OCR-D

Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)

KONZEPT ZUR VOLLTEXTTRANSFORMATION DER VD 2023

02.10.2023

Koordinierungsprojekt

Inhalt

1 ABSTRACT

Eine effiziente Volltexttransformation der Verzeichnisse der im deutschen Sprachbereich erschienenen Drucke des 16. bis 18. Jahrhunderts (VD 16, VD 17, VD 18, insgesamt: VD) setzt ein koordiniertes Vorgehen voraus. Dieses Konzept unterbreitet dazu Empfehlungen und einen Plan zur technischen und organisatorischen Durchführung einer koordinierten Volltexttransformation der VD. Im ersten Teil werden die Ausgangslage und die Vorarbeiten von OCR-D, die Bedarfe und Gegebenheiten der VD-Träger- und Teilnehmerbibliotheken sowie die zur Verfügung stehenden technischen Möglichkeiten bilanziert. Um das zwischen den VD-Bibliotheken[1] (vor allem die federführenden bzw. Trägerbibliotheken) und OCR-D abgestimmte Vorgehen zu erleichtern, sollen vorhandene Organisationsstrukturen genutzt, und neue Strukturen, wie zum Beispiel eine Kontaktgruppen zwischen den VD-Gremien und der OCR-D-Koordinierung geschaffen werden. Im daran anschließenden Teil werden Empfehlungen gegeben und ein Vorschlag vorgestellt, der im Fall der Bewilligung einer vierten OCR-D-Projektphase ausgeführt werden kann. Von den VD-Trägerbibliotheken geplante VD-übergreifende Strukturen[2] sollen federführend bei der Organisation der Volltexttransformation in den VD-Partner-Bibliotheken beteiligt werden. Die Realisierung der Volltexttransformation wird auf Basis eines zentralen Dienstes und dezentraler Dienste in den VD-Bibliotheken geleistet. Rund 90 % des heute verfügbaren digitalen Bestandes können mit einem standardisierten Workflow volltextdigitalisiert werden. Vor allem dieser Bestand kann durch einen zentralen, performanten Dienst prozessiert werden. Die dezentralen Dienste setzen eine bedarfs- und forschungsbezogene Volltextdigitalisierung um. Für diese Aufgabe sind eine hohe Bestandskompetenz sowie die Anwendung von angepassten und komplexen Workflows notwendig. Durch OCR-D wird die erforderliche Software und Support bereitgestellt. Die Bereitstellung sowie die fortlaufende Verwendung der OCR-D-Ergebnisse, vor allem der Software, aber auch der Daten und der Expertise auch nach Abschluss der Förderung, werden im separat vorgelegten „Konzept zur Verstetigung der OCR-D-Ergebnisse“ im Detail dargestellt.

Das vorliegende Konzept knüpft an das Konzept zur Volltexttransformation der VD von 2020[3] an und konkretisiert vor allem im Bereich der technischen Umsetzung die Planungen zur Volltextdigitalisierung der VD. Es wird in der noch laufenden dritten Förderphase des OCR-D-Projekts in den kommenden Monaten noch weiter ausgearbeitet und stellt damit heute einen konsolidierten Zwischenstand dar. Derzeit befinden sich die Abläufe zur Integration der in Kooperation mit Google erstellten Bilddigitalisate und Volltexte in einem gemeinsamen VD-Suchportal bereits in Klärung. Bezüglich des Ablaufs der retrospektiven Volltextdigitalisierung kann festgehalten werden, dass VD-Titel, die durch Google prozessiert wurden, nicht erneut von OCR-D bearbeitet werden. Stattdessen sollen die Google-Volltexte, um Redundanzen zu vermeiden, nachgewiesen und nutzbar gemacht werden.

[1] Hier und im Folgenden schließt der Begriff VD-Bibliotheken nicht nur federführende bzw. Trägerbibliotheken ein, sondern auch in den VD-Arbeiten teilnehmende Bibliotheken. Bei Fragen um Koordinierung und das Angebot eines zentralen Dienstes für die Massenverarbeitung wird in der Regel explizit von den federführenden und Trägerbibliotheken der VD (BSB München, HAB Wolfenbüttel, SB Berlin, SUB Göttingen) gesprochen und diese explizit so genannt.

[2] Die angesprochenen, in Vorbereitung befindlichen VD-Strukturen und Planungen („VD Zukunft“) werden von den VD-Trägerbibliotheken geplant und verantwortet. Sie sind nicht Teil des hier vorliegenden Konzepts, werden aber in der abgestimmten Planung als maßgebliche Faktoren einbezogen.

[3] Vgl. OCR-D Koordinierungsprojekt: Konzept zur Volltexttransformation der VD vom 31.07.2020.

2 AUSGANGSLAGE UND VORARBEITEN

2.1 OCR-D

2.1.1 VORARBEITEN

OCR-D, die Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR), verfolgt das Ziel der Volltexttransformation der VD . Dazu wird die Volltexttransformation auf Basis der OCR „in ihre einzelnen Prozessschritte zerlegt, die in der OCR-D-Software als eigenständige Softwareprogramme (sog. Prozessoren) implementiert sind. Dies ermöglicht es, optimale Workflows für die zu prozessierenden Drucke zu erstellen und damit wissenschaftlich verwertbare Volltexte zu generieren.“[4] Nach der ersten Phase von OCR-D (2015–2017), die dem Eruieren des Stands der Technik und der relevanten Stakeholder sowie der Entwicklung eines Funktionsmodells für OCR diente, wurden in Phase II (2018–2020) Standards und interoperable Software entwickelt und Lücken im Funktionsmodell durch gezielte Entwicklung innerhalb von acht Modulprojekten geschlossen. In der aktuellen Phase III von OCR-D (2021–2024) wird die Software für den produktiven Einsatz optimiert und es werden von vier Implementierungsprojekten skalierende Lösungen für verschiedene Einsatzszenarien entwickelt. Das Ziel der Volltextdigitalisierung der gesamten VD rückt somit näher und kann nun konkret geplant und in Angriff genommen werden.

2.1.2 AUSGANGSLAGE DER VD

Für die konzeptionelle Vorbereitung dieser VD-Volltextdigitalisierung wurde im Konzept zur Volltexttransformation der VD aus dem Jahr 2020 eine Analyse des VD-Bestandes und der Organisation der Digitalisierung in den VD-Bibliotheken (Workflowsyteme) vorgenommen. Im Ergebnis wurden drei Szenarien der koordinierten Volltextdigitalisierung ermittelt (synchrone Neudigitalisierung, nachträgliche Volltexterstellung, typografische Bündelung), auf deren Grundlage Aufgaben und Schwerpunkte für Phase III von OCR-D definiert wurden. Allen voran steht die Integration von OCR-D in weit verbreitete Software für die Digitalisierung, die in dieser Phase von vier Implementierungsprojekten und drei Modulprojekten bearbeitet wurde. Die Implementierungsprojekte sind:

- OCR-D in Kitodo: Implementierung von OCR-D mit besonderem Fokus auf die Integration in Kitodo
- OPERANDI: Massiv skalierbare Implementierung von OCR-D auf Basis von Hochleistungsrechnern (High Performance Computing Clustern, HPC)
- ODEM: Implementierung von OCR-D unabhängig von Digitalisierungssoftware mit besonderem Fokus auf Retro-Volltextdigitalisierung von VD 18-Materialien
- OCR4all-libraries: OCR-Lösung mit grafischer Benutzeroberfläche

[4] https://ocr-d.de/de/about.html. Vgl. dazu auch das OCR-D Funktionsmodell https://ocr-d.de/assets/Funktionsmodell.svg

Die Modulprojekte sind:

- Workflow für werkspezifisches Training auf Basis generischer Modelle mit OCR-D sowie Ground-Truth-Aufwertung
- Erkennung von Schriftartgruppen zur OCR-Verbesserung
- OLA-HD Service – Ein generischer Dienst für die Langzeitarchivierung historischer Drucke

In den Projekten wurden Erfahrungen gesammelt und Lösungen für die benötigten Schnittstellen zwischen OCR-D und vorhandenen Digitalisierungsworkflows entwickelt. Durch die bisherigen Ergebnisse der Implementierungsprojekte stehen bereits vielfältige Lösungen und Möglichkeiten bereit, OCR-D in Bibliotheken einzusetzen.

Mit den federführenden und Trägerbibliotheken der VD-Projekte wurden durch die Staatsbibliothek zu Berlin (SBB) die Umsetzung des Volltextnachweises und die Verankerung von Zuweisungen thematisiert sowie verschiedene Vorschläge unterbreitet. Eine gemeinsame Lösung der Volltextnachweise ist in den Verbünden anzustreben. Die Überführung von Nachweisen bisher bereits vorliegender Volltextdigitalisate der VD-Bibliotheken, v.a. der Bayerischen Staatsbibliothek München (BSB) und demnächst der Universitäts- und Landesbibliothek Halle (ULB) ist im Anschluss an eine übergreifende Lösung vorgesehen.

Die folgenden Darstellungen und Lösungsstrategien basieren auf den Ergebnissen der vorhergehenden sowie der aktuellen Projektarbeit, auf bereits vorliegenden Konzeptionen und Abstimmungen mit den VD-Träger- und Teilnehmerbibliotheken. Dazu wurden u.a. die folgenden Schritte unternommen:

- Teilnahme an VD-17-Partnertreffen
- Durchführung einer Umfrage unter VD-17-Bibliotheken,
- gezielte Interviews mit Bibliotheken: Hierbei wurden besonders solche Bibliotheken befragt, die keine federführenden oder Trägerbibliotheken sind, aber mindestens an einem VD mitarbeiten und Bestände digitalisieren,
- Auswertung von aktuellen Bestandsnachweisen der VD (vgl. Kap. 2.3 Quantitative Angaben zu den VD-Beständen)

2.1.3 Ergebnisse des VD-17-Partnertreffens 2023 (Umfrage und Interviews)

Im Vorfeld des VD 17-Partnertreffens 2023 führte das OCR-D-Koordinierungsprojekt eine Umfrage unter den VD-Bibliotheken durch, um u.a. besser abschätzen zu können, in welcher Form sich die VD-Bibliotheken an einer koordinierten Volltexterschließung beteiligen können, und welche Faktoren dabei zu berücksichtigen sind.[5]

[5] Vgl. Kapitel 5.1 VD-Umfrage 2023: Ergebnisse.

Nach Auswertung der Antworten unter Hinzuziehung der Umfrage von 2020[6] kann festgestellt werden:

- Vor allem die heterogene Bibliotheks- und Bestandsstruktur setzt flexible Lösungsansätze voraus.
- Bei den vorhandenen OCR-Kompetenzen in den Einrichtungen und möglichen Ressourcen (IT- und Organisationsstruktur) sind große Unterschiede zwischen den betroffenen Einrichtungen erkennbar.[7] Es liegen zumeist wenige interne Kenntnisse zur Volltextdigitalisierung vor.
- Die Bilddigitalisierungsworkflows werden in den meisten Fällen durch Workflowsysteme unterstützt.[8]
- Diese Systeme werden in manchen der Einrichtungen intern gehostet, in anderen auch als extern gehostete Version genutzt.
- Beim Workflowsystem Kitodo wird in vielen Fällen nicht die aktuellste Version verwendet. Die Migration auf eine aktuelle Version ist zwar zumeist geplant, aber die Realisierung steht häufig noch aus.

In den zusätzlich geführten Interviews wurden folgende Aspekte angesprochen:

- Arbeitsorganisation der Volltextdigitalisierung
- Integration der OCR-D-Lösung in den Digitalisierungsworkflow der Einrichtung
- Gewährleistung der Volltextdigitalisierung mit entsprechenden Arbeitskräften oder Dienstleistern
- koordinierte Volltextdigitalisierung vor dem Hintergrund der in den Einrichtungen festgelegten oder geplanten sammlungsbezogenen Forschungsstrategien
- Welche Ergebnisse und Möglichkeiten bietet OCR-D mit dem Abschluss des Projektes und welche aufzubauenden Fördermöglichkeiten wären notwendig?

Die Interviews zeigen den ausgeprägten Wunsch der befragten Institutionen, mit OCR-D die Volltextdigitalisierung ihrer Bestände voranzutreiben. Die konkrete Einführung muss dabei in den Einrichtungen sehr individuell gestaltet werden und wird in einigen Häusern durch das Fehlen von spezialisierten Fachkräften erschwert. Eine Möglichkeit zur Integration in den Workflow ist die Einbettung von OCR-D in Kitodo und in Goobi durch in der aktuellen Förderphase laufenden Implementierungsprojekte *OCR-D in Kitodo* und (für Goobi, da Teil des Projekts) *OCR4all-libraries* sowie *OPERANDI.* 2020 nutzten die interviewten Bibliotheken noch zu einem nennenswerten Teil Goobi (23,5% gegenüber Kitodo mit 29,4%),[9] vereinzelt DWork. 2023 lag die Nutzung von Kitodo bei den befragten Einrichtungen bei 64%.

2.1.4 Ergebnisse des DFG-Rundgesprächs 2023[10]

Vom 24. bis 25. August 2023 fand in der SUB Göttingen das DFG-Rundgespräch zur Zukunft der Produktion und Bereitstellung von Volltexten für die nationalbibliografischen Verzeichnisse VD 16, VD 17 und VD 18 statt. Die Diskussion ergab, dass eine initiale Förderung der umzusetzenden

[6] Vgl. OCR-VD-Konzept 2020.
[7] Vgl. ebenda, Kapitel 2.3.1 Bisherige Erfahrungen mit OCR [Stand 2020: 61% Erfahrungen mit OCR, Stand 2023 61%; Stand 2020 inhouse-aktive OCR-Anwendung 50%, Stand 2023 63%, Stand 2023 45% keine OCR-Kenntnisse vorhanden].
[8] Vgl. ebenda, Kapitel 2.2 (Bild-) Digitalisierung in den VD-Bibliotheken [Stand 2020: 29,4% Kitodo-Nutzung und 23,5% Goobi-Nutzung, Stand 2023: 64% Kitodo-Nutzung].
[9] Vgl. Anm. 8.
[10] Hier ausschließlich Ergebnisse bezüglich der Volltexttransformation. Vgl. insbesondere zu den Themen Verstetigung von OCR-D sowie zu VD Zukunft: OCR-D-Koordinierungsprojekt: Konzept zur Verstetigung der OCR-D-Ergebnisse und Protokoll des DFG-Rundgesprächs.

Volltexttransformation als notwendig angesehen wird, um Infrastrukturen und Personalkompetenzen aufzubauen. Dazu wird für eine mögliche vierte Förderphase von OCR-D eine 24-monatige Projektlaufzeit empfohlen. Weiter wurden Fördermöglichkeiten von synchroner Bild- und Volltextdigitalisierung angeregt und die Integration von Volltexterstellung in die Förderstrukturen und Förderbedingungen befürwortet.

In Bezug auf die seit 16 Jahren durch Kooperation mit Google vorgenommene Volltexttransformation an der Bayerischen Staatsbibliothek (BSB) wurde festgestellt, dass die Ergebnisse qualitativ und technisch anschlussfähig an OCR-D sind. Die OCR wird iterativ erstellt und damit bestmögliche Volltexte zur Verfügung gestellt. Eine Versionierung erfolgt hier derzeit zwar nicht, jedoch ist dazu eine Lösung innerhalb von VD Zukunft denkbar. Eine aufwändige Neuprozessierung zum Beispiel von Beständen des VD 16 ist zurzeit nicht notwendig. Bei der Integration der Volltexte in eine zukünftige gemeinsame Lösung für die Suche in allen drei VD sind jedoch gemeinsam mit OCR-D, BSB und VD-Community die Interoperabilität der Daten und der Schnittstellen zu prüfen und sicherzustellen. Diese Aufgabe wird als Teil des Projekts VD Zukunft angesehen. Für die koordinierte Volltextdigitalisierung bedeutet dies, dass Redundanzen durch entsprechenden Abgleich der vorhandenen Volltexte einfach vermieden werden können.

Des Weiteren wurde im Hinblick auf ein gemeinsames VD-Portal festgehalten, dass weitere Abstimmungen mit allen VD-Stakeholdern notwendig sind, um die technischen, katalogisierungsbezogenen und lizenzrechtlichen Fragen gemeinsam zu klären. Hervorgehoben wurde auch der Wunsch nach einer Versionierung von Volltexten, um sie als Forschungsdaten nutzbar/zitierbar zu machen und die langfristige Nachvollziehbarkeit daraus gewonnener Ergebnisse sicherzustellen. Die Themen Versionierung und die Zitierbarkeit mit persistenten Identifikatoren werden im Rahmen von OCR-D von dem Modulprojekt OLA-HD[11] adressiert.

Bei der Diskussion des hier in einer weiter entwickelten Fassung vorliegenden Konzepts zur VD-Transformation der VD zeichnete sich ab, dass sowohl ein *zentraler* OCR-D-Dienst, den die Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) und die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), aufgrund ihrer Erfahrungen in Implementierungsprojekt OPERANDI[12], erwägen, als auch ein *dezentrales* Deployment in den Häusern gewünscht werden. Die dezentrale Installation und Einrichtung eines OCR-Dienstes auf Basis der OCR-D-Software in den Häusern wird vor allem eine bedarfs- und forschungsbezogene Volltextdigitalisierung von komplexen Vorlagen betreffen, die einer im Hinblick auf die Originale fachkundigen Begleitung bedürfen. Dieses Vorgehen soll über die Referenzimplementierung der skalierenden Web-API in der OCR-D/core-Software bzw. über die entwickelten Implementierungsprojekte realisiert werden. Die Logistik, um erstellte Volltexte in die Systeme in den Häusern zurückzuspielen, muss VD-übergreifend abgestimmt werden. Dazu werden die Metadatensätze sowie die Datenbanken der VD erweitert und eine Datenlogistik für die Bilddigitalisate ebenso wie für die Auslieferung der OCR-Texte aufgebaut. Breite Zustimmung findet die Vorgehensweise, sämtliche VD-Digitalisate zentral mit Standardworkflows im Volltext zu digitalisieren – auch, um die Kosteneffektivität eines zentralen Dienstes zu maximieren – und nach Qualitätsprüfung ggf. mit vorlagenspezifischen Workflows dezentral in den Häusern neu zu verarbeiten.

---

[11] Vgl. https://ola-hd.ocr-d.de/

[12] Vgl. https://www.sub.uni-goettingen.de/projekte-forschung/projektdetails/projekt/operandi-ocr-d-performance-optimisation-and-integration/

2.2 Verzeichnisse der im deutschen Sprachraum erschienenen Drucke des 16. bis 18. Jahrhunderts

VD 16, VD 17 und VD 18 verfolgen das Ziel der Erstellung einer kooperativen retrospektiven Nationalbibliografie für die im deutschen Sprachraum oder in deutscher Sprache erschienenen Drucke. Diese drei eigenständigen, z.T. seit 1969 laufenden Projekte werden jeweils durch federführende bzw. Trägerbibliotheken[13] vertreten. Beteiligt sind außerdem eine Vielzahl von Partnerbibliotheken mit entsprechendem Bestand. Entsprechend der Druckproduktion der drei Sammlungszeiträume sind die VD hinsichtlich des Materials heterogen. Es geht dabei einerseits um Bestände, die als mehr oder weniger zahlreiche Mehrfachexemplare in vielen Einrichtungen vorhanden sind. Dem gegenüber stehen umfangreiche seltene bzw. unikale Bestände, v.a. mit regionalem oder besonderen Gattungsbezug (z. B. Flug-, Gelegenheits- und Amtsdruckschriften).

Der Fortschritt und die Vollständigkeit des bibliografischen Nachweises sowie der Bilddigitalisierung der einzelnen VD ist aufgrund ihrer Materialmenge, Laufzeiten und ihrer jeweiligen Genese unterschiedlich. Schätzungen für das VD 16 gehen von 120.000 Titeln aus;[14] davon sind bisher rund 90 % bibliografisch erfasst und 60 % bilddigitalisiert, wobei der Bestand der BSB (und damit 34 %) bereits über eine Kooperation mit Google als Volltext vorliegt. Das VD 17 wird nach neueren Hochrechnungen geschätzt 460.000 Titel umfassen, von denen 45 % bilddigitalisiert und der BSB-Bestand mit einem Anteil von 14 % bereits als Volltexte vorliegen. Dagegen soll das VD 18 geschätzt 600.000 Titel umfassen,[15] von denen rund 40 % derzeit bibliografisch erfasst und bilddigitalisiert sind. Volltexte für das VD 18 liegen aus der BSB (derzeit 9 % der Zielmenge) und demnächst auch aus der ULB Halle ca. 35.000 Volltexte vor.[16] Allein im Hinblick auf den Umfang erscheint das VD 18 als die größte Aufgabe für die Volltextdigitalisierung; jedoch ist das Material des VD 16 und VD 17 heterogener und komplexer, woraus sich spezifische Herausforderungen für eine Verarbeitung mit OCR-D ergeben. Es sind daher für die VD spezifische Schwerpunkte (z.B. Unterstützung von sammlungsbezogenen Forschungsstrategien) und Kategorien (z.B. die Bestandsgröße, die Diversität des Materials und Textsorten, die Infrastruktur) bei der Bewältigung der Volltextdigitalisierung zu beachten.

Die Realisierung der Volltextdigitalisierung in den VD-Bibliotheken erfordert ein hohes Maß an Koordination und Kooperation. Neben der Koordinierung der nationalbibliografisch-erschließenden Aufgabe und der Bilddigitalisierungsaufgaben, die in den letzten Jahren bereits im Fokus der VD-Arbeiten standen, müssen nun zusätzlich die Planung der Volltextdigitalisierung sowie die Auswahl und Priorisierung der im Volltext zu digitalisierenden Materialien organisiert werden. Für die Realisierung der OCR sind die VD-Bibliotheken u. a. durch Vermittlung von strategischen Partnerschaften zu unterstützen. Solche Partnerschaften können im fachlichen Austausch bestehen ebenso wie in der konkreten gegenseitigen Unterstützung bei der Umsetzung der Volltextdigitalisierung. So können Häuser, die selbst nicht in der Lage sind, OCR-D lokal zu betreiben, ihre Bilddigitalisate zum Zweck der OCR-Erkennung an andere Bibliotheken übergeben oder der oben angesprochenen zentralen Lösung zugeführt werden.

---

[13] VD 16: BSB München; VD 17: SB Berlin, BSB München, HAB Wolfenbüttel, VD 18: SUB Göttingen.

[14] Vgl. Kühne, Andreas: Die Drucke des 16. Jahrhunderts im deutschen Sprachbereich: Untersuchungen zur Weiterführung des VD 16. In: ZfBB 41 (1994), S. 32–59, hier S. 42f. Zitiert nach: Möncke, Gisela: Das „Verzeichnis der im deutschen Sprachbereich erschienenen Drucke des 16. Jahrhunderts“ (VD 16) als Teil einer deutschen retrospektiven Nationalbibliographie. In: ZfBB 51 (2004), S. 207-212, hier S. 208.

[15] Vgl. Rundgespräche zur Zukunft der nationalbibliographischen Verzeichnisse (VD). Bericht der veranstaltenden VD17-Trägerbibliotheken (Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Bayerische Staatsbibliothek München, Herzog August Bibliothek Wolfenbüttel). In: ZfBB 69 (2022) 1–2, S. 82–91, hier S. 83.

[16] Im VD 18 erfolgte seit Projektbeginn eine synchrone bibliografische Erfassung mit Bilddigitalisierung.

2.3 Quantitative Angaben zu den VD-Beständen

Der aktuelle Stand der VD wird in Abb. 1 dargestellt.

Die Daten stammen aus folgenden Quellen:

- Erfasste und bilddigitalisierte Titel VD 16: E-Mail-Korrespondenz mit Claudia Fabian (BSB) vom 09.08.2023.
- Schätzung Gesamtbestand VD 16: Kühne, Andreas: Die Drucke des 16. Jahrhunderts im deutschen Sprachbereich: Untersuchungen zur Weiterführung des VD 16. In: ZfBB 41 (1994), S. 32–59, hier S. 42f. Zitiert nach: Möncke, Gisela: Das „Verzeichnis der im deutschen Sprachbereich erschienenen Drucke des 16. Jahrhunderts“ (VD 16) als Teil einer deutschen retrospektiven Nationalbibliographie. In: ZfBB 51 (2004), S. 207-212, hier S. 208.
- Erfasste und bilddigitalisierte Titel VD 17: SB Berlin: VD17-Partnertreffen. Folien vom 12.06.2023.
- Revidierte Schätzung Gesamtbestand VD 17 und Schätzung Gesamtbestand VD 18: Rundgespräche zur Zukunft der nationalbibliographischen Verzeichnisse (VD). Bericht der veranstaltenden VD17-Trägerbibliotheken (Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Bayerische Staatsbibliothek München, Herzog August Bibliothek Wolfenbüttel). In: ZfBB 69 (2022) 1–2, S. 82–91, hier S. 83.
- Aktuelle Schätzung Gesamtbestand VD 17: SB Berlin.
- Erfasste und bilddigitalisierte Titel VD 18: E-Mail-Korrespondenz mit Christian Fieseler (SUB Göttingen) vom 13.08.2023.
- Volltextdigitalisierte Titel sind in allen VD nicht zentral erfasst, jedoch beziehen sich die Angaben hier auf die Bestände der BSB und der ÖNB, die durch die Google-Partnerschaft im Volltext digitalisiert wurden. Für das VD 18 kommen seit Neuestem 35.905 Volltextdigitalisate hinzu, die hier addiert wurden. Inwiefern es hier Doppelungen zwischen BSB und ULB Halle gibt, muss noch erfasst werden.

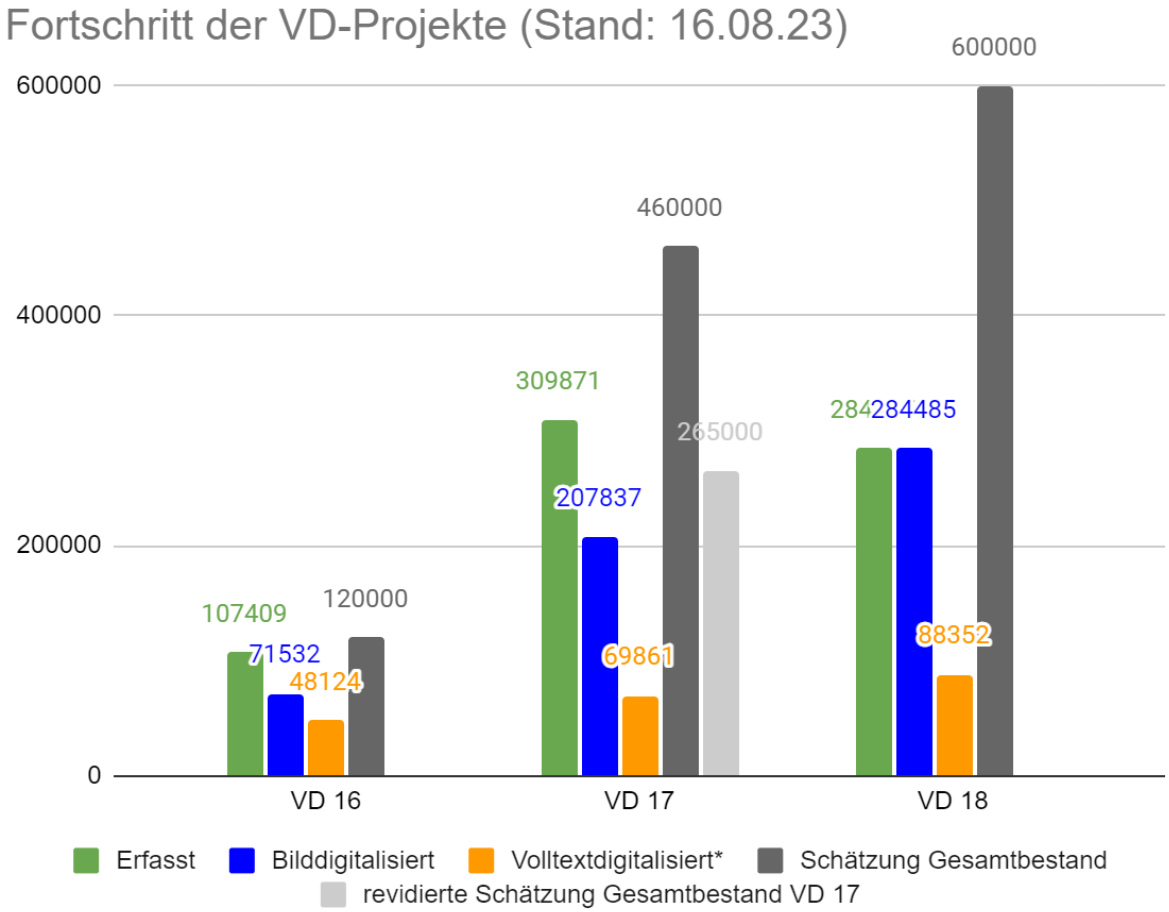

Abbildung 1: Stand der VD-Projekte.

Für die Planung der VD-Volltextdigitalisierung insgesamt ist ausschlaggebend, wie sich die Werke sowohl quantitativ in der Gesamtheit als auch bezogen auf die jeweiligen Häuser verteilen. Um einen Überblick darüber zu erhalten, wurden für die Bibliotheken mit den größten erfassten Beständen Abfragen in den jeweiligen Datenbanken durchgeführt. Relevant ist vor allem, wie viele Teile der Bestände sich in den jeweiligen Bibliotheken ohne Mehrfachexemplare befinden. Außerdem wurde erfasst, für wie viele dieser Titel bereits Bilddigitalisate hinterlegt sind. In BSB und ÖNB Wien liegen durch die Google-Partnerschaften zusätzlich Volltexte für die gesamten bilddigitalisierten Bestände vor.

Die jeweiligen nachgewiesenen Bilddigitalisate im VD 16 und VD 17 sind eigene Manifestationen, die nicht an die physischen Exemplare bzw. die Institutionen gekoppelt sind. Es handelt sich also um die Anzahl der Titel, für die mindestens ein Bilddigitalisat vorliegt, das aber auch außerhalb der größten acht Bibliotheken erstellt worden sein könnte.

2.3.1 VD 16

Bibliothek	Anzahl der Titel	Davon Anzahl der Titel mit Link auf (Bild-)Digitalisat[17]	Volltexte
Gesamt (ohne Mehrfachexemplare)	107.409	71.532	
BSB	41.927	41.148	41.148
HAB	40.711	29.899	
SBB	29.502	28.945	
ULB Halle	21.481	20.427	
ÖNB Wien	20.661	20.118	20.118
FB Gotha	15.602	12.631	
RS Zwickau	12.940	9746	
UB der LMU	12.016	10.520	

Tabelle 1: VD 16-Bibliotheken mit den größten Beständen (Stand: 10.08.2023).[18]

Ohne Berücksichtigung erfasster Mehrfachexemplare halten die acht größten VD 16-Bibliotheken zusammen 84.521 Titel (78 % des gesamten VD 16-Bestandes). Die Anzahl dieser Titel mit Link auf (mindestens) ein Bilddigitalisat beträgt 66.666 von insgesamt 71.532 bekannten Titeln mit Bilddigitalisat (93 %). Eine Sonderstellung haben die BSB und die Österreichische Nationalbibliothek (ÖNB) aufgrund großer Mengen bereits vorhandener Bild- und Volltextdigitalisate durch die Google-Partnerschaften. Während mit dem BSB-Bestand DFG-geförderte Bilddigitalisierungsvorhaben abgeglichen und von einer erneuten Bilddigitalisierung ausgeschlossen werden, um Mehrfacharbeiten zu vermeiden, ist dies für den ÖNB-Bestand nicht der Fall.

Die BSB und die Herzog August Bibliothek (HAB) besitzen einen Großteil des VD 16-Bestandes; die quantitativen Auswirkungen der Hinzunahme weiterer Bibliotheken sind jeweils vergleichsweise gering in Relation zur Gesamtzahl der zu verarbeitenden Werke. Demnach kann eine Volltextdigitalisierung der gesamten Bestände der größten Bibliotheken im VD 16 über 90 % aller vorhandenen bilddigitalisierten Titel abdecken. Um diese Abdeckung signifikant zu erhöhen, sind weitere Digitalisierungsvorhaben notwendig. Jedoch ist dabei zu beachten, dass innerhalb des DFG-Programms Digitalisierung und Erschließung nur für eine breite wissenschaftliche Zielgruppe relevante Materialien digitalisiert werden können.[19] Den größten Umfang an noch nicht im Bild digitalisiertem VD 16-Bestand hält die HAB. Für eine effektive Volltexterstellung dieses Segments und ähnlicher Bestände bedarf es daher einer Förderung, die sich auf Bild- und Volltextdigitalisierung bezieht.

[17] Das Bilddigitalisat muss nicht in der jeweiligen Institution liegen, um hier mitgezählt zu werden.
[18] Suche in https://bvbat02.bib-bvb.de/TouchPoint_touchpoint/start.do?SearchProfile=Altbestand&SearchType=2l anhand des jeweiligen Sigels, vgl. https://www.bsb-muenchen.de/fileadmin/pdf/historische_drucke/vd16_bibliotheken_sigelliste_201611.pdf
[19] https://www.dfg.de/download/pdf/foerderung/programme/lis/digitalisierung_erschliessung_hinweise_antragstellung.pdf

2.3.2 VD 17

Bibliothek	Anzahl der Titel	Davon Anzahl der Titel mit Link auf (Bild-)Digitalisat[20]	Volltexte
Gesamt (ohne Mehrfachexemplare)	309.871	219.185[21]	
HAB	97.367[22]	66.928	
BSB	69.861	63.572	63.572
SLUB	67.034	54.471	
UB Erfurt/FB Gotha	66.940	45.115	
SBB	65.084	54.455	
ULB Halle	63.559	54.563	
SUB Göttingen	57.954	49.582	
UB Leipzig	31.878	2755	
HAAB	29.814	24.584	
StB Nürnberg	18.969	12.256	
RS Zwickau	17.778	3207	
ThULB	13.336	12.108	
UB Rostock	12.426	11.861	

Tabelle 2: VD 17-Bibliotheken mit den meisten Beständen (Stand: 16.08.2023).

Ohne Berücksichtigung erfasster Mehrfachexemplare halten die 13 größten VD 17-Bibliotheken zusammen 292.548 (94% des gesamten VD 17-Bestandes) Titel. Die Anzahl dieser Titel mit Link auf (mindestens) ein Bilddigitalisat beträgt 209.323 von insgesamt 219.185 bekannten Titeln mit Bilddigitalisat (96%).

Die Bestände des VD 17 sind zwar etwas weniger verteilt als im VD 16, jedoch ist auch hier eine weitere Bilddigitalisierung oder kombinierte Bild- und Volltextdigitalisierung[23] notwendig, wenn ein großer Teil zugänglich gemacht werden soll. Es liegen insgesamt 102.076 VD 17-Titel ohne Digitalisat vor,[24] darunter insbesondere 69.027 Unika.[25]

[20] Das Bilddigitalisat muss nicht in der jeweiligen Institution liegen, um hier mitgezählt zu werden.
[21] Diese Zahl ergibt sich aus der Zählung aller vorliegenden Digitalisate. Dazu zählen auch Digitalisate, die u.a. nicht den DFG-Digitalisierungsrichtlinien entsprechen oder die als sogenannte Schlüsselseiten angeben sind.
[22] Hier und in den folgenden Zeilen: https://kxp.k10plus.de/DB=1.28/
[23] Ein weiterer Antrag der HAB auf DFG-Förderung der Digitalisierung von zusätzlichen VD17-Titeln ist in Vorbereitung.
[24] Vgl.SB Berlin: VD 17-Partnertreffen. Präsentation am 12. Juni 2023., S. 2. Diese Zahl ergibt sich aus der Einbeziehung von nicht DFG-Digitalisierungsrichtlinien konformen Digitalisaten und digitalen Schlüsselseiten.
[25] Vgl. ebenda, S. 6.

2.3.3 VD 18

Bibliothek	Anzahl der Titel	Davon Anzahl der Titel mit Link auf (Bild-)Digitalisat[26]	Volltexte
Gesamt (ohne Mehrfachexemplare)	284.485[27]	284.485	
BSB	52.447	52.447	52.447
ULB Halle	47.392	47.392	35.905
SUB Göttingen	46.890	46.890	
SLUB	37.592	37.592	
SBB	30.209	30.209	
UB Tübingen	10.638	10.638	
UB Rostock	9107	9107	
HAB	6617	6617	
THULB	5379	5379	

Tabelle 3: VD 18-Bibliotheken mit den meisten Beständen (Stand: 31.03.2023).

Das VD 18 verzeichnet für die Titel nicht alle bzw. mehrere Exemplare, sondern stets nur ein Exemplar der Bibliothek, die den Titel bearbeitet hat. Pro VD-Nummer sind also nur eine Bibliothek und ein Exemplar verzeichnet. Daher sind keine Mehrfachexemplare nachvollziehbar. Wenn eine am VD18 mitwirkende Bibliothek OCR-D nicht implementieren oder bestimmte Titel nicht verarbeiten kann, ist es umso wichtiger, die Digitalisate entsprechend anderen Einrichtungen oder einem zentralen Volltextservice zur Verfügung zu stellen.

Die neun Bibliotheken mit den umfangreichsten VD18-Beständen halten zusammen 246.271 (85% des gesamten bisher erfassten VD 18-Bestandes) Titel. Im VD 18 sind für jeden erfassten Titel Bilddigitalisate vorhanden.

Für das VD 18 ergibt sich zur Zeit eine deutlich geringere Anzahl an insgesamt dort teilnehmenden Bibliotheken (21).[28] Zu beachten ist hier allerdings der entsprechend große Anteil des noch nicht erfassten (geschätzten) Gesamtbestandes. So kann zum aktuellen Zeitpunkt eine Volltexttransformation auf wenige Einrichtungen konzentriert und dennoch der größte Teil des katalogisierten und bilddigitalisierten Bestandes erfasst werden.

[26] Im VD 18 werden alle erfassten Titel bilddigitalisiert.
[27] Stand: 31.03.2023. E-Mail-Korrespondenz mit Christian Fieseler vom 13.08.2023.
[28] https://kxp.k10plus.de/DB=1.65/SET=1/TTL=1/START_WELCOME

2.4 Quantitative Aspekte und Bedeutung für die Volltexttransformation

Aus den vorangegangenen Auswertungen ergeben sich Hinweise auf Konstellationen von Bibliotheken, deren Beteiligung an der Volltexttransformation unerlässlich ist, um weite Teile der VD-Bestände via Volltext zugänglich zu machen. Dies umfasst teilweise auch noch die (Bild-)Digitalisierung vorhandener Bestände. Um über 90 % der bisher erfassten Titel aus den jeweiligen VD abbilden zu können, ist auch die Einbeziehung von kleineren Beständen notwendig, die sich in einer Vielzahl von Bibliotheken befinden. Da die Erschließung und Bilddigitalisierung der VD noch nicht abgeschlossen sind, wird voraussichtlich auch die Volltexterschließung noch nicht zeitnah abgeschlossen werden können und für mehrere Jahre eine wichtige Aufgabe bleiben.

Für die Bestimmung der zu erwartenden Aufwände einer koordinierten Volltexterschließung ist zunächst die Anzahl der Seiten bzw. Bilddigitalisate der Werke relevant, da die OCR-Verarbeitung seitenweise erfolgt. Eine detaillierte Auflistung für alle VD kann zur Zeit nur schätzungsweise angeben werden. Für das VD 17[29] wurde ein arithmetisches Mittel von 132 (Median: 28) Seiten berechnet und für die VD insgesamt ein Durchschnitt von 130 bis 150 angenommen,[30] wobei aktuell über die konkrete Verteilung der Seiten auf die Häuser keine Aussage getroffen werden kann.

Es kann von folgenden überschlagenen Größenordnungen für den *geschätzten Gesamtbestand* der VD ausgegangen werden:

Kenngröße	**Größenordnung**
Seitenumfang pro Titel im Durchschnitt	130 - 150
Anzahl Titel in allen VD[31]	1 Million
Gesamtseitenumfang (VD16,17,18)	130 - 150 Millionen
OCR-D Workflowschritte	5
OCR-Ergebnis pro PAGE-XML-Datei	150 kB
OCR-Ergebnis aller PAGE-XML Dateien einschließlich Zwischenschritten	62 - 70 TB
Bilddigitalisat (Durchschnitt)	10 MB
Bildmaterial insgesamt (einschließlich Zwischenschritten)	6,3 - 7,3 PB

Tabelle 4: Größenordnungen der VD-Volltexttransformation.

2.5 Qualitative Aspekte und Bedeutung für die Volltexttransformation

Eine weitere Herausforderung für die Volltexttransformation der VD ist der inhaltlich und damit verbunden typografisch heterogene Bestand, der in Hinblick auf für die OCR relevante Merkmale analysiert werden muss. Wie im Konzept zur Volltexttransformation der VD von 2020[32] bereits erläutert, sind einzelne Textgattungen (wie etwa Leichenpredigten) für die OCR mit einem erhöhten Aufwand

[29] Parsing der json-Dateien im Dump: https://git.hab.de/beyer/vd17-dump (Stand: 11.08.2023). Der Median beträgt 28 (ungeachtet der Werke mit verzeichneten 0 Seiten), jedoch sind hier weitere Analysen für aussagekräftige Zahlen notwendig. So enthalten 768 Werke in der Liste laut Daten 0 Seiten und es sind Ausreißer nach oben enthalten.
[30] Die Berechnung für VD 16 und VD 18 wird dadurch erschwert, dass diese Datenbanken keine Datenbankdumps zur Verfügung stellen.
[31] Bei den Berechnungen nicht berücksichtigt sind Titel, die ggf. aus restauratorischen Gründen nicht digitalisiert werden können.
[32] Vgl. Konzept zur Volltexttransformation der VD, Kapitel 3.3.1 Prozessierungsszenarien: Szenario 3 (Typographische Bündelung).

verbunden. So können eine komplexe typografische Gestaltung (Layout, Schriftmischung, Marginalien), Sprachmischungen, Störungen wie Benutzungsspuren (bspw. handschriftliche Annotationen) sowie Beschädigungen der Seite vorkommen. Dadurch sind für die Volltextdigitalisierung Anpassungen in folgenden Bereichen vorzunehmen:

- OCR-Workflow bzw. Kombinationen von Prozessoren
- Auswahl von speziellen Segmentierungs- und Zeichenerkennungsmodellen,
- gegebenenfalls Erstellung von GT und werks- oder gattungsspezifisches Training der OCR

Allerdings kann als Zwischenergebnis der Prozessierung innerhalb des OCR-D-Implementierungsprojekts ODEM für das VD18 festgehalten werden, dass für ca. 90 % des Materials ein generisches Modell (trainiert mit Fraktur und Antiqua) ausreicht. Das verdeutlicht, wie wichtig eine genaue Analyse der eigenen Bestände für die Aufwandsabschätzung ist.

3 EMPFEHLUNGEN

Ziel der koordinierten Volltexterschließung ist die vollständige Bereitstellung des VD-Bestandes in digitaler Form (Text und Bild). Dabei sind der Zugang (Metadaten, URN) sowie die verwendeten Formate entsprechend den *Praxisregeln Digitalisierung* zu realisieren.

Für die koordinierte Volltexttransformation sind Zuständigkeiten in der Verteilung der OCR-Bearbeitung von einzelnen Titeln und das Einpflegen der Nachweise in die VD-Datenbanken unter den beteiligten VD-Bibliotheken zu regeln sowie weiterhin die Unterstützung bei der Projektkonzeption von haus- und bestandsübergreifenden Volltexterschließungen (z. B. bei gattungs- und/oder layoutspezifischen Vorhaben) sowie eine enge Abstimmung mit OCR-D für die technische Bereitstellung der OCR zu organisieren. Dafür ist eine VD-Struktur zu schaffen, die aus Vertretern der jeweiligen VD-Trägerbibliotheken besteht und die damit verbundenen Aufgaben löst sowie ein kompetenter Partner für die jeweils beteiligten Bibliotheken ist. Für die Realisierung dieses Zieles sind sowohl organisatorische als auch institutionelle Voraussetzungen notwendig. Aus diesen Gründen werden die folgenden Empfehlungen ausgesprochen:

3.1 Schaffung von Rahmen- und Förderbedingungen zur Umsetzung der Volltexttransformation des kulturellen Erbes (VD)

Es wird empfohlen, für die Volltexttransformation der VD angemessene Rahmen- und Förderbedingungen zu schaffen:

- Der eingeleitete VIGO-Prozess zur Revision, Redaktion und Erstellung der Praxisregeln Digitalisierung/Volltextdigitalisierung ist zu intensivieren.
- Die vom OCR-D Projekt erarbeiteten Richtlinien, Spezifikationen und Empfehlungen sind in den Kanon damit verbundener Förderbedingungen aufzunehmen.
- Für die fachliche, vermittelnde und technische Betreuung der anstehenden Volltextdigitalisierung ist die institutionelle Schaffung eines Kompetenzdienstes Digitalisierung/Volltextdigitalisierung zu planen und einzurichten.
- In den VD-Bibliotheken ist die digitale Erfassung (bibliografisch und im Bild) zu intensivieren und fortzuführen. Die vorhanden VD-Nachweissysteme sind um die Möglichkeiten eines Zugangs zum digitalen Volltext zu erweitern. Zur Realisierung dieser Aufgaben ist im Rahmen der nationalbibliografischen Erschließung die Bereitstellung weiterer Fördermittel notwendig.

- Für die Realisierung der Volltextdigitalisierung wird eine „DFG-Aktionslinie VD-Volltextdigitalisierung“ vorgeschlagen. Diese Linie soll auf drei Bestands-Ebenen eine Antragstellung ermöglichen:

    a) Synchrone Digitalisierung (Bild- und Volltextdigitalisierung) – vornehmlich einfache Workflows
    b) Nachträgliche Volltextdigitalisierung – vornehmlich einfache Workflows
    c) Typographie- bzw. gattungsspezifische Volltexterstellung – komplexe Workflows

Der Bestand der Ebenen a) und b) zeichnet sich durch eine einfache typografische Gestaltung aus (u.a. durch Überschriften, Absätze, klar erkennbaren Spaltendruck, Kolumenentitel, Fußnoten sowie durch geringe Schrift- und Sprachmischung). Der Bestand der Ebene c) wird durch eine umfangreiche typografische Gestaltung und eine komplexe Schrift- und Sprachmischung gekennzeichnet sowie Bestände, die nach einer Evaluation nicht den für a) und b) zugrunde gelegten Maßstäben entsprechen.

Es sind mehrere Wege der Volltextprozessierung förderfähig:

- Nutzung eines zentralen OCR-Services, der OCR-D-Software im Einsatz hat (Kooperationspartner, Dienstleister)
- Dezentrale Implementierung der OCR-D-Software in der eigenen Einrichtung; die Volltextdigitalisierung ist Bestandteil des Digitalisierungsworkflows und wird über eine geeignete Workflow-Software gesteuert und prozessiert
- Im Rahmen einer Kooperation mit anderen Einrichtungen eine gemeinsame bzw. arbeitsteilige OCR-D-Volltextdigitalisierung in eigener technischer Verantwortung

3.2 Realisierung der Volltextdigitalisierung des vorliegenden gesamten digitalen (Bild-)Bestandes durch den zentralen OCR-D-Dienst

Im Rahmen einer OCR-D-Phase IV können infrastrukturelle Voraussetzungen für die VD-Volltextdigitalisierung geschaffen, die Bereitstellung des OCR-D-Dienstes geleistet und die retrospektive Volltextdigitalisierung der vorhandenen bilddigitalisierten Bestände (dann aktueller Stand) begonnen werden. Dieser Bestand umfasst derzeit:

VD	**Bestand (Titel) abzüglich bereits vorhandener Volltexte der BSB**	**Bestand (Seitenzahl bei Ø 132 Seiten)**
16	25.548 ohne ÖNB: 24.196	3,8 Mio. ohne ÖNB: 3,4 Mio.
17	145.751	19,2 Mio.
18	193.824	25,6 Mio.

Tabelle 5: Nachgewiesene VD-Bestände (bilddigitalisiert) in den jeweils größten VD-Bibliotheken mit geschätzter Gesamtzahl an Seiten abzüglich der BSB-Bestände (Schätzung einer Größenordnung für die koordinierte Volltexttransformation).

- Der OCR-D-Dienst stellt und gewährleistet die notwendige Datenlogistik (Transport der Bilddigitalisate, Umsetzung der OCR, Nachweis und Speicherung der OCR-Ergebnisse) in Abstimmung mit den VD und den Geberbibliotheken der Digitalisate.
- Für die Realisierung der retrospektiven Volltextdigitalisierung ist die Bestandskompetenz zu nutzen. In Abstimmung mit dem VD-Gremien ist auf Grundlage von technischen und bibliografischen Metadaten eine Priorisierung und Gruppierung der Bestände durchzuführen. Diese Sets können entsprechend dieser ermittelten Metadaten in einfachen oder komplexen

Workflows verarbeitet werden. So kann bei der Anwendung von einfachen Workflows eine unnötig komplexe Segmentierung[33] vermieden werden.

- Die Anwendung von komplexeren Workflows ist in Abstimmung mit den Geberbibliotheken der Digitalisate zu realisieren. Unter Nutzung der Bestandskompetenz kann entweder notwendiger GT erstellt oder ein angepasster Workflow eingerichtet werden. Es wird empfohlen, die Volltexttransformation dieser Vorlagengruppe bei entsprechender Kapazität in der VD-Bibliothek oder in Kooperation mit anderen Bibliotheken im Rahmen eines OCR-Schwerpunkts durchzuführen.

3.3 Einrichtung und Integration von angepassten Workflows

- Die VD-Bibliotheken zeichnen sich durch ihre Bestandskompetenz und Kenntnis der Nutzungs- und Nachfrageseite seitens der Wissenschaft aus. Diese Kompetenz ist für die Entwicklung sowie Integration von angepassten Workflows zur Volltextdigitalisierung zu nutzen. Für die Umsetzung können folgende Möglichkeiten der Bereitstellung und Einrichtung einer OCR genutzt werden:
  - Nutzung eines Dienstleister-Angebotes
  - Nutzung des zentralen OCR-D-Dienstes
  - Dezentrale Nutzung der OCR im Haus oder in Kooperationen
- Durch eine Massenvolltextdigitalisierung des gesamten VD-Bestands – abzüglich bereits vorhandener Volltexte – können die Ergebnisse als Teil der Qualitätssicherung bzw. zur Identifizierung der Problemfälle genutzt werden. So können die Problemfälle anschließend mit speziell angepassten Workflows bearbeitet und die Ergebnisse sukzessiv verbessert werden.
- Die Einrichtung von spezifischen Workflows umfasst die Erstellung von GT zum Training von OCR-Modellen, zur Verbesserung von OCR-Modellen, zur Evaluation oder zur Verbesserung der Textqualität. Der GT ist dabei nach den OCR-D-GT-Richtlinien zu erstellen und öffentlich zur Verfügung zu stellen. Dabei können die OCR-D-Ressourcen (GT-Repositorium) genutzt werden.
- Empfohlen wird die Bildung von OCR-Schwerpunktzentren. Diese Zentren bestehen aus Einrichtungen, die die Volltexttransformation einer bestimmten Gruppe von Vorlagen abgestimmt durchführen. Neben den bereits angesprochenen Textgattungen können solche Schwerpunkte bspw. bestimmte Sprach- oder Schriftkombinationen, thematische Sammlungen oder die Anforderungen existierender Forschungs- oder Infrastrukturvorhaben sein, wie bspw. FID/NFDI- oder Editionsprojekte.

3.4 Einrichtung einer digitalen Bestandspflege in den Bibliotheken

Durch die Volltexttransformation und die Bereitstellung des digitalen Bestandes ergeben sich neue Aufgaben in der Datenpflege. Diese Aufgaben sind in den Workflows der Einrichtungen zu berücksichtigen und umfassen folgende Schwerpunkte:

- Die digitale Bestandspflege umfasst nach festgelegten Intervallen eine Reprozessierung. Durch eine kontinuierliche und koordinierte Verbesserung von OCR-Modellen und deren Anwendung in der OCR ist eine Verbesserung der Textqualität und der Erschließungstiefe möglich, was sich

[33] Die Analyse und Erweiterung der Metadaten eines VD-Bestandes (>70.000 Werke aus VD 16, VD 17 und VD 18) aus der SBB mit einem Spaltenklassifikator hat ergeben, das 90% der Werke einspaltig und zu 40% einsprachig vorliegen. Vgl. Gerber, Mike: Spaltenklassifikator und Analyse des SBB-Bestandes. https://qurator-data.de/~mike.gerber/2022-09%20select%20documents%20for%20mass%20digitization/Select%20documents%20for%20potential%20mass%20digitization.html

sowohl an den Bedarfen der Nutzenden im Allgemeinen als auch speziell der Forschenden orientiert.

- Intervalle der Reprozessierung können nach bestimmten Zeitabschnitten oder nach festgelegten Qualitätskriterien bzw. Verbesserungen bei den OCR-Prozessoren erfolgen. Zur Einschätzung der Erkennungsqualität sollen überprüfbare Methoden und bewährte Erschließungsmittel (GT) verwendet werden. Die so geschaffenen Erschließungsmittel (GT) sollen kollaborativ genutzt, kooperativ ausgetauscht und der Community zur Verfügung gestellt werden.
- Wie in Kap. Qualitative Aspekte und Bedeutung für die Volltexttransformation dargelegt, ist die Diversität der Vorlagen in den VD eine besondere Herausforderung für die massenhafte OCR. Um die jeweils adäquaten Workflows für die Erzielung bestmöglicher OCR-Ergebnisse auszuwählen, bedarf es der Bestandsexpertise. Neben der Qualität der Ergebnisse ist dem Ressourcenbedarf (u.a. Maschinenlaufzeit) Rechnung zu tragen. Dabei ist zu beachten, dass den ressourcenintensivsten Aspekt der OCR-Verarbeitung die Layouterkennung darstellt. Die leistungsfähigsten Verfahren dazu verwenden neben Heuristiken teilweise mehrere Machine-Learning-Modelle und, wenn vorhanden, auch GPU-Ressourcen, um komplexe Vorlagen mit mehreren Spalten, Initialen, Marginalien und Ähnlichem korrekt zu segmentieren. Für einfache Vorlagen, d.h. einspaltige, einsprachige Werke, ist eine derart aufwändige Segmentierung hingegen in der Regel nicht notwendig. Für diese Vorlagen kann durch Verwendung eines Prozessors mit einfacherer Segmentierung zumeist ein Ergebnis von ebenso guter Qualität bei deutlich geringerer Rechenintensität erreicht werden. Es ist daher ratsam, die im Volltext zu digitalisierenden Werke nach Merkmalen wie einsprachig/mehrsprachig, komplexe Typografie/einfache Typografie, einspaltig/mehrspaltig[34] etc. zu gruppieren, was über die Bestandskompetenz und andererseits teilweise über die Metadaten der Bestände (Sprachen, Gattungen), geleistet werden kann.
- Da der Bestand an Werken in den VD zwar umfangreich, aber selbst unter Beachtung der bisher nur auf Schätzungen beruhenden Zahlen von noch zu erfassenden und im Bild zu digitalisierenden Werken dennoch begrenzt ist, stellt zumindest dessen einmalige Volltextdigitalisierung eine zwar aufwändige, aber endliche Aufgabe dar. Je nach Grad der Kooperation und Priorisierung wird in absehbarer Zeit (laut Verstetigungskonzept ca. fünf Jahre) die initiale Retro-Volltextdigitalisierung abgeschlossen und Volltextdigitalisierung als zusätzlicher Schritt neben Erfassung und Bilddigitalisierung etablierbar sein. Dennoch müssen Volltextdigitalisierung und damit zusammenhängende Technologien, Kenntnisse und Abläufe als dauerhafte Aufgaben verstanden und langfristig in die Haushaltsplanung integriert werden. Mit einer Förderung in einer vierten Förderphase von OCR-D kann die dazu notwendige Kompetenz aufgebaut werden.
- Zum Aufgabenspektrum der digitalen Bestandspflege ist auch die Realisierung von neuen Dienstleistungen z.B. Verbesserung der Erschließungstiefe des Bestandes und neue Recherchemöglichkeiten zu zählen. Diesem Spektrum ist nicht nur gemein, dass sie Volltext zwingend erfordern, sondern auch, dass sie sehr gut mit der OCR-D-Architektur kombinierbar sind. Ein einmal eingerichteter Geschäftsgang für die Volltextdigitalisierung eröffnet somit Perspektiven für diverse zukünftige Aufgaben, deren Implementierung umso weniger kompliziert wird, je besser die Integration von OCR-D auf technischer, logistischer und personeller Ebene in den kommenden Jahren gelingt. Viele der nachgelagerten Verarbeitungsschritte verwenden ähnliche Technologien (z.B. KI) und erfordern ähnliche Kompetenzen wie die OCR. Entsprechend geschultes Personal wird so auch bei diesen

[34] Vgl. ebenda.

zukünftigen Aufgaben wertvoll sein. Dieser Umstand sollte in der unmittelbaren und langfristigen Ressourcen- und Haushaltsplanung berücksichtigt werden.

3.5 Förderbedarfe

Es wird daher empfohlen, einen zentralen OCR-D-Dienst und dezentrale OCR-Dienste in den Einrichtungen einzurichten. Mit der Schaffung von OCR-Diensten wird OCR-Kompetenz aufgebaut, die für die Umsetzung der Volltexttransformation und der zukünftigen digitalen Bestandspflege notwendig ist. Die mittel- und langfristigen Implikationen dieser Dienste werden im separat vorgelegten Verstetigungskonzept im Detail dargestellt (siehe Konzept zur Verstetigung der OCR-D-Ergebnisse).

Die Förderbedarfe sollten die folgenden Ebenen umfassen:

1. In einer OCR-D-Phase IV
    a. Förderung für den Aufbau von OCR-Diensten. Bei einem zentralen OCR-D-Dienst sind Anforderungen in den Bereichen der Datenlogistik und Abrechnung, sowie ein hoher Durchsatz der Volltexttransformation zu berücksichtigen.
    b. Beginn der Volltextdigitalisierung der vorhandenen Bilddigitalisate der VD (dann aktueller Stand).
    c. VD-Trägerbibliotheken und Bibliotheken integrieren OCR-Dienste in vorhandene Digitalisierungsworkflows. Dabei kann neben der vorrangigen OCR-D-Technik auch andere geeignete Software eingesetzt werden
    d. Zur digitalen Bestandspflege berücksichtigen VD-Träger- und Partner-Bibliotheken Methoden zur Einschätzung der Erkennungsqualität und die Erstellung von Erschließungsmitteln (GT). Die notwendige Kompetenz wird dazu mit Hilfe einer zentralen OCR-D-Koordinierung aufgebaut und kooperativ unterstützt.
2. Im Rahmen der vorgeschlagenen „DFG-Aktionslinie VD-Volltextdigitalisierung“ werden VD-Trägerbibliotheken und Partnerbibliotheken in der Umsetzung der bibliographischen und bild- und volltextdigitalisierenden Arbeit gefördert. Zusätzlich ist es zu prüfen, ob die zentralen Dienstleister auch für die volltextdigitalisierende Arbeit gefördert werden können.
3. Im Rahmen der nationalbibliografischen Aufgaben wird die Verzeichnung des digitalen Bestandes in den VD-Datenbanken ausgebaut und erweitert. Eine Erweiterung der Suchmöglichkeiten, die den so entstandenen Volltext angemessen nutzen soll bei den Planungen zur Optimierung der VD berücksichtigt werden.
4. Mit der Schaffung eines Kompetenzdienstes Volltextdigitalisierung werden Sachkenntnisse, Erfahrungen und Expertise im Bereich der Volltext-Digitalisierung von bibliothekarischen Beständen konzentriert. In Abstimmung mit den VD-Gremien, der OCR-D-Koordinierung, der Deutschen Digitalen Bibliothek, der NFDI sowie fach- und sachkompetenter Organisationen arbeitet dieser Dienst in den Bereichen Richtlinien, Qualitätssicherung sowie Vermittlung und Kollaboration bezogen auf Expertise zur Volltextdigitalisierung.

4 AUSBLICK UND ROADMAP

Auf Grundlage der Projektergebnisse und der Diskussionen sowie deren Ergebnisse im Rahmen des VD-Rundgesprächs empfiehlt OCR-D für die Umsetzung der koordinierten Volltextdigitalisierung:

1. Schaffung von Rahmen- und Förderbedingungen zur Umsetzung der Volltexttransformation des kulturellen Erbes (VD)
2. Realisierung der Volltextdigitalisierung des vorliegenden gesamten digitalen (Bild-)Bestandes durch den zentralen OCR-D-Dienst
3. Einrichtung und Integration von auf die jeweiligen Vorlagen bzw. Bestände angepassten Workflows
4. Durch die Volltexttransformation ist die digitale Bestandspflege in den Workflows einzurichten.

Die Umsetzung dieser Empfehlungen wird in enger Zusammenarbeit mit den VD-Gremien realisiert. Im Rahmen einer zweijährigen Phase werden notwendige Organisationsstrukturen und Dienste in folgenden Bereichen geschaffen:

1. Etablierung von VD- und OCR-D-Koordinierung
2. Integration, Aufbau der OCR-Kompetenz in den Bibliotheken und Einrichtungen
3. Etablierung des zentralen OCR-D-Diensts und Installation der OCR-D Software im Workflow der VD-Einrichtungen
4. koordiniertes Release-Management der OCR-D-Software sowie Aufbau von Supportstrukturen

Bei der Realisierung dieser Aufgaben arbeiten das OCR-D- und die VD-Gremien in enger Abstimmung zusammen:

Anfang 2024

Aufgaben für OCR-D:

- Gründung einer langfristigen Struktur (vgl. Verstetigungskonzept) mit den Aufgaben:
    - Bereitstellung, Pflege und abgestimmte Weiterentwicklung der OCR-D-Software
    - Einrichtung eines technischen Supports:
        - Erhalt der bereits regelmäßigen Calls (Tech Call, GT-Call)
        - Fixing kritischer Bugs
        - Hilfestellung bei Problemen der Integration und Anwendung von OCR-D-Software
    - Release-Management
    - Koordination der Erstellung von GT und OCR-Modellen

Aufgaben für die VD-Trägerbibliotheken und federführende Bibliotheken:

- Die OCR-D-Koordinierung und die VD-Gremien setzen die Planung und Konzeption der DFG-Förderlinie VD-Volltextdigitalisierung um.
- Durch die VD-Gremien wird eine Anpassung/Erweiterung des Metadatenbankformats in den VD vorgenommen:
    - zur Identifizierung und Nachweisführung von Bild- und Volltextdigitalisaten
    - zur Kennzeichnung innerhalb eines bereits vorhandenen Titels, damit Volltexte zu den einzelnen Titeln als vorgemerkt bzw. in Bearbeitung (und durch welche Einrichtung) eingetragen werden können. Hierbei könnte unterschieden werden:

        - Titel für die Massendigitalisierung mit einfachen Workflows
        - Komplexe Vorlagen
        - Titel für spezifische Vorhaben, für die eine höhere Genauigkeit erforderlich ist (sammlungsbezogene Forschungsdaten, Editionen)
- Konzept eines automatisierten Meldeverfahrens für neu hinzukommende Bilddigitalisate und Volltexte

Nach Q1 2024

- Die OCR-D-Koordinierung und die VD-Gremien legen Sets für retrospektive Volltextdigitalisierung fest (Bündelung von einspaltigen Monographien, Gattungen, komplexen Dokumenten)
- Einrichtung von automatisierten Meldeverfahren für die Volltexte an die VD
- Veröffentlichung der Förderlinie: OCR-D-Phase IV, mit den Teilen Interessensbekundung und nachfolgender Antragstellung
    - Antragsberatung und Antragsworkshops durch OCR-D-Koordinierung
    - Antragstellung der VD-Bibliotheken, anschließende Begutachtungs- und Bewilligungsverfahren der DFG
- Installation der OCR-D-Software in einzelnen Einrichtungen (Bibliotheken, Digitalisierungszentren, Dienstleister usw.)
    - Umsetzung der beantragten Projekte in den Einrichtungen
    - Aufbau der OCR-Kompetenz
    - Anpassung des Geschäftsgangs (Erweiterung der Digitalisierung um den Task Volltextdigitalisierung)

Nach Q4 2024[35]

- Einrichtung des zentralen OCR-D-Diensts und Realisierung der Datenlogistik mit zentralem OCR-D-Dienst
- Übergabe Listen mit zu verarbeitenden Titeln an den zentralen OCR-D-Dienst zur Volltexttransformation
    - Auslieferung der digitalen Volltexte an die Geberbibliotheken
    - Verzeichnung der Volltexte in den VD durch das automatisierte Meldeverfahren an die VD

[35]Da nach aktuellem Stand das Implementierungsprojekt OPERANDI bis November 2024 läuft, sind die Arbeiten eines zentralen Dienstes für die Massenvolltextdigitalisierung später anzusiedeln. Bei einer vorangehenden Koordinierung wie hier beschrieben, ist das unproblematisch.

5 APPENDIX

5.1 VD-Umfrage 2023: Ergebnisse

Der Fragebogen wurde gezielt an die VD-Bibliotheken über die VD17-Mailingliste versendet. Der Rücklauf umfasste 17 Antworten.

Die Fragen inkl. zusammengefasster Antworten lauteten:

1. Ist in Ihrer Einrichtung bereits OCR-Know-how vorhanden? Gibt es einen OCR-Workflow: in der Konzeption oder bereits in der Umsetzung und wenn ja, auf welcher Basis?

In 10 Einrichtungen ist bereits OCR-Know-how vorhanden; in 7 gibt es einen OCR-Workflow (überwiegend unter Einbeziehung von Dienstleistern); 3 weitere Einrichtungen planen einen OCR-Workflow.

2. Ist Ihre Einrichtung daran interessiert, die OCR-D-Software selbst zu implementieren, Know-how aufzubauen und zu pflegen, um die Volltextdigitalisierung der eigenen Bestände vollständig selbst durchzuführen?

9 Bibliotheken sind daran interessiert, die OCR-D-Software zu implementieren bzw. zu nutzen. Für den Nutzungsfall wurde häufiger die Integration in Goobi bzw. Kitodo genannt. Die Antworten ließen sich nicht immer eindeutig zuordnen, da die Bibliotheken die Konditionen (Aufwand, Personal und Infrastruktur) verständlicherweise momentan noch nicht abschätzen können.

3. Können Sie sich vorstellen, für die Volltextdigitalisierung Ihrer Bestände mit anderen Institutionen oder Infrastruktureinrichtungen bzw. Dienstleistern zu kooperieren? Welche Kriterien sind für Sie ausschlaggebend?

Grundsätzlich zeigen sich bis auf 3 Bibliotheken alle Befragten aufgeschlossen für Kooperationen. Als Kriterien wurden am häufigsten Kostenreduktion und gemeinsame Nutzung von IT-Infrastruktur genannt. Voraussetzung für erfolgreiche Kooperation ist, dass Qualitätsstandards von OCR-D eingehalten werden.

4. Es wird Bestandssegmente oder Gattungen (z.B. Leichenpredigten) geben, für die eine/mehrere gesonderte Prozessierungen notwendig werden. Wie hoch schätzen Sie diesen digitalisierten Bestand in Ihrer Einrichtung ein? Gibt es weitere Besonderheiten? Können Sie sich in diesen Fällen ein gattungsspezifisches oder typografisch getriggertes Vorgehen unabhängig von der bestandshaltenden Einrichtung vorstellen?

Überwiegend wird ein „gebündeltes“ Vorgehen für typografisch und vom Layout her schwierigen Vorlagen für sinnvoll erachtet; auch die gattungsspezifische Herangehensweise ist vorstellbar. Als besonders zu berücksichtigende Vorlagen wurden in diesem Zusammenhang genannt:

- Personale Gelegenheitsschriften (v.a. Funeralschriften) – hier liegt bereits eine große Menge digital in den verschiedenen Bibliotheken vor; weitere werden folgen.
- Flugschriften, Panegyrik (Herrscherlob)
- Gesangbücher
- Dissertationen
- Schreibkalender

- Historische Sprachvarianten und Mehrsprachigkeit der Vorlagen bzw. Mehrsprachigkeit auf einzelnen Seiten
- Außerhalb des VD-Bereichs: Illustrierte Presse, Menükarten, Kochbücher, Karten

5. Können Sie sich vorstellen, an der Erfassung von Ground Truth (GT)[36] im Verbund mit der entstehenden OCR-D-Community mitzuwirken? Gibt es bereits Erfahrungen in der Erfassung von Ground Truth?

7 Bibliotheken haben bereits über verschiedene Projekte Erfahrungen im Bereich der GT-Erfassung; 10 Einrichtungen haben diese Erfahrung nicht. Allerdings können sich 2 von diesen 10 vorstellen, bei entsprechender Personalkapazität GT zu erfassen. Die Kapazitätsfrage wurde von fast allen Bibliotheken benannt, das Interesse ist mit insgesamt acht von 17 Bibliotheken aber durchaus vorhanden.

5.2 Nachgewiesene VD-Bestände (bilddigitalisiert)

VD	Bestand (Titel)	Bestand (Seitenzahl bei Ø 132 Seiten)
16	66.666 ohne ÖNB: 65.344	8.8 Mio. ohne ÖNB: 8,6 Mio.
17	209.323	27,6 Mio.
18	246.271	32,5 Mio.

Tabelle 6: Nachgewiesene VD-Bestände (bilddigitalisiert) in den jeweils größten VD-Bibliotheken mit geschätzter Gesamtzahl an Seiten.

Abzüglich der Anzahl an bereits vorhandenen BSB-Volltexten ergeben sich folgende Bestandsgröße, die für eine retrospektive Volltextdigitalisierung vorgesehen sind:

VD	Bestand (Titel) abzüglich bereits vorhandener Volltexte der BSB	Bestand (Seitenzahl bei Ø 132 Seiten)
16	25.548 ohne ÖNB: 24.196	3,8 Mio. ohne ÖNB: 3,4 Mio.
17	145.751	19,2 Mio.
18	193.824	25,6 Mio.

Tabelle 7: Nachgewiesene VD-Bestände (bilddigitalisiert) in den jeweils größten VD-Bibliotheken mit geschätzter Gesamtzahl an Seiten abzüglich der BSB-Bestände (erste Schätzung einer Größenordnung für die koordinierte Volltexttransformation).

Die geschätzten Seitenzahlen basieren aktuell auf einem arithmetischen Mittel der Seiten pro Werk[37]. Bei der konkreten Volltextdigitalisierung ist der Seitenumfang des VD-Bestands in den einzelnen Einrichtungen anhand der Metadaten zu ermitteln. Neben dem Umfang der zu prozessierenden Bestände ist in Abstimmung mit OCR-D und den VD-Gremien eine Analyse des Materials (Sprache/Gattung/Textsorte) wesentlich. Mit dieser Analyse werden Sets von Vorlagengruppen gebildet, die spezifischen Workflows (einfach oder komplex) übergeben werden.

[36] Ground Truth bezeichnet intellektuell erstellte Transkriptionen für die Evaluation und das Training von OCR.
[37] Dieses Mittel wurde auf Grundlage einer Analyse des VD 17 ermittelt.

6 REFERENZEN

- DFG-Praxisregeln "Digitalisierung". Aktualisierte Fassung 2022. https://zenodo.org/record/7435724
- Gerber, Mike: Spaltenklassifikator und Analyse des SBB-Bestandes. https://qurator-data.de/~mike.gerber/2022-09%20select%20documents%20for%20mass%20digitization/Select%20documents%20for%20potential%20mass%20digitization.html
- OCR-D: https://ocr-d.de/
- OCR-D Koordinierungsprojekt: Konzept zur Volltexttransformation der VD vom 31.07.2020.
- Rundgespräche zur Zukunft der nationalbibliographischen Verzeichnisse (VD). Bericht der veranstaltenden VD17-Trägerbibliotheken (Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Bayerische Staatsbibliothek München, Herzog August Bibliothek Wolfenbüttel). In: ZfBB 69 (2022) 1–2, S. 82–91.
- SB Berlin: VD 17-Partnertreffen. Präsentation am 12. Juni 2023.
- VD 16:
    - Datenbank: https://bvbat02.bib-bvb.de/TouchPoint_touchpoint/start.do?SearchProfile=Altbestand&SearchType=2
    - Konkordanz: https://www.bsb-muenchen.de/fileadmin/pdf/historische_drucke/vd16_bibliotheken_sigelliste_201611.pdf
- VD 17: https://kxp.k10plus.de/DB=1.28/DB=1.28/?COOKIE=U999,K999,D1.28,E6216f742-4,I0,B9994++++++,SY,QDEF,A,H12,,73,,76-78,,88-90,NGAST,R45.118.185.245,FN / https://git.hab.de/beyer/vd17-dump/-/tree/master/json
- VD 18: https://vd18.gbv.de/viewer/index/