation bei Metadaten dar. Sie beziehen sich auf Einträge in bibliothekarischen Katalogen und Datenbanken. Diese können und werden sich über die Zeit ändern, so wie sich Katalogisierungsstandards und Datenschemata wandeln. Durch die Zusammenarbeit der Bibliotheken im VD-Bereich sind solche Anpassungen Ausdruck der Weiterentwicklung. Damit im Fall einer Katalogänderung das Objekt einer konkreten digitalen Sammlung aktuell bleibt, müssen dafür technisch und organisatorisch Möglichkeiten existieren.

2.2 Verbesserung von Volltext

Beim Volltext erreicht die Problematik Stillstand versus Entwicklung eine neue Dimension. Ein anderes System, ja selbst das gleiche System mit anderen Parametern, könnte *sofort* die Daten von heute verbessern. Bei dieser latenten Unsicherheit ist es verständlich, wenn lieber abgewartet wird. Bevor viel Zeit in die massenhafte Volltextgenerierung gesteckt wird, soll erst das 99 %-System gefunden werden. Digitale Sammlungen, die sich dagegen frühzeitig der Herausforderung „Volltext“ stellten, stehen dann Jahre später vor dem Problem, dass die Qualität der vorhandenen Daten als unzureichend empfunden wird. Dann steht ein großer Wechsel an.

Seit einigen Jahren gibt es Anstrengungen, komplette Volltextdatenbestände auszutauschen. In Deutschland startete 2016 an der Staats- und Universitätsbibliothek Bremen (SuUB) ein erstes Projekt zur nachträglichen Verbesserung von Fraktur-Volltexten (Nölte et al. 2016). Ähnliche Ansätze wurden unter Namen wie „Rerunning OCR“ in Luxemburg (Schneider und Maurer 2021) oder „Re-OCR“ in Finnland (Kettunen 2019) im Kontext der Massendigitalisierung historischer Zeitungen entwickelt, um komplette Volltextdatenbestände auszutauschen. An der ULB wurde dieses Konzept im Rahmen eines Projektes zur Digitalisierung Historischer Zeitungen umgesetzt, um ca. 100.000 Ausgaben und Beilagen mit einem adaptierten Modell noch einmal zu prozessieren (Hartwig 2022).

Andere Vorschläge beziehen die eigentlichen (End-)Nutzer der digitalen Sammlungen ein. Bei diesem Ansatz geht es darum, Volltexte bei Bedarf kollaborativ zu verbessern (Hertling und Klaes 2022). In der BSB existieren Workflows, bei denen Nutzer*innen Fehler im Volltext melden, welche in Zusammenarbeit von Staatsbibliothek und Google korrigiert werden (Kempf 2015, S.272). Da der Volltext als

DOI: 10.15460/apimagazin.2025.6.1.221

Ausgangsmaterial für weitere Anwendungen dient, führen Änderungen in diesem Bereich zu erheblichen Auswirkungen in nachgelagerten Systemen. Während die ursprüngliche Bilddigitalisierung durchaus als Einbahnstraße verstanden werden kann, sind Arbeiten an Katalogeinträgen und Volltexten als work-in-progress zu verstehen.

2.3 Datensouveränität und Integration

Ein weiteres Argument, sich als Einrichtung selbstständig in der Erstellung der Volltexte zu engagieren, ist die Frage nach dem Eigentümer. Wem gehören die Volltextdaten? Wofür und im welchem Umfang dürfen sie genutzt werden? Dienstleister rufen möglicherweise sehr günstige Preise auf, können dann aber vorgeben, unter welchen Bedingungen die Ergebnisse genutzt werden. Für öffentliche Einrichtungen ist es schwierig zu erklären, dass diese Daten von der Allgemeinheit bezahlt wurden, aber nur eingeschränkt zur Verfügung stehen.

Wird der Volltext selbst erstellt, hat die Einrichtung die Datensouveränität in der Hand. Nutzungsbedingungen können von der Bibliothek bestimmt werden. So gibt es in Österreich bei der ÖNB Überlegungen, abseits der etablierten Digitalisierung mit dem Technikgiganten Google/Alphabet, die Volltexterzeugung für bestimmte Bereiche wieder in die eigene Hand zu nehmen (Kaiser 2023, S. 207). Wenn nur der Volltext extern erzeugt wurde, muss die Besitzerin des Retrodigitalisates klären, wie dieser Stand in die ursprüngliche digitale Sammlung zurückfließt. Zur Integration gehören in der Regel mehrere Ebenen, die neben der Aktualisierung des Objektes auch Such- und Nachweissysteme umfassen. Diese Aufwände sind zu bedenken: Volltexterzeugung bedeutet in jedem Fall fortwährende Datenpflege.

3 Volltextgenerierung separat

Abgesehen von dem Wunsch, die Datenqualität je nach Möglichkeit zu verbessern, gibt es technische und organisatorische Schwierigkeiten. In der Massendigitalisierung stellt häufig schon die Menge ein Problem dar. Ebenso müssen organisatorische und technische Rahmenbedingungen festgelegt werden, um reguläre Arbeitsabläufe nicht zu beeinträchtigen.

3.1 Bisherige Erfahrungen der ULB

Volltexte wurden an der ULB zuerst 2014 im Rahmen des Pilotprojektes zur Digitalisierung historischer Zeitungen über einen externen Dienstleister bereitgestellt (Sommer et al. 2014). Ab 2019 erfolgte die Erzeugung von Volltextdaten, wiederum für historische Zeitungen, in Eigenregie über die Integration von „Tesseract-OCR“ im Rahmen eines laufenden Projektes und nicht unabhängig. Allerdings bestand die Abschlussaufgabe darin, die beim ersten Lauf generierten Volltextdaten mit einem angepassten Tesseract-OCR-Modell zu verbessern. Das zu dieser Zeit benutzte System erlaubte über eine Webschnittstelle (OAI-PMH) Zugriffe auf Metadaten und

DOI: 10.15460/apimagazin.2025.6.1.221

Bilder der bereits vorhandenen Volltexte. Nach der Erzeugung wurden die neuen Daten über einen Hotfolder-Mechanismus angeliefert. Auf diese Art und Weise erhielten 550.000 Zeitungsseiten in zwei Monaten neuen Volltext (Hartwig 2022, S. 17).

Parallel dazu nahm die ULB 2019/2020 an einer Teststellung für OCR-D-Komponenten teil, als eine von neun Pilotbibliotheken (Engl 2020, S. 16). Bei dieser Gelegenheit konnten erste Erfahrungen mit dem System gesammelt werden.

3.2 Anforderungen ODEM

Ab 2021 beteiligte sich die ULB als eines von drei Implementationsprojekten an der dritten Phase des koordinierten Förderprojektes OCR-D (OCR-D 2023). Ziel der Implementationsprojekte war die Integration von OCR-D-Komponenten in bestehende Digitalisierungsabläufe. Im Gegensatz zur bisherigen Praxis für Projekte wurden keine neuen Retrodigitalisate erstellt. Im Blickpunkt stand nicht die Verbesserung, sondern die nachträgliche Erzeugung von Volltext für einen speziellen Bestand.

Grundsätzlich baute das Vorgehen auf den Erfahrungen des Projektes zur Digitalisierung historischer Zeitungen auf. Bereits dort hatte sich gezeigt, dass eine rechenintensive Aufgabe wie die Volltexterzeugung durch eine parallele Vorgehensweise profitiert. OCR-D war zu diesem Zeitpunkt im Kern ein strikt sequenzieller Workflow, aber er konnte natürlich als Teil eines umfassendes Prozesses mehrfach, gleichzeitig und für jede Druckseite isoliert ausgeführt werden.

Dieser umschließende Prozess sollte:

- definiert parallel ausführbar sein,
- automatisch und robust laufen,
- offene Standards und Schnittstellen für Datenformate und Austausch einsetzen,
- für die Volltexterzeugung nur Komponenten aus dem OCR-D-Umfeld verwenden,
- bestehende Abläufe möglichst nicht beeinträchtigen,
- einfach skalierbar sein, und
- alle Datenformate der beteiligten Systeme verarbeiten und erzeugen.

Das neue Projekt wurde „OCR-D Erweiterung für Massendigitalisierung“, kurz: „ODEM“ getauft. Im Ziel stand die Umsetzung eines isolierten, parallelen OCR-D-Workflows für jede *einzelne* Seite eines Buches. Zur Vorbereitung werden alle Seiten geladen, dann parallel verarbeitet und die Ergebnisse im Nachgang wieder in die Ausgangsdaten integriert. Mit Seitenparallelität ist zwar ein zusätzlicher, rechentechnischer Overhead verbunden, aber im Falle eines Fehlers geht nur eine einzige Seite verloren. Dieser Punkt war besonders wichtig. Im sequenziellen Ablauf bedeutet ein *einziger Fehler* durch *eine einzelne Seite* den vollständigen Prozessabbruch.

DOI: 10.15460/apimagazin.2025.6.1.221

Es spielte keine Rolle, ob alle anderen Seiten zuvor erfolgreich waren.

Im Verlauf des Projektes zeigte es sich, dass Seiten mit wenig oder gar keinem Text, z. B. Illustrationen oder Bauzeichnungen einen Workflow mit Layouterkennung vor fatale Probleme stellen. Das ist nachvollziehbar: Texterkennungssysteme sind darauf trainiert, Textzeichen zu finden. Entsprechen Seiten schon für den menschlichen Betrachter nicht einer „normalen“ Buchseite, passen sie auch für die KI nicht in erlernte Muster. Darum wurden zusätzliche Filtermöglichkeiten für Strukturmetadaten implementiert. Der Prozess läuft stabiler, wenn er „schwierige“ Seiten erkennt und aussortiert. Zusätzliche Konfigurationsmöglichkeiten wurden umgesetzt, um Ressourcen und Laufzeit eines Prozesses auf Seitenebene zu kontrollieren und ggf. einen Verarbeitungsabbruch für eine Seite zu erzwingen. Zusätzlich zur Seitenparallelität wurden mehrere ODEM-Instanzen gestartet, die sich jeweils von einem kleinen, integrierten Dateiserver alle notwendigen Informationen zum nächsten Datensatz laden und dann Buch für Buch und Seite für Seite verarbeiten.

Nach erfolgreichen Tests wurden Metadaten genutzt, um automatisch über Sprachinformationen zur Laufzeit eine passende Modellkonfiguration zu ermitteln. Das erhöhte den Automatisierungsgrad beträchtlich. War zunächst angedacht, a-priori Teillisten von Digitalisaten je nach Sprache zu erstellen und zu verwalten – wobei Kombinationen von Sprachen auch diesen Ansatz vor Probleme stellte – , wählt der ODEM-Prozess auf Werksebene eigenständig die passende Konfiguration. Technisch gesehen, ist die eigentliche Volltexterstellung nur ein kleiner Teil des Projektes, weil viele OCR-D Funktionalitäten genutzt werden können. Die Implementation von ODEM konzentriert sich darauf, den Workflow für große Mengen vorhandener Digitalisate stabil auszuführen. Ein hoher Automatisierungsgrad wird erreicht, indem vorhandene Metadaten genutzt werden.

3.3 Qualitätsbestimmung von VD18 ODEM

Im Projekt musste die Qualität des Volltextes methodisch exakt bestimmt werden. Um diese Frage zu beantworten, ist zunächst ein Maßstab notwendig. Wird OCR von externen Dienstleistern erzeugt, empfiehlt die DFG in den Praxisregeln statistische Stichproben (Altenhöner et al. 2023, S. 31). Für die Bewertung interner, von der Einrichtung selbst generierter Volltextdaten macht die DFG keine Vorgaben.

Eine Projektmitarbeiterin erstellte dazu Korpora aus zufällig gewählten Buchseiten und arbeitete diese zu Referenzdaten (Groundtruth, GT) um. Nach intensiven Diskussionen mit dem OCR-D-Koordinierungsteam wurde diese Datensätze mehrfach umgearbeitet. Über den Zeitraum von 18 Monaten kamen insgesamt 1.600 Datensätze zusammen. Das ergibt für den VD18-Bestand der ULB folgende Sprachverteilung:

 DOI: 10.15460/apimagazin.2025.6.1.221

Tab.1: Häufigkeit der Sprachen im ODEM-Referenzkorpus (total 1.600 Seiten)

Sprache (Modell)	N_{Seiten}	Prozent %
dt. Fraktur (ger)	*1.026*	*64,15*
Latein Antiqua (lat)	*325*	*20,31*
Latein und dt. Fraktur (lat, ger)	*93*	*5,81*
Französisch (fre)	*72*	*4,50*
Latein und Griechisch (lat+grc)	*12*	*0,75*

Diese Seiten wurden zu GT-Daten gemäß OCR-D Groundtruth-Richtlinien (OCR-D 2024) transformiert und veröffentlicht (ODEM 2024a; ODEM 2024b). Sie dienten und dienen zur Optimierung des OCR-Workflows und für die Berichterstattung gegenüber der DFG. Auch nach Projektende wurde an diesen Daten weitergearbeitet. Volltext, selbst wenn er manuell erstellt wird, ist nie hundertprozentig korrekt, sondern immer auch eine Frage nach der Auslegung des Originals (Liesken 2000, S.147). Auch Groundtruth ist work-in progress.

Eine Analyse der Sprachannotationen ergab, dass 90 % der Seiten in die Kategorien „deutsch Fraktur“, „Latein“ oder eine Mischung der beiden Sprachen fielen. Für zukünftige Volltextvorhaben in diesem Bereich sollte besondere Aufmerksamkeit auf einem kombinierten Modell für Fraktur und Latein gelegt werden, weil damit der Großteil der VD18-Retrodigitalisate verarbeitet werden könnte.

3.4. ODEM im Einsatz

Durch die Skalierbarkeit des Verfahrens wurden fast 90 % der 45.000 VD-18 Drucke wie geplant in 10 Monaten verarbeitet. In Spitzenzeiten verarbeiteten acht ODEM-Instanzen jeweils 8 Seiten parallel, so dass bis zu 64 Seiten gleichzeitig prozessiert wurden. 2023/2024 erzeugte ODEM an der ULB Daten für ca. 280 persische Zeitschriften. Im Herbst 2024 startete die Volltexterzeugung für 25.000 Titel der VD17-Sammlung der ULB mit insgesamt 3,4 Mio. Seiten.

Außerhalb der ULB haben Experimente mit anderen Bibliotheken gezeigt, dass dort für den Einsatz zusätzliche Anpassungen erforderlich wären. Zudem sind einige Annahmen auf Datenebene, z. B. über die Art der Identifikation von Digitalisaten, über die Struktur der abgerufenen Metadaten und die erforderlichen Exportformate sehr unterschiedlich. ODEM stellt durch vielfältige Validierung Integrität und Konsistenz automatisch sicher, setzt diese jedoch auch voraus.

4 Fazit

ODEM hat gezeigt, dass Volltexte für Bestandsdaten mit OCR-D-Komponenten unabhängig von laufenden Projekten erzeugt und aktualisiert werden können. Die konkrete Implementation wurde im Projektverlauf auf der Internetplattform „GitHub“ veröffentlicht und wird von der ULB gepflegt und angepasst (ODEM 2024c). Zukünftig sollen auch Volltexte für Zeitungen mit ODEM erzeugt und verbessert werden.

DOI: 10.15460/apimagazin.2025.6.1.221

Wer in die Volltexterzeugung einsteigt, steht zunächst vor einer steilen Lernkurve. Zum Glück gibt es OCR-D, wo sich viele Volltext-Enthusiasten und Digital Humanities versammelt haben und Einsteigern Hilfestellung bei vielen Fragen rund um OCR, Daten, Workflows und Systeme geben können. Volltext separat zu erzeugen, bedeutet natürlich auch, dass eine zusätzliche Komponente in den Alltag der Digitalisierung integriert werden muss. Ressourcen sind für den Betrieb eines zusätzlichen Systems erforderlich. Klar ist auch, dass bei Laufzeiten, die mehrere Monate in Anspruch nehmen, die grundlegenden Prozesse isoliert laufen müssen. Mit ODEM konnte an der ULB eine derartige Lösung implementiert werden.

Separat bedeutet, dass Volltext losgelöst von einem konkreten Projekt erzeugt werden kann. Oft spielt Volltext nur eine Rolle, wenn ein neues Digitalisierungsprojekt in Aussicht steht, um zusätzliche Mittel zu beantragen. Aber was passiert nach Projektende? Es gilt, längerfristig und nachhaltig zu planen und die Volltextgenerierung in der Einrichtung zu verankern. Die digitale Sammlung ist ein Bestand, der nicht nur erweitert, sondern mit dem gearbeitet wird. Und das gilt natürlich auch für das Retrodigitalisat als dem Grundbaustein solcher digitalen Sammlungen. Daten verändern sich im Digitalen viel schneller als im Magazin.

DOI: 10.15460/apimagazin.2025.6.1.221

Literatur

ALTENHÖNER, Rainer, BERGER, Andreas, BRACHT, Christian, KLIMPEL, Paul, MEYER, Sebastian, NEUBURGER, Andreas, STÄCKER, Thomas, STEIN, Regine, 2023. *DFG-Praxisregeln "Digitalisierung"* [online]. Aktualisierte Fassung 2022. [Zugriff am: 20.10.2024]. Zenodo. Verfügbar unter: DOI: 10.5281/zenodo.7435724

CZMIEL, Alexander, IORDANIDIS, Martin, JANCZAK, Pia, KURZ, Susanne, 2005. *Retrospektive Digitalisierung von Bibliotheksbeständen. Evaluierungsbericht über einen Förderschwerpunkt der DFG* [online]. Köln: DFG. [Zugriff am: 15.10.2024]. Verfügbar unter: https://www.dfg.de/resource/blob/168798/d-d73d7ea39ab183aa846b39e78efed3c/retro-digitalisierung-eval-050406-data.pdf

ENGL, Elisabeth, 2020. OCR-D kompakt: Ergebnisse und Stand der Forschung in der Förderinitiative. In: *Bibliothek Forschung und Praxis* [online]. 44(2), S. 218-230. [Zugriff am: 15.10.2024]. Verfügbar unter: DOI: 10.1515/bfp-2020-0024

HALLER, Manfred, 2007. *Digitalisierung und Erschließung der im deutschen Sprachraum erschienenen Drucke des 18. Jahrhunderts*. Halle (Saale): Univ.- und Landesbibliothek Sachsen-Anhalt. Schriften zum Bibliotheks- und Büchereiwesen in Sachsen-Anhalt Bd. 88. ISBN 978-3-86010-968-7

HARLOFF, Jan, 2001. *Grundlagen der Retrodigitalisierung von Texten und Bildern* [online]. Frankfurt am Main: Bibliotheksschule [Zugriff am: 15.10.2024]. Verfügbar unter: https://core.ac.uk/download/pdf/11877884.pdf

HARTWIG, Uwe, 2022. Evaluation von Volltextdaten mit Open-Source-Komponenten. In: *o-bib. Das offene Bibliotheksjournal* / Herausgeber VDB [online]. 9(4), S. 1-21. [Zugriff am: 15.10.2024]. ISSN 2363-9814. Verfügbar unter: DOI: 10.5282/o-bib/5888

HERTLING, Anke und KLAES, Sebastian, 2022. Volltexte für die Forschung: OCR partizipativ, iterativ und on Demand. In: *o-bib. Das offene Bibliotheksjournal* / Herausgeber VDB [online]. 9(3), S. 1-11. [Zugriff am: 15.10.2024] ISSN 2363-9814. Verfügbar unter: DOI: 10.5282/o-bib/5832

KAHLE, Philip, COLUTTO, Sebastian, HACKL, Günter, MÜHLENBERGER, Günter, 2017. Transkribus – a Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In: *2017 14th IAPR International Conference on Document Analysis and Recognition* [online]. IEEE. [Zugriff am: 15.10.2024]. Verfügbar unter: https://ieeexplore.ieee.org/abstract/document/8270253/

KAISER, Max, 2023. Digitale Sammlungen als offene Daten für die Forschung: Strategische Zielsetzungen der Österreichischen Nationalbibliothek. In: *Bibliothek Forschung und Praxis* [online]. 47(2), S. 200-212. [Zugriff am: 15.10.2024]. ISSN 0341-4183. Verfügbar unter: DOI: 10.1515/bfp-2023-0021

KANN, Bettina und HINTERSONNLEITNER, Michael, 2015. Volltextsuche in historischen Texten. In: *Bibliothek Forschung und Praxis* [online]. 39(1), S. 73-79. [Zugriff am: 15.10.2024]. ISSN 0341-4183. Verfügbar unter: DOI: 10.1515/bfp-2015-0004

KEMPF, Klaus, 2015. Data Curation oder (Retro) Digitalisierung ist mehr als die Produktion von Daten. In: *o-bib. Das offene Bibliotheksjournal* / Herausgeber VDB [online]. 2(4), S. 268–278. [Zugriff am 15.10.2024] ISSN 2363-9814. Verfügbar unter: DOI: 10.5282/o-bib/2015H4S268-278

KETTUNEN, Kimmo Tapio, 2019. Open Source Tesseract in Re-OCR of Finnish Fraktur from 19th and Early 20th Century Newspapers and Journals – Collected Notes on Quality Improvement In: *Digital Humanities in the Nordic Countries* [online]. Copenhagen, Denmark. [Zugriff am: 15.10.2024]. Verfügbar unter: https://researchportal.helsinki.fi/files/131861408/25_paper.pdf

KIESSLING, Benjamin, TISSOT, Robin, Stokes, Peter, STÖKL BEN EZRA, Daniel, 2019. eScriptorium: An Open Source Platform for Historical Document Analysis. In: *2019 international conference on document analysis and recognition workshops (icdarw)* [online]. S.19-24 [Zugriff am: 15.10.2024]. Verfügbar unter: DOI: 10.1109/ICDARW.2019.10032.

LIESKEN, Herrmann, 2000. Retrodigitalisierung : Eine Zwischenbilanz. In: *BFB - Bibliotheksforum Bayern*. Jg. 28, Nr. 2, S. 132-153. ISSN 0340-000X

NÖLTE, Manfred, BULTMANN, Jan Paul, SCHÜNEMANN, Maik, BLENKLE, Martin, 2016. Automatische Qualitätsverbesserung von Fraktur-Volltexten aus der Retrodigitalisierung am Beispiel der Zeitschrift „Die Grenzboten“. In: *o-bib. Das offene Bibliotheksjournal* / Herausgeber VDB [online]. 3(1), S. 32-55. [Zugriff am: 15.10.2024]. ISSN 2363-9814. Verfügbar unter: DOI: 10.5282/o-bib/2016H1S32-55

OCR-D, 2024. *Ground Truth Richtlininien.* [online] OCR-D, 31.01.2024 [Zugriff am: 08.11.2024]. Verfügbar unter: https://ocr-d.de/de/gt-guidelines/trans

OCR-D, 2023. *OCR-D Phase III.* [online] OCR-D, 31.01.2024 [Zugriff am: 08.11.2024]. Verfügbar unter: https://ocr-d.de/de/phase3

DOI: 10.15460/apimagazin.2025.6.1.221

ODEM, 2024a. *OCR Groundtruth ULB VD18 German Fraktur – OCR-D Phase III*. [online] 20.10.2024 [Zugriff am: 08.11.2024]. Verfügbar unter: https://github.com/ulb-sachsen-anhalt/ulb-groundtruth-eval-odem-ger

ODEM, 2024b. *OCR Groundtruth ULB VD18 Latin – OCR-D Phase III*. [online] 20.10.2024 [Zugriff am: 08.11.2024]. Verfügbar unter: https://github.com/ulb-sachsen-anhalt/ulb-groundtruth-eval-odem-lat

ODEM, 2024c. *OCR Workflows based on OCR-D*. [online] 20.10.2024 [Zugriff am: 08.11.2024]. Verfügbar unter: https://github.com/ulb-sachsen-anhalt/ocrd-odem

REUL, Christian, CHRIST, Dennis Christ, HARTELT, Alexander, BALBACH, Nico, WEHNER, Maximilian, SPRINGMANN, Uwe , WICK, Christoph, GRUNDIG, Christine, BÜTTNER, Andreas, PUPPE, Frank, 2019. OCR4all – An open-source tool providing a (semi-) automatic OCR workflow for historical printings. In: *Applied Sciences* [online]. Vol. 9, no. 22. [Zugriff am: 15.10.2024]. Verfügbar unter: https://doi.org/10.3390/app9224853

RICE, Stephen V., JENKINS, Frank R., NARTKER, Thomas A., 1996. *The fifth annual test of OCR accuracy.* [online] 03.2012. University of Nevada: Information Science Research Institute. [Zugriff am 15.10.2024]. Verfügbar unter: https://www.stephenvrice.com/images/AT-1996.pdf

SCHNEIDER, Pit und MAURER, Yves, 2021. Rerunning OCR: A Machine Learning Approach to Quality Assessment and Enhancement Prediction. In: *Journal of Data Mining and Digital Humanities* [online] arXiv preprint arXiv:2110.01661. [Zugriff am: 15.10.2024]. Verfügbar unter: https://arxiv.org/pdf/2110.01661

SMITH, Ray, 2007. An overview of the Tesseract OCR engine. In: *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2, S. 629–633. [Zugriff am: 15.10.2024]. Verfügbar unter: https://static.googleusercontent.com/media/research.google.com/de//pubs/archive/33418.pdf

SOMMER, Dorothea, HEILIGENHAUS, Kay, PANKRATZ, Manfred, WIPPERMANN, Carola, 2014. Zeitungsdigitalisierung: eine neue Herausforderung für die ULB Halle. In: *ABI Technik*. Bd. 34, 2, S. 75–85. ISSN 0720-6763

STÄCKER, Thomas, 2019. Die Sammlung ist tot, es lebe die Sammlung! Die digitale Sammlung als Paradigma moderner Bibliotheksarbeit. In: *Bibliothek Forschung und Praxis* [online]. 43(2), S. 304-310. [Zugriff am: 15.10.2024]. ISSN 1865-7648. Verfügbar unter: DOI: 10.1515/bfp-2019-2066

DOI: 10.15460/apimagazin.2025.6.1.221

WICK, Christopf, REUL, Christian, PUPPE, Frank, 2019. Calamari − A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. In: *Journal for Language Technology and Computational Linguistics* [online]. arXiv preprint arXiv:1807.02004 [Zugriff am: 08.11.2024]. Verfügbar unter: https://jlcl.org/article/download/219/217