OCR-D

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

KONZEPT ZUR VOLLTEXTTRANSFORMATION DER VD – AKTUALISIERTE FASSUNG OKTOBER 2024

14.10.2024

OCR-D-Koordinierungsprojekt

Inhalt

1 Einleitung 1

2 Vorarbeiten und Ausgangslage 1

2.1 Vorarbeiten 1

2.2 Quantitative Ausgangslage 2

2.3 Abstimmung mit den VD-Trägerbibliotheken: VD-Zukunft, VD-Portal, VD-Index 3

3 Realisierung 4

3.1 VD-Bestand mit vorhandenen Bilddigisalitaten 4

3.2 VD-Bestand ohne Bilddigitalisate, ohne Katalogisierung sowie Spezialfälle 6

3.3 VD-Bestand mit vorhandener OCR 7

4 Empfehlungen an die DFG 8

4.1 Empfehlungen nach Abschluss der dritten Förderphase 8

4.2 Vierte Phase: Durchführung der Volltextdigitalisierung der VD und Abschluss der Verstetigung 9

4.2.1 Vierte Förderphase: Grundlagen 9

4.2.2 Vierte Förderphase: Projektplan 10

4.2.3 Zusammenfassung der Meilensteine 12

4.2.2 Vierte Förderphase: Kosten 12

5 Executive Summary 13

6 Referenzen 14

1 EINLEITUNG

Eine effiziente Volltexttransformation für die Werke, die in den Verzeichnissen der im deutschen Sprachraum erschienenen Drucke des 16. bis 18. Jahrhunderts (VD 16, VD 17, VD 18, insgesamt: VD) erschlossen sind, setzt ein koordiniertes Vorgehen voraus. Dieses Konzept unterbreitet dazu Empfehlungen und einen Plan zur technischen und organisatorischen Durchführung.

Zunächst werden die Vorarbeiten und die Ausgangslage von OCR-D, d. h. die Bedarfe der Community, die Anforderungen und Kompetenzen der VD-Trägerbibliotheken und der an den VD beteiligten, weiteren Bibliotheken[1] sowie die zur Verfügung stehenden technischen Möglichkeiten bilanziert. Im daran anschließenden Teil folgen Erläuterungen und Kennzahlen zur Realisierung der Volltexttransformation für die drei wichtigsten Szenarien. Darauf aufbauend werden konkrete Empfehlungen für die Umsetzung der Volltexttransformation gegeben. Die zentrale Empfehlung zur Förderung einer vierten OCR-D-Projektphase mit dem Hauptziel der Volltextdigitalisierung der VD-Bestände wird daran anschließend in einem eigenen Kapitel erläutert. Zuletzt fasst eine Executive Summary die wichtigsten Aussagen des Konzepts zusammen.

Das vorliegende Konzept zur VD-Volltexttransformation basiert auf dem 2023 eingereichten Konzept[2] und berücksichtigt die Rückfragen und Monita dazu aus dem Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme (AWBI) von 2023. Das Konzept zur VD-Volltexttransformation steht in engem Zusammenhang mit dem Konzept zur Verstetigung der OCR-D-Software, das parallel vorgelegt wird.

2 VORARBEITEN UND AUSGANGSLAGE

2.1 VORARBEITEN

OCR-D, die Koordinierte Förderinitiative zur Weiterentwicklung von Optical Character Recognition (OCR)-Verfahren, verfolgt vorrangig das Ziel der Volltexttransformation des VD-Bestands. In der ersten Projektphase von OCR-D (2015–2017) wurden der Stand der Technik, die relevanten Stakeholder sowie ein Funktionsmodell für OCR eruiert. In der zweiten Phase (2018–2020) wurden Standards und interoperable Software entwickelt, während Lücken im Funktionsmodell durch gezielte Entwicklungen innerhalb von acht Modulprojekten geschlossen wurden. In der aktuellen Phase III (2021–2024) wurde die Software für den produktiven Einsatz optimiert, und es wurden von vier Implementierungs- und drei Modulprojekten skalierbare Lösungen für verschiedene Einsatzszenarien erarbeitet. Damit kann nun eine Volltextdigitalisierung des VD-Bestands konkret geplant und umgesetzt werden.

[1] Hier und im Folgenden schließt der Begriff VD-Bibliotheken nicht nur federführende VD-Bibliotheken bzw. VD-Trägerbibliotheken ein, sondern auch alle anderen an den VD-Arbeiten teilnehmenden Bibliotheken. Bei Fragen zur Koordinierung und zum Angebot eines zentralen Dienstes für die Massenverarbeitung werden in der Regel hinsichtlich der VD explizit die Trägerbibliotheken der VD, Bayerische Staatsbibliothek München (BSB), Herzog August Bibliothek Wolfenbüttel (HAB), Staatsbibliothek zu Berlin (SB Berlin) und die Staats- und Universitätsbibliothek Göttingen (SUB), adressiert.
[2] Konzept zur Volltexttransformation der VD 2023: https://ocr-d.de/konzepte/VD-Volltexttransformation%202023.pdf, Anlage dazu: https://ocr-d.de/konzepte/Anlage_VD-Volltexttransformation%202020.pdf

2.2 Quantitative Ausgangslage

Die Planungen stützen sich auf inzwischen konsolidierte quantitative Angaben zu den vorhandenen VD-Katalogisaten, Bilddigitalisaten und Volltexten:[3]
Insgesamt liegen in den drei VD 736.734 katalogisierte Titel vor. Für 586.663 dieser Titel liegen außerdem Bilddigitalisate vor, und für 193.072 Titel wurden bereits Volltexte erstellt. Demnach existieren für 393.591 Titel Bilddigitalisate, aber keine Volltexte (siehe Tabelle 1). Das sind diejenigen Titel, auf die sich das vorliegende Konzept bei der VD-Volltexttransformation zunächst konzentriert. Ferner werden für die Volltextdigitalisierung von 150.071 katalogisierten, aber noch nicht bilddigitalisierten Titeln sowie für die 443.266 noch nicht katalogisierten Titel (potenzielle VD-Nova) in diesem Konzept entsprechende Vorschläge unterbreitet. Für diese zusammen 593.337 Titel des VD-Bestands fehlen noch die Bilddigitalisate (Stand September 2024). Diese können mittelfristig im Rahmen weiterer Projekte entweder integriert, also sowohl bild- als auch volltextdigitalisiert werden,[4] oder zunächst bild- und später separat volltextdigitalisiert werden, sobald die Bilddigitalisierung erfolgt ist. Für die Erreichung von annähernder Vollständigkeit wird eine Fortführung der Bilddigitalisierung notwendig bleiben. Es wird aber empfohlen, bei noch anstehenden Bilddigitalisierungsprojekten grundsätzlich von vornherein die Volltextdigitalisierung einzubeziehen. Ausgehend von der durchschnittlichen Zahl von 132 Seiten pro Titel, die für die vorliegenden Bilddigitalisate des VD 17 ermittelt wurde[5], ergeben sich bei 586.663 Titeln aus allen VD mit Bilddigitalisat (aber ohne OCR) rechnerisch etwa 77.439.516 Seiten, für die eine OCR durchgeführt werden muss. Werden die Titel, die bereits volltextdigitalisiert sind, sowie Mehrfachexemplare berücksichtigt, und wird dazu eine Fehlertoleranz von 10 % eingerechnet, so ergibt sich ein Umfang von 58 Millionen Seiten für die Volltextdigitalisierung auf der Basis der vorhandenen Bilddigitalisate. Mit dieser Summe wird im Weiteren kalkuliert.

	VD16	VD17	VD18	Gesamt
Titel im Katalog	108.311	313.486	314.937	736.734
Titel mit Bilddigitalisat	71.994	222.915	291.754	586.663
Titel mit OCR	41.148	63.572	88.352	193.072
Titel mit Bilddigitalisat, ohne OCR	30.846	159.343	203.402	393.591
Titel im Katalog ohne Bilddigitalisat	36.317	90.571	23.183	150.071
Noch nicht katalogisierte Titel	11.689	146.514	285.063	443.266

Tabelle 1: Zahlen zum Bestand der Titel für VD16, VD17 und VD18. Für die Zahlen „Titel mit Volltext“ wurden eigene Recherchen zugrunde gelegt. Alle anderen Zahlen stammen aus der aktuellen Veröffentlichung von Lauer et al. (2024)[6].

Das OCR-D-Implementierungsprojekt ODEM hat überprüft, mit welchen Erkennungsquoten und Fehlerraten die Texterkennung mit einem generischen Modell der OCR-D-Software (trainiert auf Fraktur und Antiqua) für VD-Bilddigitalisate und vergleichbare Quellen durchgeführt werden kann. Ein vielversprechendes Zwischenergebnis für das VD18 zeigt, dass für etwa 90 % des Materials zufriedenstellende Ergebnisse geliefert werden. Das verdeutlicht, dass die weit überwiegende Mehrzahl

[3] Lauer, Limbach et al. (2024), S. 508.
[4] Für ein solches Projekt befindet sich beispielsweise ein Antrag zur Förderung der Bild- und Volltextdigitalisierung von VD17-Beständen der HAB durch die DFG derzeit in der Begutachtung.
[5] Parsing der json-Dateien im Dump: https://git.hab.de/beyer/vd17-dump (Stand: 10.07.2024). Der Durchschnitt beträgt 132,5 Seiten.
[6] Lauer, Limbach et al. (2024), wie Anm. 3).

der Bestände mit einem Standard-OCR-Workflow effizient verarbeitet werden können (vgl. dazu auch das Quiver-Dashboard[7] für Ergebnisse zur Erkennungsrate).
Besonderheiten wie Schriftmischungen, handschriftliche Annotationen oder beschädigte Seiten stellen für die OCR-Prozesse Herausforderungen dar. Spezielle Textgattungen, wie beispielsweise Leichenpredigten, erfordern aufgrund ihrer komplexen Typografie und ihres Layouts eine gezielte Anpassung des OCR-Workflows, um hohe Texterkennungsraten zu erzielen. Daraus ergibt sich, dass in den meisten Fällen eine hohe Genauigkeit mit einer Standardlösung erreicht werden kann, was den Gesamtaufwand erheblich reduziert. Bei der verbleibenden Restmenge ist eine sorgfältige und angepasste Herangehensweise u.a. mit dezentralen Lösungen erforderlich.

2.3 Abstimmung mit den VD-Trägerbibliotheken: VD-Zukunft, VD-Portal, VD-Index

In der zurückliegenden Projektphase haben die VD-Trägerbibliotheken und die OCR-D-Koordinierung auf verschiedenen Ebenen und in unterschiedlichen Formaten ihre Bedarfe vorgestellt und ihre Planungen abgestimmt. Dabei wurden ihre jeweiligen Zuständigkeiten definiert und die zur Realisierung der Planung notwendigen Schritte festgelegt. Ein Umstand, der den Austausch erleichtert, besteht darin, dass die meisten Einrichtungen aus der OCR-D-Koordinierung und ein Teil der im OCR-Bereich insgesamt engagierten Institutionen sich mit den VD-Trägerbibliotheken und den VD-Bibliotheken überschneidet. Im engeren Kreis der OCR-D-Koordinierung und der VD-Trägerbibliotheken betrifft das die SBB zu Berlin, die SUB Göttingen und die HAB Wolfenbüttel. Maßgeblich im OCR-Bereich aktive Häuser wie die BSB München oder die ULB Halle, die nicht Teil der OCR-D-Koordinierung sind, sind dabei ebenso engagiert.

Beim Austausch der VD-Bibliotheken zur Optimierung der VD, vor allem mit dem Ziel eines gemeinsamen VD-Portals, sind in der hierzu engagierten Arbeitsgruppe „VD Zukunft“ die meisten der oben genannten Häuser vertreten und bringen sich in die Planungen mit ein.

Bei den VD-Partnertreffen der letzten Jahre hat die OCR-D-Koordinierung regelmäßig über ihre Arbeit berichtet, Fragen zu OCR-Bedarfen eingebracht und konnte so wertvolle Anregungen aus dem Kreis der VD-Bibliotheken in die eigene Planung einbeziehen. Zwei Veranstaltungen haben besonders konkrete Ergebnisse für die planerische „Schnittstelle“ zwischen der Volltexterstellung und den VD-Bibliotheken hervorgebracht: Zum einen das Göttinger DFG-Rundgespräch am 24. und 25.08. 2023 sowie die Sitzung der VD-Partner vom 8. April 2024 in Berlin. Bei dieser Präsenzveranstaltung mit Berichten über die „Planung zur Vervolltextung der VD“ Seitens OCR-D und über die Planungen und den Stand des Gesamtprojektes VD Zukunft Seitens der VD-Trägerbibliotheken wurden außerdem substantielle Zuarbeiten für die OCR-D-Konzepte durch eine Umfrage[8] unter den VD-Bibliotheken zu ihren Bedarfen und Planungen im OCR-Bereich und durch die Klärung von Volltext-relevanten Fragen[9] der OCR-D-Koordinierung mit den VD-Trägerbibliotheken angestoßen.

[7] https://ocr-d.de/quiver-frontend/#/workflows?view=table
[8] So wurden Antworten auf Fragen nach bereits integrierten OCR-Anteilen in den Digitalisierungsworkflows, nach der Implementierung von OCR-D, den Interessen und Kompetenzen bzgl. Ground Truth, dem Beratungsbedarf, dem Interesse an einem zentralen OCR-Dienst, dem für den Einsatz eines solchen Dienstes gewünschten Zeithorizonts und für OCR-Prozesse vorhandenem Personal gesammelt.
[9] Dabei kamen Aussagen zu OCR-D-Monita zustande, die auch Punkte aus dem Feedback des AWBI betreffen, darunter zur Bereitstellung der von Google generierten und von der BSB übernommenen Volltexte, zu den Maßnahmen der VD-Trägerbibliotheken für die Koordination der Volltextdigitalisierung, zu dem inzwischen angestoßenen Diskussionsprozess für die Formatumsetzung von Volltextnachweisen, die von den VD-Trägerbibliotheken übernommene Verantwortung im Arbeitsfeld der Volltexterstellung.

In der Folge fanden unter maßgeblicher Mitwirkung aus der Wissenschaft und den VD-Trägerbibliotheken - mit zunächst teilweise von einander abweichenden Vorstellungen mehrere Gesprächsformate zwischen den oben genannten Einrichtungen zu den Fragen rund um ein konkretes Projekt zur Gestaltung eines gemeinsamen VD-Portals bzw. für einen gemeinsamen Index statt. Die Überlegungen umfassen nicht nur eine gemeinsame Suchmöglichkeit in allen drei VD (bei denen auch Volltexte einbezogen werden sollen), sondern auch Bestrebungen zur Angleichung der abweichenden Datenmodelle, die sich bei der Nutzung der Daten schon länger als Hindernis erwiesen haben.[10] Dabei bringen sich zunehmend die Deutsche Digitale Bibliothek (DDB) und das FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH mit in die Planungen ein. Nach mehreren Videokonferenzen, folgte am 3. Juli 2024 der in Deutschen Nationalbibliothek, Frankfurt am Main, ein Präsenzworkshop, weitere Online-Folgetermine und zuletzt die Institutionalisierung einer Schreibgruppe zum Verfassen eines Projektantrags für das VD-Portal (Kickoff am 22. Oktober 2024).

3 REALISIERUNG

3.1 VD-Bestand mit vorhandenen Bilddigisalitaten

Die oben (Tabelle 1) referenzierten Quantitäten zeigen, dass insgesamt 586.663 Titel (etwa 80 % der katalogisierten VD-Bestände) bereits über Bilddigitalisate verfügen. Ein Teil der Titel, wie beispielsweise die von der Bayerischen Staatsbibliothek München (BSB) bearbeiteten, wurden bereits vollständig volltextdigitalisiert. Die verbleibenden Titel mit Bilddigitalisaten, jedoch ohne OCR, umfassen etwa 58 Millionen Seiten. Die Volltextdigitalisierung dieser Seiten soll nach dem vorliegenden Konzept zeitnah gestartet und rasch umgesetzt werden.

Würde man den OCR-Prozess sequentiell mit einer Verarbeitungsgeschwindigkeit von 1 Seite pro Minute durchführen, würde die Bearbeitung von 58 Millionen Seiten ungefähr 110 Jahre in Anspruch nehmen. Die Implementierung der OCR-D-Software in einer Hochleistungsrechner (HPC)-Umgebung ermöglicht dagegen eine erhebliche Beschleunigung der Volltextdigitalisierung durch bessere Leistung und Parallelisierung, wie es im OPERANDI-Projekt von der Niedersächsischen Staats- und Universitätsbibliothek Göttingen (SUB) und der Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) erfolgreich getestet wurde.[11]

Pro Seite werden im Durchschnitt bei Berücksichtigung aller Parameter 45 Sekunden benötigt (vgl. Tabelle 2, folgende Seite). Durch die Parallelisierung von 25 Workflows und bei einem Uptime-SLA für die technische Infrastruktur von 95 % können die genannten 58 Millionen Seiten in etwa 42 Monaten verarbeitet werden. Für die Realisierung dieses Prozesses sind außerdem Zeitaufwände für Datenlogistik, technische Wartung und Fehlerbehebung zu berücksichtigen, wodurch realistisch mit einer Gesamtbearbeitungszeit der vorhandenen Bilddigitalisate von vier Jahren zu rechnen ist.

[10] Lauer, Limbach et al. (2024), S. 515f.
[11] Zu den Ergebnissen siehe OPERANDI auf Github: https://github.com/subugoe/operandi?tab=readme-ov-file#operandi; OLA-HD Website: https://ola-hd.ocr-d.de/; OLA-HD auf Github: https://github.com/subugoe/OLA-HD-IMPL#

HPC-Ergebnisse der Leistungstests			
	Workspace klein	Workspace medium	Workspace large
Identifier	PPN1023134829	PPN180446581X	PPN55554432X_0001
Schriftart	Fraktur	Fraktur	Fraktur
Jahr	1743	1779	1708
Seitenzahl	140	412	1303
Durchführungszeit	57m 50s	3h 6m 30s	11h 24m 31s
Durchführungszeit pro Seite	24,79s	27,16s	31,52s

Tabelle 2: Durchführungszeit mit HPC bei 100 % Uptime

Die über die Digitalisierungssoftware und die implementierten Schnittstellen (METS-Import) gesammelten Bilddigitalisate sowie Strukturdaten werden in einem bibliotheksspezifischen Speicher abgelegt. Von dort aus können die Daten, einschließlich der benötigten Workflowparameter, an einen OCR-as-a-Service-Dienst gesendet und dort prozessiert werden. Nach Abschluss der Verarbeitung werden die Strukturdaten um einen zusätzlichen Dateiverweis auf den generierten Volltext erweitert.

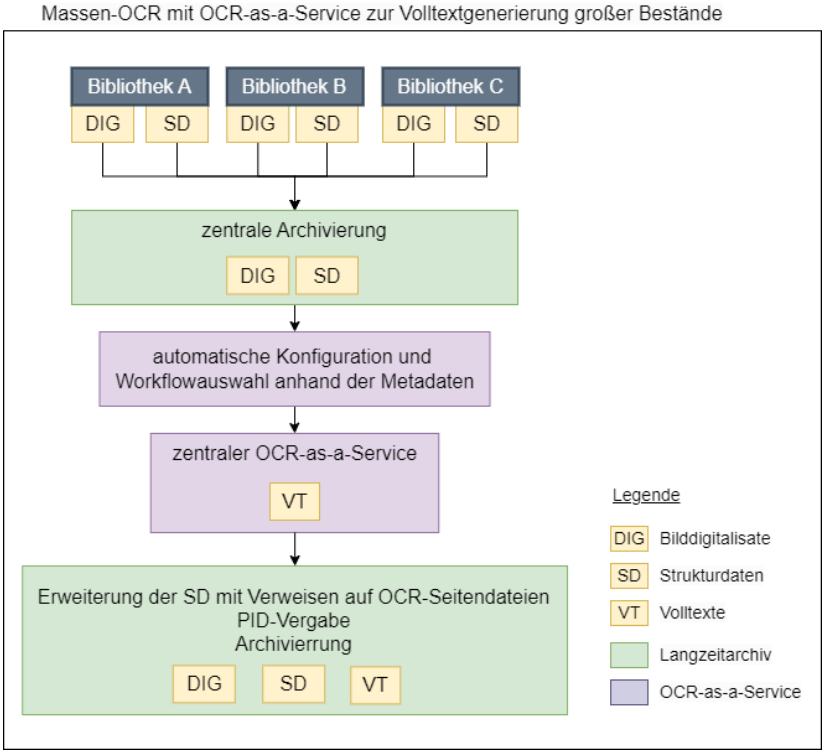

Abbildung 1: Workflow der Massen-OCR mit OCR-Service zur Volltextdigitalisierung großer Bestände

Die produzierten Volltexte werden zusammen mit den Digitalisaten und den ergänzten METS-Dateien zentral in dem dazu eingerichteten Archiv gespeichert. Dabei erhalten die generierten Volltexte einen Persistent Identifier (PID), der eine eindeutige Identifizierung und den dauerhaften Zugriff auf diese Daten gewährleistet. Über dieses Archiv können die Daten sowohl manuell als auch maschinell abgerufen und lokal gespeichert werden. So können die Daten sowohl den Bibliotheken als auch anderen Nutzern bzw. Vorhaben (z. B. VD-Suchraum) bereits während des Projektverlaufs bereitgestellt werden (zur grafischen Darstellung dieses Ablaufes siehe Abbildung 1).

Als ein möglicher Anbieter, der seine Leistungsfähigkeit für die Aufgabe unter Beweis gestellt und sich dieser Aufgabe im Rahmen von OCR-D angenommen hat, steht die GWDG zur Verfügung.

3.2 VD-Bestand ohne Bilddigitalisate, ohne Katalogisierung sowie Spezialfälle

Für 150.071 Titel des katalogisierten VD-Bestands fehlen mit Stand von September 2024 die Bilddigitalisate. Dies entspricht ca. 19 Millionen Seiten, für die OCR erst nach bzw. während der Retrodigitalisierung erstellt werden kann. Weiter wird mit 443.266 Titeln gerechnet, die bisher nicht erfasst sind (VD-Nova)[12]. Dies entspricht rund 78,5 Millionen Seiten, wobei die tatsächliche Anzahl der Seiten, für die Volltext zu erstellen sind, aufgrund von Mehrfachexemplaren voraussichtlich geringer ausfallen wird. Außerdem sind besondere Fälle zu betrachten, wie z. B. Leichenpredigten, die aufgrund ihrer komplexen Typografie, Mehrsprachigkeit oder wegen besonderen Layouts eine gezielte Anpassung erfordern, um hohe Texterkennungsraten zu erzielen. Für die Volltextdigitalisierung dieser VD-Bestände soll neben OCR-as-a-Service (zentral) auch On-Premise-Deployment (dezentral/ lokal) bzw. eine Kombination aus beiden Modellen zum Einsatz kommen. Eine detaillierte Beschreibung der Modelle ist im Konzept zur Verstetigung der OCR-D-Software dargestellt (Kapitel 5).

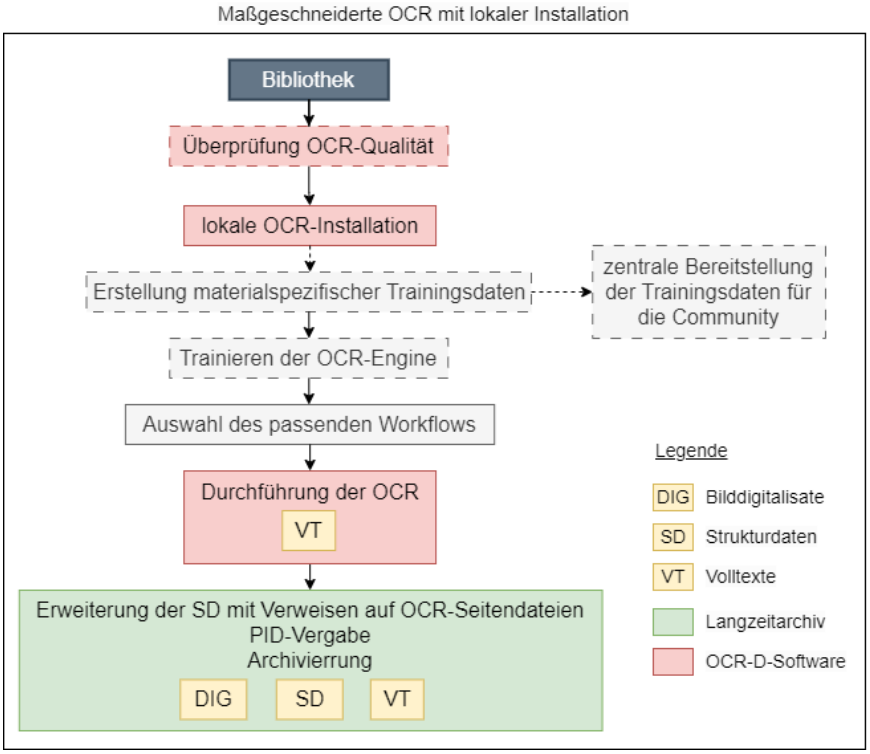

Abbildung 2: Workflow für maßgeschneiderte OCR-Generierung mittels lokaler Installation der OCR-D-Software

Während OCR-as-a-Service für Massenverarbeitung optimiert ist und einen hohen Durchsatz bei guter Erkennungsrate liefert, bietet das On-Premise-Deployment lokale Integration, bessere individuelle Anpassbarkeit und, mit Hilfe von gezielten Trainingsmöglichkeiten und Ground Truth (GT), eine potentiell höhere Erkennungsrate für sinnvoll gewählte Korpora (zur grafischen Darstellung dieses

[12] Lauer, Limbach et al. (2024), S. 2; Beyer, Jürgen: How Complete are the German National Bibliographies for the Sixteenth and Seventeenth Centuries (VD16 and VD17)? In: The Book Triumphant, 2011, S. 57–77.

Ablaufes siehe Abbildung 2.) Somit eignet sich On-Premise-Deployment für diejenigen Einrichtungen und Vorhaben, bei denen die lokale Integration in bestehende Systeme und Workflows besonders wichtig ist, oder für Fälle, in denen für wissenschaftliche Auswertungen spezifische Anforderungen erfüllt werden müssen. Insbesondere bei komplexerem Layout oder besonderer Typographie, die eine maßgeschneiderte Verarbeitung erfordern, ist ein On-Premise-Deployment von Vorteil.

Zudem ist ein hybrider Ansatz möglich: Dabei wird zunächst OCR-as-a-Service für die Massenverarbeitung genutzt, um daran anschließend Korpora von unzureichend erkannten Titeln mit einem On-Premise-Deployment gezielt nachzubearbeiten. Die Nutzung besonderer Bestandskompetenz ermöglicht dabei die Erstellung von spezialisierter GT und eines angepassten Workflows. Die Ergebnisse dieser Form der Volltexttransformation sollen durch Bereitstellung der GT sowie gegebenenfalls angepasster Software oder Workflows der Community zugänglich gemacht werden. Die GT sollte gemäß den OCR-D-GT-Richtlinien[13] erstellt und öffentlich verfügbar gemacht werden. Hierbei können die bestehenden OCR-D-Ressourcen, insbesondere das GT-Repository[14], genutzt werden, um eine breite Nachnutzung und Weiterentwicklung durch die Community zu fördern. Lokale Installationen können auch im Rahmen der Bild- und Volltextdigitalisierung bisher nicht bilddigitalisierter Bestände oder bei der Digitalisierung und Erschließung von VD-Nova eingesetzt werden. Über das Förderprogramm 'Digitalisierung und Erschließung' besteht die Möglichkeit, bei der DFG sammlungsbezogen eine Finanzierung für die Volltextdigitalisierung zu beantragen.[15]

3.3 VD-Bestand mit vorhandener OCR

Bei den Planungen wurde darauf geachtet, bestehende OCR-Bestände zu berücksichtigen, um Doppelarbeit (redundante Volltextdigitalisierungen) zu vermeiden: So existieren bereits OCR-Daten für Teile der VD-Bestände, wie etwa aus der Google-Kooperation der Bayerischen Staatsbibliothek München (BSB)[16] und ihrer bayerischen Partnereinrichtungen sowie von Institutionen, die eigene OCR-Workflows umgesetzt haben (z. B. Universitäts- und Landesbibliothek Sachsen-Anhalt (ULB Halle)). Durch die OCR-D-Koordinierung wurden mit diesen Einrichtungen Standards, Schnittstellen und Erschließungsfragen abgestimmt, um gemeinsame Bereitstellungs-, Such- und Nachnutzungsmöglichkeiten sicherzustellen und redundante Aufwände zu vermeiden.

In dem DFG-geförderten Göttinger Rundgespräch am 24. und 25.08. 2023 und im Austausch mit den VD-Trägerbibliotheken hat sich die BSB dazu bereit erklärt, ihre Bild- und Textdaten für eine gemeinschaftliche VD-Präsentation bereitzustellen. Zudem werden die Google-Volltexte jährlich neu prozessiert, um Qualitätsverbesserungen zu integrieren und bisher nicht bearbeitete Texte hinzuzufügen. Auch diese Ergebnisse der iterativen Prozessierung sollen bereitgestellt werden.

Außerdem hat die ULB Halle im Rahmen eines OCR-D-Projekts in erheblichem Umfang Volltexte für Bilddigitalisate aus dem VD18 erstellt:[17] Im Rahmen dieses Projektes wurden zwischen 2021-2024 über 6,25 Millionen Seiten mittels der in den OCR-D Projektphasen entwickelten Tools um Volltexte angereichert. Zu Messung der Qualität wurde ein Referenzkorpus aus 1.600 GT-Seiten mit Text- und

[13] https://github.com/OCR-D/gt-guidelines
[14] https://github.com/OCR-D/gt_structure_text/releases
[15] https://www.dfg.de/de/foerderung/foerdermoeglichkeiten/programme/infrastruktur/lis/lis-foerderangebote/digitalisierung-erschliessung
[16] Über Google bzw. die BSB München prozessierte Volltexte werden nicht durch OCR-D erneut prozessiert.
[17] https://bibliothek.uni-halle.de/ueber-uns/projekte/abgeschlossene-projekte/

Layoutinformationen aufgebaut, die in Zusammenarbeit mit der OCR-D-Koordinierung und den Trägerbibliotheken in das zentrale Repositorium eingepflegt und mit den Metadaten des VD-Katalogs verknüpft werden.

Ein weiterer Akteur, die Österreichische Nationalbibliothek (ÖNB), liefert Bilddigitalisate und Katalogdaten zum VD16 und verfügt über Volltexte für etwa 20.118 Titel aus ihrer eigenen Google-Kooperation. Da jedoch eine erhebliche Überschneidung mit der Sammlung der BSB im VD16-Bereich hinsichtlich der Mehrfachexemplare besteht, sind hier nur begrenzt Synergien zu erwarten.

4 EMPFEHLUNGEN AN DIE DFG

4.1 Empfehlungen nach Abschluss der dritten Förderphase

Mit dem Abschluss der dritten Phase der OCR-D sind die Ziele dieses Projektabschnitts erreicht: Die konzeptionelle und technische Vorbereitung der Volltexttransformation der VD sind abgeschlossen. Die Anzahl der VD-Daten, mit denen die Volltextdigitalisierung unmittelbar gestartet werden soll, ist ermittelt (58 Mio. Seiten), ein zentral angebotener Service für die Massenbearbeitung und lokale Möglichkeiten der Volltextdigitalisierung sind konzeptionell vorbereitet und technisch erprobt, die Evaluation der Gesamtdauer von einer HPC-gestützten Massenvolltextdigitalisierung ist ermittelt (42 Monate), die Verstetigung der Software mit Vereinsstrukturen ist in die Wege geleitet (Kitodo e. V.) und die Speicherung, die Versionierung und die Archivierung der Volltexte ist ermöglicht (OLA-HD).
Vor diesem Hintergrund empfiehlt die OCR-D-Koordination der DFG die Förderung der systematischen Volltextdigitalisierung von VD-Beständen mit vorhandenen Bilddigitalisaten im Rahmen einer vierten Phase der OCR-D-Koordination (dazu im Folgenden). Nur auf diesem Weg können die datenlogistischen und zeitlichen Herausforderungen angesichts der großen Mengen qualitativ befriedigend und wirtschaftlich vertretbar gelöst werden können. Zudem können die erstellten Volltexte nur so zeitnah und basierend auf Open Source-Strukturen für die weitere wissenschaftliche Nutzung bereitgestellt werden.
Darüber hinaus empfiehlt die OCR-D-Koordination eine erleichterte Förderung für die Volltextdigitalisierung derjenigen VD-Bestände ohne Bilddigitalisat (sofern dabei die Volltexterstellung eingeschlossen ist) und für die Neukatalogisierung bisher unkatalogisierter VD-relevanter Bestände, der VD-Nova (einschließlich Bilddigitalisierung, sofern dabei ebenfalls die Volltexterstellung durchgeführt wird) über das Förderprogramm „Digitalisierung und Erschließung“. Der Nachweis dafür, dass für derartige Projekte die wissenschaftliche Relevanz bzw. ein wissenschaftliches Interesse vorliegt, kann angesichts vielfach geäußerter und publizierter Voten aus der Wissenschaft als erbracht gelten.[18] Die Volltextdigitalisierung kann in diesen Fällen entweder durch lokale OCR-D-Installationen oder den Einsatz des zentralen OCR-Dienstes erfolgen, um eine effiziente und umfassende Erschließung der Werke zu gewährleisten.

[18] Zur wissenschaftlichen Relevanz der VD wurden zahlreiche Stellungnahmen publiziert, u. a. Beyer et al. (2022), S. 85f, S. 90; Lauer, Limbach et al. (2024), S. 507f oder Stäcker (2018), S. 131. Stäcker (2018) beschreibt die VD „als Forschungsgrundlage [...], die im europäischen und internationalen Zusammenhang ihresgleichen sucht" (S. 131), Lauer, Limbach et al. (2024) charakterisieren sie als „unverzichtbare(n) Arbeitsmittel aller Wissenschaften, die sich mit der frühen Neuzeit befassen (S. 507f), während Bayer et al. (2022) von den mit Teilnehmenden aus der Wissenschaft durchgeführten VD-Rundgesprächen berichten, dass für die VD „die Bedeutung als wissenschaftlicher Fundus“, als „wichtige historische und sozialgeschichtliche Quelle“ (S. 85f) und als „relevante Arbeitsinstrumente“ (S. 90) betont wurde.

Als Bedingungen für eine DFG-Förderung sollen der offene Zugang zu Text-, Bild-, Meta- und GT-Daten im Ergebnis von Projekten mit VD-Volltexterstellung sowie die Nutzung persistenter Adressen (DOI, URN oder Handle), ebenso wie die Einhaltung der Praxisregeln Digitalisierung zugrunde gelegt werden.
Weiter wird empfohlen, das Optimieren der Volltexterschließung durch iterative OCR-Bearbeitung zu fördern. Diese soll in begründeten Fällen und nach einer noch festzulegenden Frequenz[19] entweder erneut zentral oder bei abgestimmtem Vorgehen auch dezentral erfolgen.
Weitere Empfehlungen richten sich nicht an die DFG, sondern an Institutionen, die im Bereich der Erschließung, Bild- und Volltextdigitalisierung engagiert sind. Sie werden in den zuständigen Gremien adressiert und hier nicht eigens aufgeführt.[20]

4.2 Vierte Phase: Durchführung der Volltextdigitalisierung der VD und Abschluss der Verstetigung

4.2.1 Vierte Förderphase: Grundlagen

In der empfohlenen vierten OCR-D-Förderphase[21] werden die vorbereiteten Maßnahmen zur umfassenden Volltextdigitalisierung der VD in der Praxis umgesetzt. In dieser Projektphase wird die Governance von einem Koordinierungsteam gesteuert werden, das Beiträge der Community mittels der im Kitodo-Verein etablierten Strukturen und Gremien vermittelt.[22] Die Volltexte für alle verfügbaren Bilddigitalisate der VD werden erstellt, einschließlich der dazu notwendigen Vor- und Nacharbeiten (Harvesting, Erstellung von Identifiern, Speicherung/Archivierung der Bild- und Textdaten, Zurückspielen von Meta- und Strukturdaten in die Nachweissysteme). Außerdem wird die dauerhafte Nutzung der OCR-D-Software durch die eingeleitete Verstetigung, möglichst im Rahmen des Kitodo-Vereins, sichergestellt. Der Schwerpunkt liegt dabei auf der systematischen Volltextdigitalisierung von VD-Beständen mit vorhandenen Bilddigitalisaten sowie auf dem Abschluss der Verstetigung der Ergebnisse aus den OCR-D-Projekten. Die folgenden Ziele stehen dabei im Mittelpunkt:

1. Vollständige Volltextdigitalisierung der VD-Bestände mit Digitalisat und ohne vorliegende OCR innerhalb von vier Jahren durch den Einsatz von HPC und Parallelisierung sowie die sukzessive Bereitstellung und die Archivierung der OCR-Daten über OLA-HD.
2. Begleitung der Verstetigung der OCR-D-Software im Kitodo-Verein durch Wissenstransfer von Projektinhalten und -strukturen in die Vereinsstrukturen und projekthafte Unterstützung des Vereins während der Übergangsphase seitens OCR-D.
3. Sicherstellung der Nutzung von OCR-D in verschiedenen Digitalisierungsworkflows, damit die Volltextdigitalisierung in künftigen DFG-geförderten Digitalisierungsvorhaben berücksichtigt werden kann.

Ein weiteres Ziel der vierten Phase ist die Unterstützung des geplanten Vorhabens „VD-Suchraum“.[23] Durch die zentrale Erfassung und Bereitstellung der Bild-, Struktur- und OCR-Daten kann das Vorhaben gezielt unterstützt werden. Die erfassten Daten werden dem Projekt sukzessiv zur Verfügung gestellt,

[19] Im Fall der Volltexterstellung mittels der Google-Kooperation der BSB München erfolgt sie jährlich.
[20] Zum Beispiel wird empfohlen, dass der eingeleitete VIGO-Prozess zur Revision, Redaktion und Erstellung der Praxisregeln Digitalisierung fortgesetzt und möglichst intensiviert werden soll. Dabei sollen die von der OCR-D-Koordinierung erarbeiteten Richtlinien, Spezifikationen und Empfehlungen berücksichtigt werden.
[21] Die Institutionen der bisherigen Kooperationspartner in der OCR-D-Koordinierung erklären ihre Bereitschaft dazu, für eine etwaige Ausschreibung zur Förderung der vierten Phase (ggf. neben anderen Bewerbungen) entsprechende Anträge einzureichen und sich im Fall der Förderung weiter zu engagieren..
[22] Weitere Angaben zu Governance und Communitybeteiligung finden sich in Kapitel 6.6 des Verstetigungskonzepts.
[23] Lauer, Limbach et al., S. 516

um die Suchindizes für VD-Bestände zu erstellen, umfassend zu erweitern und die Suchmöglichkeiten - auch für Volltexte - kurzfristig zugänglich zu machen. Die an der OCR-D-Koordinierung beteiligten Einrichtungen sind in die Planungsphase des Vorhabens "VD-Suchraum" involviert und können im Rahmen der empfohlenen vierten Förderphase von OCR-D potenzielle Synergien mit dem Projekt "VD-Suchraum" gezielt befördern. Dies ermöglicht eine koordinierte Nutzung und Weiterentwicklung der bestehenden Infrastruktur, um die notwendige Datenintegration zu leisten und den Abschluss beider Vorhaben zu erreichen.

4.2.2 Vierte Förderphase: Projektplan

AP 1: Projektmanagement und Kommunikation

Im ersten Arbeitspaket werden die organisatorischen Strukturen des Projekts aufgebaut. Das beinhaltet die Zusammenstellung von des Teams, die Zuweisung von Verantwortlichkeiten und die Einführung von neuen bzw. Fortsetzung von etablierten von Kommunikationsformaten. Zentrale Kommunikationswege mit internen und externen Stakeholdern werden etabliert, um erzielte Fortschritte transparent zu dokumentieren, das Berichtswesen abzusichern und die Grundlage für eine offene inhaltliche Kommunikation aller Beteiligten zu schaffen. Ein zentraler Aspekt ist der enge und stetige Austausch mit den Trägerbibliotheken, um deren Anforderungen und Rückmeldungen kontinuierlich in den Projektverlauf zu integrieren.

- AP 1.1: Einrichten der Projektorganisation (Projektstruktur, Teams, Verantwortlichkeiten)
- AP 1.2: Etablierung regelmäßiger Kommunikationsformate (Meetings, Protokolle, Berichte)
- AP 1.3: Stakeholder-Management und externe Kommunikation (Workshops, Webinare, Dokumentation)
- AP 1.4: Zentrale Erfassung und Koordination von Daten aus VD-Bibliotheken. Dies umfasst das Sammeln von optimierten OCR-Workflows und GT-Material zur Qualitätssicherung und Workflow-Optimierung. Diese Daten werden von den VD-Bibliotheken bereitgestellt, um die kontinuierliche Verbesserung der OCR-D-Verfahren sicherzustellen.
- AP 1.5: Überwachung des Fortschritts und Qualitätssicherung durch Feedback-Schleifen und regelmäßige Abstimmungen mit den beteiligten Bibliotheken.

– Meilenstein 1 (MS1): Projektorganisation vollständig eingerichtet, erste Statusberichte erstellt (Monat 3)

AP 2: Volltextdigitalisierung der VD-Bestände

Dieses Arbeitspaket zielt auf die systematische Volltext-Digitalisierung der VD-Bestände, die bisher zwar mit Bilddaten versehen sind, für die aber noch keine Volltexte vorliegen. Hierzu wird die OCR-D-Software auf einer HPC-Infrastruktur implementiert, um eine parallele Verarbeitung der Bestände zu ermöglichen. Die Volltextdaten werden anschließend zentral gespeichert, archiviert und regelmäßig qualitätsgesichert.

- AP 2.1: Aufnahme und Analyse der zu digitalisierenden Bestände (VD ohne OCR)
- AP 2.2: Überführung der OCR-D-Software in Produktivbetrieb
- AP 2.3: Durchführung der OCR-Prozesse für VD-Bestände, inkl. Qualitätssicherung
- AP 2.4: Zentrale Speicherung und Archivierung der OCR-Daten in OLA-HD (einschließlich PID)

– Meilenstein 2 (MS2): 50 % der VD-Bestände ohne OCR digitalisiert und archiviert (Monat 24)

– Meilenstein 3 (MS3): Vollständige OCR-Digitalisierung der VD-Bestände abgeschlossen (Monat 42)

AP 3: Verstetigung der OCR-D-Software im Kitodo-Verein

Im Rahmen dieses Arbeitspakets erfolgt der Übergang von Release-Management und Softwareentwicklung der OCR-D-Software in die Strukturen des Kitodo-Vereins. Dies beinhaltet den Wissenstransfer durch Schulungen und Workshops sowie technische Unterstützung bei der Integration der OCR-D-Software in die bestehenden Kitodo-Workflows. Ziel ist es, die langfristige Nutzung und Weiterentwicklung der Software im Kitodo-Verein sicherzustellen und sie breiter in der Community (GLAM-Institutionen und Wissenschaft) nutzbar zu machen.

- AP 3.1: Wissenstransfer und Schulungen für die Nutzung von OCR-D im Kitodo-Verein
- AP 3.2: Unterstützung bei der technischen Integration von OCR-D in Kitodo-Workflows
- AP 3.3: IT-Unterstützung für die Anpassung und Wartung von OCR-D in Kitodo
- AP 3.4: Dokumentation und Übergabe der Projektstrukturen an den Verein

– Meilenstein 4 (MS4): Kitodo ist vollständig in der Lage, OCR-D zu nutzen und zu verwalten (Monat 18)

AP 4: Zentrale Speicherung und Bereitstellung der Daten

Hier wird die systematische Bereitstellung der OCR-Daten realisiert. Die erfassten Daten werden - sobald sie aufbereitet sind - sukzessive zur Verfügung gestellt, um die Suchindizes der VD-Bestände kontinuierlich und zeitnah zu erweitern und zu optimieren. Über dieses Arbeitspaket wird auch das zukünftige Vorhaben "VD-Suchraum" unterstützt.

- AP 4.1: Bereitstellung der OCR-D-Daten in OLA-HD
- AP 4.2: Unterstützung bei der Erweiterung der Suchmöglichkeiten (OCR-basierte Volltextsuchen)
- AP 4.3: Koordination der Synergien zwischen OCR-D und VD-Suchraum (Datenintegration)

– Meilenstein 5 (MS5): Erster Datensatz für den VD-Suchraum verfügbar und integriert (Monat 24)

– Meilenstein 6 (MS6): Suchindex und Volltextsuche für alle OCR-D-VD-Daten im Suchraum verfügbar (Monat 42)

AP 5: Qualitätssicherung und Monitoring

Dieses Arbeitspaket stellt sicher, dass die Qualität der OCR-Daten auf hohem Niveau bleibt. Es werden automatisierte Fehlerprüfungen implementiert, um frühzeitig potenzielle Probleme zu identifizieren. Manuelle Nachbearbeitungen und Workflow-Optimierungen, basierend auf GT-Daten, werden durchgeführt, um eine kontinuierliche Verbesserung der OCR-D-Workflows zu gewährleisten. Zum Abschluss dieser vierten Projektphase wird der erreichte Stand evaluiert. Die Evaluation und daraus abgeleitete Vorschläge für das weitere Vorgehen werden den Gremien der DFG (AWBI) zur Entscheidung vorgelegt.

- AP 5.1: Implementierung von automatisierten Checks zur Fehlerquote der OCR-Daten
- AP 5.2: Manuelle Qualitätskontrolle und Nachbearbeitung (OCR4all, LAREX, Ground Truth)
- AP 5.3: Optimierung der Workflows durch Feedback-Schleifen und Nutzung von Quiver für Training

– Meilenstein 7 (MS7): Verbesserung der OCR-D-Workflows durch GT-Training (Monat 30)

– Meilenstein 8 (MS8): Abschluss-Evaluation und Übergabe der Empfehlungen an die DFG-Gremien (Monat 42)

4.2.3 Zusammenfassung der Meilensteine

– MS1: Projektorganisation und Kommunikation eingerichtet (Monat 3)
– MS2: 50 % der VD-Bestände ohne OCR digitalisiert (Monat 24)
– MS3: Vollständige OCR-Digitalisierung der VD-Bestände (Monat 42)
– MS4: Verstetigung von OCR-D in Kitodo abgeschlossen (Monat 18
– MS5: Erster Datensatz für den VD-Suchraum verfügbar und integriert (Monat 24)
– MS6: Suchindex und Volltextsuche für alle OCR-D-VD-Daten im Suchraum verfügbar (Monat 42)
– Meilenstein 8 MS7: Verbesserung der OCR-D-Workflows durch Ground Truth Training (Monat 30)
– MS8: Abschluss-Evaluation und Übergabe der Empfehlungen an die DFG-Gremien (Monat 42)

4.2.4 Vierte Förderphase: Kosten

Für die Laufzeit von 48 Monaten wird eine verkleinerte Koordinierungsstruktur für das OCR-D-Projekt vorgeschlagen, die eine effektive und nachhaltige Umsetzung der Projektziele gewährleistet. Die Teamstruktur setzt sich aus folgenden Rollen und Personalkapazitäten zusammen:

- 1,0 (E13) für Koordination und Kommunikation: Verantwortlich für das übergeordnete Projektmanagement, die Organisation von Kommunikationsformaten, die zentrale Sammlung und Verteilung von Informationen sowie das Stakeholder-Management.
- 1,0 (E13) für OCR-Expertise: Zuständig für die technische Implementierung und Anpassung der OCR-D-Software, Qualitätssicherung und die Integration in verschiedene Digitalisierungsplattformen sowie die Schulung der Anwendenden.
- 1,0 (E13) für zentralen Service: Diese Rolle deckt administrative und technische Aufgaben ab, darunter die Unterstützung bei der Implementierung und Archivierung der OCR-Daten, die Optimierung der Workflows und die Sicherstellung des kontinuierlichen Betriebs der Software.
- 1,0 (E10) für Metadaten, Kataloge und Datenhandling: Fokussiert auf die Verarbeitung und Integration der digitalisierten Daten, die Verwaltung der Metadaten sowie die Unterstützung bei der Qualitätssicherung der Daten.
- 0,5 (E13) IT-Unterstützung für die Verstetigung im Kitodo-Verein (für 18 Monate): Wissenstransfer, technische Unterstützung bei der Integration und Wartung der OCR-D-Software in den Kitodo-Workflows, um die langfristige Nutzung sicherzustellen.

Die geplanten Personalkosten belaufen sich auf 1.423.950 €, ergänzt durch 221.130 € für OCR-as-a-Service Sachkosten (Aufstockung der IT-Infrastruktur) sowie 30.000 € für sonstige Sachkosten.

Diese Struktur stellt sicher, dass das Projekt effizient gemanagt und alle notwendigen technischen und administrativen Aufgaben abgedeckt werden, um eine erfolgreiche Verstetigung und Integration der OCR-D-Software zu ermöglichen.

Die folgende Tabelle 3 zeigt, wie die verschiedenen Rollen auf die Arbeitspakete (AP) verteilt sind. Jede Rolle übernimmt spezifische Aufgaben, die den jeweiligen Expertisen entsprechen und zur Zielerreichung des Projekts beitragen. Die Zuordnung der Personalkapazitäten in Personenmonaten

(PM) spiegelt den geplanten Aufwand für jedes Arbeitspaket wider und gewährleistet eine ausgewogene Verteilung der Projektlast.

Rollen / AP in Personenmonate (PM)	AP1	AP2	AP3	AP4	AP5	SUMME PM pro Rolle
Koordination & Kommunikation	36	2	4	3	3	48
OCR-Experte	1	36	3	2	6	48
Betreuung zentraler Service	1	39	3	3	2	48
IT-unterstützung für die Verstetigung (Release-Management & Weiterentwicklung)	2	1	12	1	2	18
Metadaten, Datenhandling	2	12	2	24	8	48
SUMME PM pro AP	42	90	24	33	21	

Tabelle 3: Zuordnung der geplanten Projektrollen zu den Aufgaben und Aufwände in Personenmonate

5 EXECUTIVE SUMMARY

Das Konzept zur Volltexttransformation der Verzeichnisse der im deutschen Sprachraum erschienenen Drucke (VD 16, VD 17, VD 18) zielt auf eine koordinierte und effiziente OCR-Digitalisierung dieser historischen Bestände ab. Durch den Einsatz von Hochleistungsrechnern und dezentralen Lösungen sollen die etwa 58 Millionen Seiten, die bereits bilddigitalisiert vorliegen, aber noch nicht volltextdigitalisiert wurden, schnell und wirtschaftlich erfasst werden.

Ausgangslage und Vorarbeiten:

Seit 2015 entwickelt die OCR-D-Initiative Verfahren zur Texterkennung (OCR) für historische Bestände. Bisher wurden Software, Standards, Dokumentationen und technische Lösungen geschaffen, die es ermöglichen, eine großflächige Volltextdigitalisierung der VD-Bestände durchzuführen. Der Gesamtbestand umfasst über 736.000 katalogisierte Titel, von denen rund 586.000 über Bilddigitalisate verfügen. 358.894 dieser Titel müssen noch volltextdigitalisiert werden, was etwa 58 Millionen Seiten entspricht.

Realisierung:

Das Konzept schlägt einen hybriden Ansatz vor. Für die Mehrheit der Bilddigitalisate wird ein zentraler OCR-Dienst eingesetzt, der durch Hochleistungsrechner (HPC) in der Lage ist, große Mengen von Seiten parallel zu verarbeiten und sukkessive Ergebnisse innerhalb von vier Jahren zu liefern. Dies reduziert den Aufwand für die Volltextdigitalisierung erheblich. Dezentral installierte Lösungen sollen jedoch in Spezialfällen – wie bei komplexer Typografie oder handschriftlichen Annotationen – zum Einsatz kommen, um eine bessere Texterkennung zu gewährleisten. Diese duale Struktur soll sicherstellen, dass sowohl standardisierte als auch spezifische Anforderungen erfüllt werden.

Empfehlungen:

Das Konzept gibt umfassende Empfehlungen für die nächste Phase der OCR-D-Initiative. Zentral ist die Fortführung der systematischen Volltextdigitalisierung für Bestände, die bereits über Bilddigitalisate verfügen. Es wird empfohlen, dass zukünftige Bilddigitalisierungsprojekte direkt die Volltextdigitalisierung einschließen, um eine nahtlose Integration zu ermöglichen.

Zusätzlich betont das Konzept die Bedeutung von Kooperationen mit bereits existierenden OCR-Projekten, um doppelte Digitalisierungen zu vermeiden und Synergien zu nutzen. Bibliotheken sollen aktiv in die Digitalisierung eingebunden werden, um spezifische Anforderungen zu kommunizieren und um sicherzustellen, dass die Daten langfristig offen und nutzbar bleiben. Besonders hervorgehoben wird die Bedeutung der Qualitätssicherung: Manuelle und automatisierte Prüfungen sowie die Nutzung von Ground Truth (GT)-Daten sollen sicherstellen, dass die OCR-Ergebnisse kontinuierlich verbessert werden.

Vierte Projektphase:

Die vierte Förderphase der OCR-D-Initiative sieht die vollständige Digitalisierung der verbleibenden 58Millionen Seiten innerhalb von vier Jahren vor. Dies soll durch den Einsatz von Hochleistungsrechnern und parallel laufenden Workflows geschehen, was die Bearbeitungszeit erheblich verkürzt. Ziel ist es, diese Volltexte möglichst schnell und in hoher Qualität für die wissenschaftliche Forschung bereitzustellen.

Ein weiterer zentraler Aspekt der vierten Phase ist die dauerhafte Verstetigung der OCR-D-Software. Geplant ist, die Software in bestehende Strukturen, wie den Kitodo-Verein, zu integrieren, um ihre langfristige Nutzung in Bibliotheken und anderen wissenschaftlichen Einrichtungen sicherzustellen.

6 REFERENZEN

- Beyer, Hartmut, Bubenik, Claudia, Scheibe, Michaela: 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: Zeitschrift für Bibliothekswesen und Bibliographie 69 (2022), S. 82–91. DOI: http://dx.doi.org/10.3196/18642950206912112, zuletzt abgerufen am 14.10.2024
- 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
- Lauer, Gerhard, Limbach, Saskia, Reske, Christoph, Scheibe, Michaela, Weichselbaumer, Nikolaus (2024): Zukunft der VD – Vision einer forschungsadäquaten Nationalbibliographie der frühen Neuzeit. Bibliotheksdienst, 58 (9-10), 507-525. DOI: https://doi.org/10.1515/bd-2024-0076, zuletzt abgerufen am 14.10.2024
- OCR-D: https://ocr-d.de/
- OCR-D Koordinierungsprojekt: Konzept zur Volltexttransformation der VD vom 31.07.2020.
- SBB: VD 17-Partnertreffen. Präsentation am 12. Juni 2023.
- Stäcker, Thomas: Das VD17 at your fingertips: Der Masterplan. Nachgedanken zu einem paradigmatischen Digitalisierungsprogramm. In: Kooperative Informationsinfrastrukturen als Chance und Herausforderung. Festschrift für Thomas Bürger zum 65. Geburtstag. Hrsg. von Achim Bonte, Juliane Rehnolt. München 2018, S. 131–143. DOI: 10.26083/tuprints-00017470, zuletzt abgerufen am 14.10.2024
- 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/