OCR-D

Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition

KONZEPT ZUR VOLLTEXTTRANSFORMATION DER VD

31.07.2020

Koordinierungsprojekt

1. BASIS DER UNTERSUCHUNG

1.1. HINTERGRUND

Für die angestrebte Massenvolltextdigitalisierung der VD soll im Rahmen der 2. Förderphase des OCR-D-Projekts ein Konzept erstellt werden, das der DFG vorgelegt wird. Eine Umfrage unter den VD-Bibliotheken dient als Basis für einen ersten Konzeptentwurf, der im Mai 2020 mit den führenden VD-Bibliotheken diskutiert wird, bevor das Konzept finalisiert und an die DFG übergeben wird.

Die Umfrage richtet sich an zwei Zielgruppen: Zum einen werden die Trägerbibliotheken der VD zum Stand der Digitalisierungsarbeiten in dem von ihnen koordinierten VD befragt. Zum anderen werden die weiteren an den VD beteiligten Bibliotheken zu den Digitalisierungsarbeiten in ihren Häusern sowie ihrer Bereitschaft zu bzw. Erfahrung mit OCR-Projekten befragt.

1.2. VORGEHENSWEISE

Auf Basis der Vorumfrage zur ersten Teststellung der OCR-D-Software an ausgewähltenPilotbibliotheken wurden zwei Umfragen für das VD-Konzept erstellt. Die erste Umfrage (TBU) ist an die Trägerbibliotheken der VD gerichtet, die zweite Umfrage (WBU) ist für die weiteren VD-Bibliotheken bestimmt. Beide Umfragen sind weitgehend identisch und unterscheiden sich nur durch die Fragen zum Stand der Bilddigitalisierung in den VD, die lediglich den Trägerbibliotheken gestellt werden. Die TBU wurde am 3.12.2019 direkt an die 4 Trägerbibliotheken geschickt. Am gleichen Tag wurde die WBU über die VD17-Mailingliste an die übrigen beteiligten 41 Bibliotheken versandt. Darüber können fast alle VD-Bibliotheken erreicht werden (lediglich drei kleinere, ausschließlich am VD18 beteiligte Bibliotheken fehlen in dieser Liste). Als Frist wurde allen Bibliotheken der 15.1.2020 gesetzt. Am 7.1.2020 wurde eine Erinnerungsmail verschickt, am 16.1.2020 wurde die Umfrage für die weitere Teilnahme gesperrt.

1.3. RÜCKLAUF UND REPRÄSENTATIVITÄT DER UMFRAGE

An der TBU haben alle 4 Trägerbibliotheken, an der WBU 14 der 41 angeschriebenen Bibliotheken (31,7 %) teilgenommen. Diese relativ niedrige Rücklaufquote der WBU war angesichts der Tatsache, dass einige der im VD17-Verteiler enthaltenen Bibliotheken gar keine oder nur sehr geringe Digitalisierungsarbeiten (in den VD) leisten bzw. geleistet haben, zu erwarten.[1] An der Umfrage teilgenommen haben v.a. die im Bereich der Digitalisierung sehr stark bis mittel engagierten Bibliotheken. Bis auf eine Ausnahme sind die im Digitalisierungsbereich führenden Bibliotheken alle in der Umfrage vertreten.

Die Umfrage kann daher v.a. die Erfahrungen, Einschätzungen und Planungen der VD-Bibliotheken darstellen, die in großem oder mittlerem Umfang an den Digitalisierungsarbeiten beteiligt sind bzw. waren. Anzunehmen ist, dass sich vornehmlich diese Bibliotheken auch in eine Volltexttransformation der VD einbringen würden.

2. AUSWERTUNG

2.1. ERFASSUNG UND DIGITALISIERUNG IN DEN VD

2.1.1. DERZEITIGER STAND DER ARBEITEN

[1] Drei Bibliotheken haben auch per Mail eine entsprechende Rückmeldung gegeben. Eine dieser Bibliotheken hat nach expliziter Einladung doch noch an der Umfrage teilgenommen.

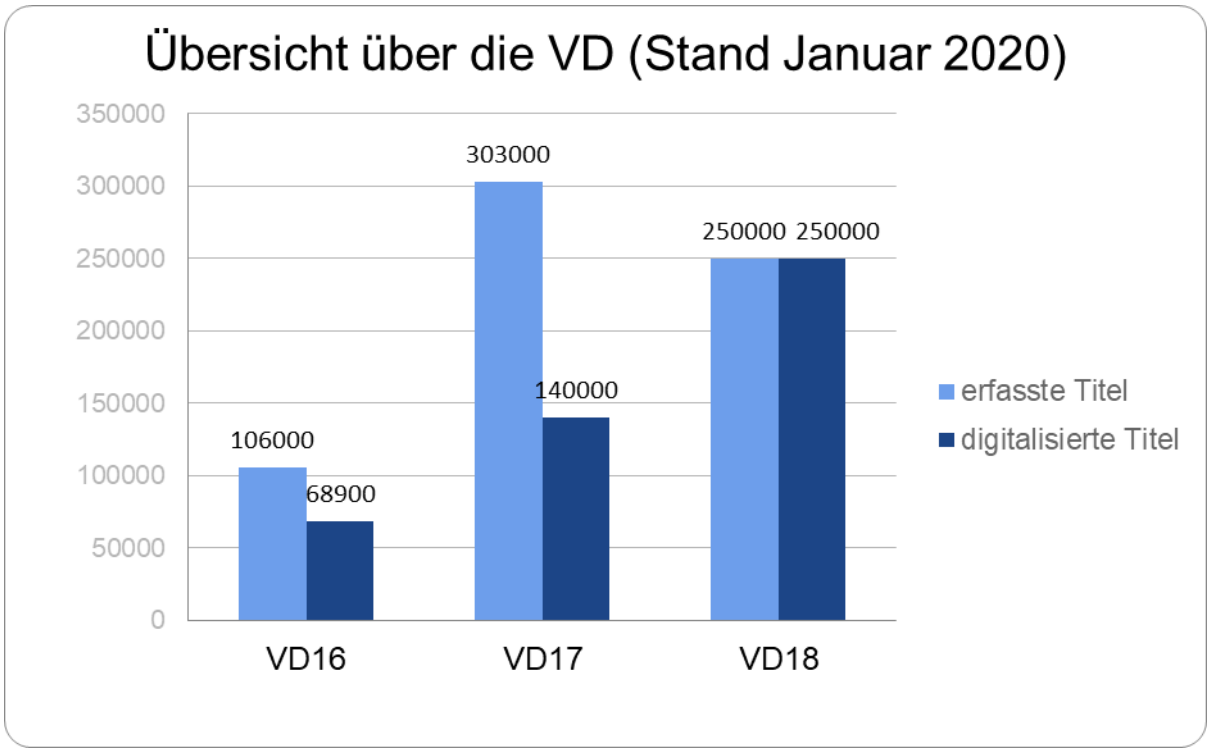


Diagramm 1 zeigt den derzeitigen Stand in der Erfassung und Digitalisierung der VD-Titel. Die Arbeiten am VD16 sind weitgehend abgeschlossen, nur vereinzelt werden noch neue Titel ergänzt. Im VD17 sind ebenfalls fast alle Titel eingetragen, davon liegt allerdings nur die Hälfte digital vor. Besonders die Bestände kleinerer Bibliotheken wurden noch nicht digitalisiert - zum Teil auch noch nicht im VD17 nachgewiesen - und sollen in einer neuen Förderlinie bearbeitet werden. Im VD18 werden die erfassten Titel stets direkt digitalisiert, weshalb hier keine Nachdigitalisierungsarbeiten nötig sind. Gleichzeitig ist dieses Verzeichnis wegen des späteren Projektbeginns und der insgesamt deutlich größeren Anzahl an Titeln noch stark im Aufbau befindlich.[2]

Digitalisiert wurden die VD-Titel in der Regel nach den jeweils aktuellen DFG-Praxisrichtlinien zur Digitalisierung (300 dpi, tiff). Eine Ausnahme, die zahlreiche Titel betrifft, sind die Google-Digitalisate der BSB, die in deutlich schlechterer Qualität vorliegen. Diese sollen ab Ende des Jahres 2020 erneut prozessiert werden, sodass deren Qualität zunehmend verbessert werden kann.

2.1.2. SCHWIERIGKEITEN BEI DEN VD-ARBEITEN

Die von den Trägerbibliotheken der VD geschilderten Probleme lassen sich in fünf Bereiche einteilen: Organisation/Abstimmung, Förderungsmöglichkeiten, Bestände in kleinen Bibliotheken, Zustand der zu digitalisierenden Titel und bibliothekarische Probleme.

1. Organisation/Abstimmung

[2] Vorsichtige Schätzungen vor Beginn der VD18-Arbeiten gingen von mindestens 600.000 zu erfassenden Titeln aus. Vgl. z.B. Siegert, Reinhart: VD18 – zum Diskussionsstand aus fachwissenschaftlicher Sicht. In: Bibliothek – Forschung und Praxis 32 (2008) H. 2, S. 203–208; Schnelling, Heiner: Die Verzeichnisse der im deutschen Sprachraum erschienenen Drucke des 16., 17. und 18. Jahrhunderts. Kontinuität und Innovation. In: Fabian, Claudia / Bubenik, Claudia (Hrsg.): Schmelze des barocken Eisbergs? Das VD17 – Bilanz und Ausblick (Bibliothek und Wissenschaft 43). Wiesbaden 2010, S. 199–211.

Die Zuteilung der zu erschließenden und zu digitalisierenden Titel unter Vermeidung von Mehrfacharbeiten ist in allen VD eine große Herausforderung. Besonders problematisch sind in diesem Zusammenhang (exemplarspezifische) Sammelbände, bei denen eine Bibliothek nur einen Teil der darin enthaltenen Titel im Rahmen der VD-Projekte bearbeitet. Beim VD18 verschärft sich die Organisationsproblematik weiter durch die Synchronisierung von Erschließung und Digitalisierung.

2. Förderungsmöglichkeiten

Sind VD-relevante Titel noch nicht katalogisiert, müssen diese aufwändig neu erfasst werden. Diese Zusatzarbeiten werden von der DFG-Förderung nicht abgedeckt und müssen in Eigenleistung erbracht werden. Arbeiten in ausländischen Einrichtungen mit VD-relevanten Beständen, wie bspw. Bibliotheken in den ehemaligen deutschen Ostgebieten, können ebenfalls nicht gefördert werden. Dies ist besonders im Fall von Unika problematisch.

3. Bestände in kleinen Bibliotheken

Kleine Einrichtungen haben zum einen teils nicht die nötige personelle und technische Ausstattung, um ihre Bestände selbst für die VD erfassen und digitalisieren zu können. Zum anderen ist die Anzahl der in einer solchen Bibliothek vorhandenen, noch für die VD benötigten Titel oft nur gering bzw. zu gering für einen DFG-Antrag.

4. Zustand der zu digitalisierenden Titel

Die VD-Titel sind bereits mehrere Jahrhunderte alt, weshalb sich die erhaltenen Exemplare in teils schlechtem konservatorischen Zustand befinden und für die Digitalisierung nicht geeignet sind. Zurückgewiesene Digitalisierungsaufträge müssen entsprechend neu vergeben werden, was den Organisationsaufwand weiter erhöht.

5. Bibliothekarische Probleme

Die Titel sind in den VD sehr tiefgehend zu erfassen, sodass diese Erschließung teils deutlich über die regulären bibliothekarischen Erschließungsarbeiten hinausgeht und vorhandene Altkatalogisate dementsprechend nachgearbeitet werden müssen. Für das VD16 wurde zudem die Bibliothekssoftware Aleph eingesetzt, bei der es wegen fehlender Redaktionszugänge zu Bearbeitungsproblemen kam.

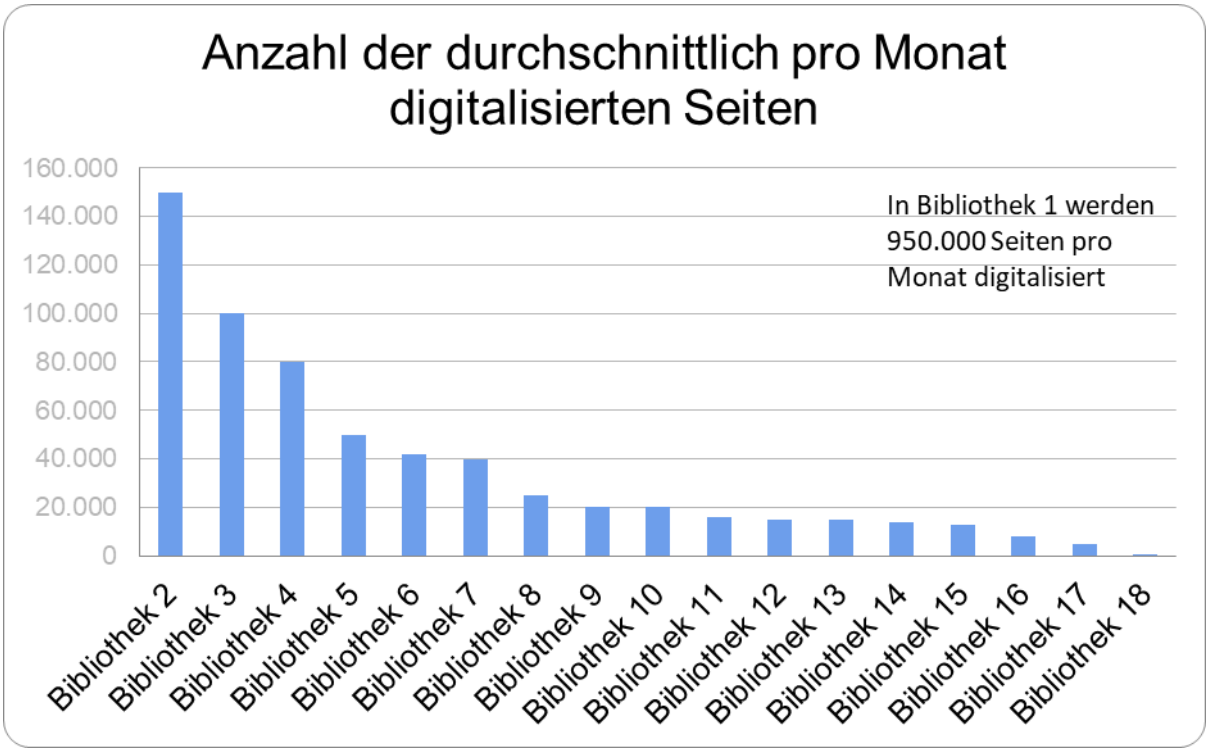
Die befragten weiteren VD-Bibliotheken sehen überwiegend keine Probleme in den VD-Projekten, lediglich 36 % der Teilnehmer, überwiegend große Digitalisierungseinrichtungen, weisen auf Schwierigkeiten hin. An organisatorischen Problemen wird im VD18 die Zuweisung der zu bearbeitenden Titel auf Grundlage eines zehn Jahre alten Katalogabzugs genannt, der inzwischen in Teilen veraltet ist. Auf Seiten der Bibliotheken ist darüber hinaus die Sichtung von Titeln für das VD18 sehr aufwändig, deren Digitalisierung aufgrund der teils unzulänglichen Set-Bildung, teils auch aus konservatorischen Gründen, nur bei einem Bruchteil tatsächlich realisiert wird.

Weitere Kritikpunkte betreffen Aufbau und Funktionen der VD-Portale. So ist es nicht möglich, die einzelnen VD nach Strukturmerkmalen zu durchsuchen. Ebenso werden die in das VD18 aufgenommenen Digitalisate von Zeitschriften-Titeln nicht nachvollziehbar zentral zusammengeführt, was deren Benutzung erschwert. Die Voraussetzungen für die Einbindung der im Bibliotheksbereich als zukunftsweisend angesehenen Volltexterkennung sowie des IIIF-Formats sind wegen fehlender Normierungen (IIIF-Viewer) sowie Ground Truth-Daten (für OCR) derzeit noch nicht geschaffen.

Zu Beginn des VD18-Projekts war auch der zu gering angesetzte Seitenpreis problematisch. Diese für die beteiligten Bibliotheken schwierige Finanzierungsgrundlage hat sich durch die im Jahr 2017 erfolgte Aufhebung des festen Seitenpreises deutlich verbessert.[3]

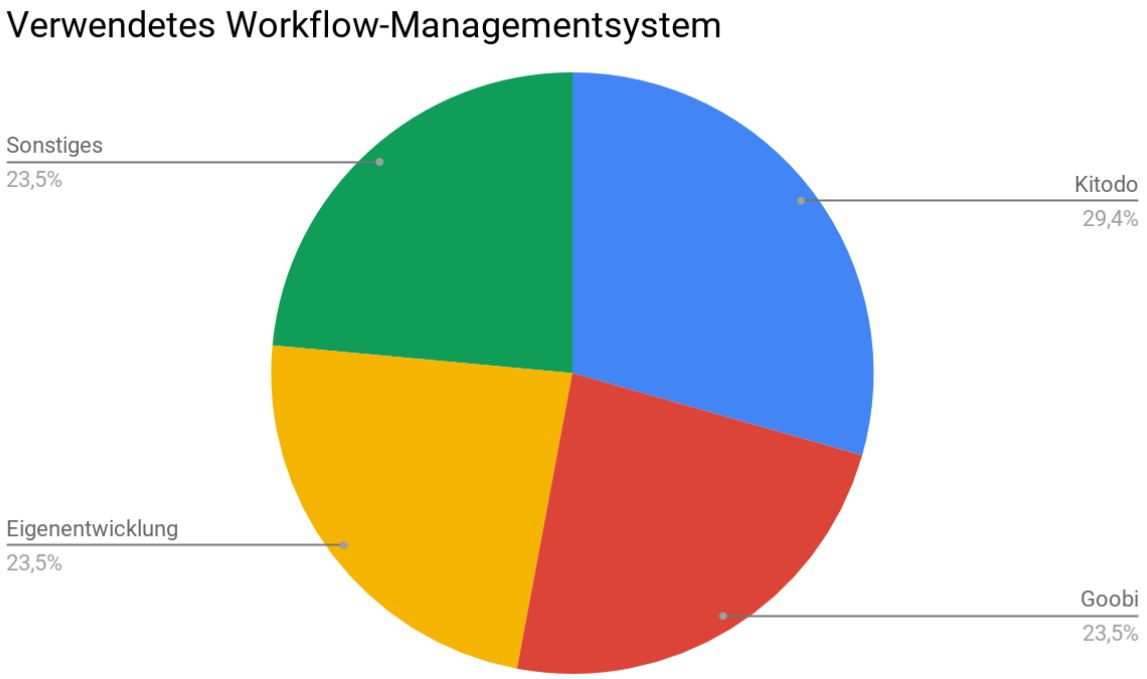
2.2. (BILD-)DIGITALISIERUNG IN DEN VD-BIBLIOTHEKEN

Die Anzahl der von den befragten Bibliotheken pro Monat digitalisierten Seiten weist eine sehr große Spannbreite auf (vgl. Diagramm 2).



Bis auf eine sehr kleine Bibliothek werden in allen befragten Einrichtungen mehrere tausend Seiten pro Monat von mehreren Stellen mithilfe einer Workflow-Software bilddigitalisiert. Überwiegend kommen dabei Kitodo bzw. Goobi zum Einsatz (vgl. Diagramm 3). Zwei Bibliotheken, die sich mit dem Umfang ihrer Digitalisierungsarbeiten am unteren Ende der Umfrageteilnehmer bewegen, verwenden neben Kitodo jeweils eine weitere Software.

[3] Vgl. Deutsche Forschungsgemeinschaft - Hinweise für Projekte zur Digitalisierung und Erschließung von im deutschen Sprachraum erschienenen Drucken



Ausgabeformat ist mit einer Ausnahme überall tiff, daneben werden meist noch weitere Formate wie jpg und pdf genutzt.

2.3. VOLLTEXTDIGITALISIERUNG IN DEN VD-BIBLIOTHEKEN

2.3.1. BISHERIGE ERFAHRUNGEN MIT OCR

In den befragten Bibliotheken werden zahlreiche Digitalisierungsprojekte unternommen bzw. derzeit geplant. Lediglich sechs der insgesamt 18 Umfrageteilnehmer erwähnen aktuelle Projekte mit Volltexterkennung. 61 % der Umfrageteilnehmer haben jedoch zumindest in zwei Projekten bereits erste Erfahrungen mit OCR gesammelt. Von diesen führt die Hälfte die Texterkennung selbst inhouse durch, während 18 %, v.a. die kleineren Einrichtungen, auf einen Dienstleister setzen. Die übrigen Bibliotheken haben bereits beide Varianten ausprobiert. Zum Einsatz kommt hauptsächlich der ABBYY Finereader, daneben auch Tesseract. Für die Erstellung der OCR durch einen Dienstleister sprechen aus Sicht der Umfrageteilnehmer der verringerte organisatorische Aufwand sowie fehlende OCR-Kenntnisse in den Einrichtungen. Durchgeführt wird die OCR hauptsächlich direkt im Anschluss an die Bilddigitalisierung, nachträglich werden nur wenige Titel für die Volltexterkennung ausgewählt.

Die mit OCR prozessierten Titel sind sehr heterogen und stammen seltener aus dem 16. oder 17., meist aus dem 18. bis 20. Jahrhundert. Fast alle befragten Bibliotheken wollen künftig OCR einsetzen bzw. ihre dahingehenden Arbeiten auf weitere oder auch alle Teile ihrer Sammlungen ausbauen. Bibliotheken, die bisher keine OCR genutzt haben, erwägen, diese v.a. für neu digitalisierte Titel inhouse sowie durch Dienstleister durchführen zu lassen.

2.3.2. OCR-VD-PROJEKTE

Aufgeschlossen gegenüber einem OCR-Projekt zu VD-Titeln an ihrer Bibliothek sind 82 % der Umfrageteilnehmer. Als Voraussetzung dafür wird insbesondere eine gute Erkennungsrate ohne bzw. nur mit automatisierter Nachkorrektur angegeben. Zudem müssten die Arbeiten gut in den bestehenden Digitalisierungsworkflow integrierbar sein und finanziell (von der DFG) gefördert werden. Der Umfang eines

derartigen OCR-Projekts wäre insbesondere von der Förderung abhängig, perspektivisch werden jedoch überwiegend sämtliche relevanten Titel als prozessierbar angegeben. Bevorzugt wird dabei die Prozessierung eigener Bestände bzw. Digitalisate, da bei diesen ein gewisser Qualitätsstandard sichergestellt ist. 46 % der an einem VD-OCR-Projekt interessierten Bibliotheken würden ggf. zusätzlich fremde Bestände prozessieren.

An eine OCR-Software werden hohe Anforderungen gestellt. Insbesondere sollte diese eine gute Erkennungsrate liefern, in der Bedienung (auch durch die Anbindung an bestehende Workflows) einfach und sowohl kosten- als auch zeiteffizient sein. Wünschenswert ist darüber hinaus eine integrierte Layout- und Strukturerkennung sowie die Möglichkeit zum weiteren Training der zur Verfügung gestellten Modelle. Einzelne Bibliotheken sehen zudem eine aktive Nutzer-Community, die gesicherte Weiterentwicklung der Software, offenen Quellcode sowie Möglichkeiten zur Inbetriebnahme auf verschiedenen Plattformen wie Linux und FreeBSD als wichtig an.

Als Herausforderung bei einer Volltexttransformation der VD werden neben der technischen Schwierigkeit der Volltexterkennung bei einer großen Vielfalt an Schriften sowie Schriftmischung insbesondere Organisation, Präsentation, Finanzierung und Personal gesehen.

1. Organisation

Die große Anzahl der zu prozessierenden Titel macht eine umfassende Koordinierung sowie weitgehend automatisierte Prozesse zu einer unabdingbaren Voraussetzung. Dies beginnt bereits bei der Zuweisung der Titel an eine Bibliothek. Zuteilungen müssen zentral erfolgen und vermerkt werden, um Mehrfacharbeiten zu vermeiden. Wird die OCR direkt nach der Bilddigitalisierung durchgeführt, sollte diese in die genutzte Workflow-Software integriert sein und automatisiert ablaufen. Müssen Titel aufgrund ihrer schlechten Erkennungsrate bzw. verbesserten technischen Möglichkeiten erneut prozessiert werden, wird dadurch zum einen der Workflow deutlich verkompliziert, zum anderen müssen Fragen der Aktualisierung, Referenzierbarkeit und Langzeitverfügbarkeit der verschiedenen Versionen geklärt werden. Sind mehrere Exemplare eines Titels im VD eingetragen, muss zudem die korrekte Zuordnung des Volltextes zum zu Grunde liegenden Exemplar bzw. Digitalisat sichergestellt werden. Sofern eine Bibliothek kein Exemplar des ihr zugewiesenen Titels besitzt, muss diese im Vorfeld der Prozessierung von der besitzenden Einrichtung Zugriff auf die zugehörigen Master-tiff erhalten.

2. Präsentation

Für Anzeige und Einbindung der Volltexte in die VD sind vorab einheitliche Ausgabeformate zu definieren. Zudem ist zu klären, in welcher Form die Volltexte präsentiert bzw. dem Nutzer zur Verfügung gestellt werden. Außerdem sollten die Volltexte in die Suchoptionen der VD eingebunden werden.

3. Finanzierung und Personal

Bei der Finanzierung der Volltexttransformation werden (zu hohe) Eigenanteile als abschreckend genannt. Zudem ist unklar, ob Titel aus bereits abgeschlossenen VD-Digitalisierungsprojekten, bei denen OCR kein Teil des ursprünglichen Antrags war, von der Bibliothek in Eigenleistung prozessiert werden müssten. Teilweise müssten auch die OCR-Kenntnisse im Haus erst noch durch Schulungen oder neues Personal verbessert werden. In diesem Zusammenhang wird die Ausschreibung verschiedener organisatorischer bzw. technischer Modelle als wünschenswert angesehen, die durch die Nutzung unterschiedlich anspruchsvoller Software oder ggf. die Vergabe an Dienstleister einem größeren Kreis an Bibliotheken die Teilnahme an OCR-Projekten erlauben.

Insgesamt wäre es wichtig, technische und organisatorische Standards für den gesamten OCR-Prozess einzuführen, um die große Datenmenge bewältigen zu können. Dies schließt auch die Überarbeitung der DFG-Praxisregeln zur Digitalisierung mit ein, in denen insbesondere die Evaluation der Erkennungsqualität noch nicht zufriedenstellend gelöst ist. Zwischen den teilnehmenden Einrichtungen sollte ein enger Austausch

bestehen, um sich gegenseitig mit Erfahrungen und Problemlösungen zu unterstützen. Sofern im Zuge der OCR-Arbeiten weitere Modelle trainiert werden, sollten auch diese gegenseitig zugänglich gemacht werden.

3. FAZIT

Die Umfrage hat einige Voraussetzungen sowie Herausforderungen bei der Volltexttransformation der VD aufgezeigt, die von OCR-D, aber auch den VD im Vorfeld zu adressieren sind.

3.1. AUFGABEN FÜR OCR-D

Die benannten Anforderungen an eine OCR-Software kann der OCR-D-Prototyp durch sein Entwicklungsstadium bedingt bisher nur teilweise erfüllen. Sie stimmen jedoch mit den Zielen von OCR-D überein und sollen im Zuge der Optimierung und Implementierung des Prototypen in einer breiten Auswahl an Einrichtungen in der 3. Projektphase ab dem Jahr 2021 alle umgesetzt werden.

Neben der Optimierung des Prototypen sollte dabei von OCR-D eine Integration in Kitodo/Goobi geleistet werden, um dem Großteil der Bibliotheken die Einbindung der OCR in die bestehenden Workflows zu erleichtern. Mit Kitodo wurde bereits ein Letter of Intent unterzeichnet, der diesen Punkt perspektivisch beinhaltet. Zudem werden die praktischen Kenntnisse über die OCR-D-Software sowie die Erkenntnisse des OCR-D-Projekts weiter verbreitet bzw. an die interessierten Bibliotheken kommuniziert. Die Erfahrungen des Projekts werden als Vorschläge für die Überarbeitung der DFG-Praxisregeln eingebracht und wären dann von künftigen Digitalisierungsprojekten zu berücksichtigen.

3.2. AUFGABEN FÜR DIE VD

Die VD bzw. VD-Bibliotheken müssten dagegen Fragen zur Integration der Volltexte klären. Dies betrifft die Bereitstellung und Anzeige der Volltexte im viewer neben dem Digitalisat und zum Download bei der besitzenden Bibliothek. In den VD als zentrale Nachweissysteme müssen Felder für die Links zu den dezentral gespeicherten Volltexten sowie den zugehörigen Metadaten (bspw. zur Provenienz, Version etc.) bereitgestellt werden. Außerdem ist zu klären, wie die Volltexte über Indexierung in die Suchmöglichkeiten der VD sowie ggf. der lokalen Bibliothekskataloge eingebunden werden sollen.

Für die Organisation der Volltexttransformation sind die vorhandenen zentralen VD-Verzeichnisse so anzupassen, dass Volltexte zu den einzelnen Titeln als vorgemerkt, in Bearbeitung und bereits vorhanden eingetragen werden können. Alle bis zum Beginn der Volltexttransformation bereits in anderen (Projekt-)Kontexten erstellten Volltexte sind in diesen Verzeichnissen zu erfassen, um Mehrfacharbeiten zu vermeiden.

3.3. MÖGLICHE UMSETZUNG DER VOLLTEXTTRANSFORMATION

Wenn die oben angeführten Aufgaben gelöst sind und insbesondere eine zufriedenstellende Erkennungsrate bei einer großen Bandbreite an VD-Titeln erreicht wird, kann mit der Volltexttransformation der VD begonnen werden. Viele der bisherigen Probleme im Aufbau der VD beziehen sich lediglich auf den Umgang mit physischen Exemplaren oder die Erfassung der Metadaten und sind für die Volltexterkennung nicht relevant. Dafür ist mit einer Verschärfung der organisatorischen Herausforderungen zu rechnen.

3.3.1. PROZESSIERUNGSSZENARIEN

Es ist davon auszugehen, dass sich die Anzahl der Einrichtungen, die sich an der Umsetzung der Volltexttransformation beteiligen würde, im niedrigen zweistelligen Bereich (ca. 10-20) bewegt. Grundsätzlich sind drei Szenarien zu unterscheiden: 1) Die Volltexterkennung noch zu erstellender Bilddigitalisate 2) Die nachträgliche Prozessierung der vorhandenen Bilddigitalisate. Diese beiden Szenarien sind weiterhin danach

zu differenzieren, ob die Volltexte von der bilddigitalisierenden Einrichtung selbst in-house oder extern angefertigt werden. 3) Die gebündelte Prozessierung von typografisch ähnlichen Titeln.

Szenario 1 (Synchrone Neudigitalisierung)

Begonnen werden könnte mit dem ersten Szenario, indem die OCR-Erstellung in die neuen VD-Digitalisierungsprojekte aufgenommen wird. Die Auswahl der zu prozessierenden Titel ergibt sich aus den aktuellen (Bild-)Digitalisierungsplänen der VD. Sofern die Volltexterkennung direkt in den Digitalisierungsworkflow eingebunden werden kann, dürften so Mehraufwände in der Organisation etc., die bei einer getrennten Bild- und Textdigitalisierung entstehen, vermieden werden. Dass Volltexterkennung vermutlich nicht von allen Bibliotheken, die sich an der Digitalisierung neuer VD-Titel beteiligen, geleistet werden kann, ist in den neuen Ausschreibungen entsprechend zu berücksichtigen. Möglich wären bspw. die Auslagerung der OCR an Dienstleister, die Einrichtung eines OCR-D-Service oder Kooperationen mit Einrichtungen, die die OCR durchführen können. Muss für die OCR-Erstellung mit einer externen Einrichtung zusammengearbeitet werden, ist die beabsichtigte Vorgehensweise im Antrag darzulegen. Ein Ausschluss von Einrichtungen ohne eigene Möglichkeiten zur OCR-Erstellung von den (Bild-)Digitalisierungsarbeiten der VD ist zu vermeiden.

Szenario 2 (Nachträgliche Volltexterstellung)

In einem zweiten Schritt können, unter Berücksichtigung der Erfahrungen mit den Neudigitalisierungen, die Titel im zweiten Szenario prozessiert werden. Da deren Bilddigitalisierungsprojekte bereits abgeschlossen sind, muss deren Volltexterkennung sowohl neu beantragt, als auch neu organisiert werden. Mit Blick auf den Wunsch der Bibliotheken, vorwiegend selbst erstellte Bilddigitalisate zu prozessieren, sowie den zu erwartenden Mehraufwand bei der Prozessierung "fremder" Digitalisate könnten die ursprünglichen Zuteilungen der VD zur Bilddigitalisierung der VD-Titel als Ausgangspunkt für die Verteilung der Titel zur Volltexterkennung genutzt werden. Bei Mehrfachdigitalisaten, wie sie v.a. im VD16 vorliegen, ist die Zuteilung weniger eindeutig. Dieses Problem dürfte bereits dadurch entschärft werden, dass nicht alle bilddigitalisierenden Einrichtungen an der Volltexterstellung teilnehmen werden. Haben mehrere Einrichtungen Interesse an der Prozessierung einzelner Titel, können als Entscheidungsgrundlage zum einen die Anzahl der von diesen Einrichtungen potentiell zu prozessierenden Titel (mit Blick auf adäquate Antragsvolumina), zum anderen die Bündelung (in Layout etc.) ähnlicher Drucke anhand der Gattungs-Normdaten dienen.

Szenario 3 (Typografische Bündelung)

Bei einzelnen Textgattungen wie bspw. Leichenpredigten ist aufgrund deren typografischer Gestaltung zu erwarten, dass annehmbare Erkennungsraten nur mit größerem Aufwand in der Workflowerstellung oder dem Training neuer Modelle erreicht werden können. Durch die gebündelte Prozessierung ähnlich gestalteter Titel wird verhindert, dass mehrere Einrichtungen parallel aufwändige Lösungen für diese anspruchsvollen Titel erarbeiten müssen. Für die Ermittlung der zu einer Gattung gehörigen Titel können die VD-Normdatensätze genutzt werden. Dieses Szenario richtet sich insbesondere an Einrichtungen, die bereits über große Kompetenz in der Volltextdigitalisierung verfügen. Die erwarteten Mehraufwände sind in der Finanzierung entsprechend zu berücksichtigen, die erarbeiteten OCR-Lösungen (Workflows und Modelle) sind in einem geeigneten Repositorium auch anderen Einrichtungen zugänglich zu machen. Als zentrale Plattform kann bspw. Zenodo genutzt werden.

3.3.2. VORGEHENSWEISE

Die Vorgehensweise kann sich am VD17 Masterplan orientieren. Bei Beantragung eines VD-Digitalisierungsprojekts bei der DFG werden die Titel, die (bild- und) volltextdigitalisiert werden sollen, in der Datenbank des jeweiligen VD als vorgemerkt markiert. Können die Titel doch nicht prozessiert werden, sind

diese wieder für die Bearbeitung durch andere Einrichtungen freizugeben. Die erstellten Digitalisate werden ebenfalls zentral verzeichnet und deren Bearbeitung als erledigt markiert.

Falls kein zentraler OCR-D-Service für die Prozessierung der bereits vorhandenen Bilddigitalisate von Bibliotheken, die sich nicht an der Volltexttransformation beteiligen können oder wollen, angeboten werden kann, sind zwei alternative Lösungen denkbar. Zum einen könnten diese zunächst unberücksichtigt bleiben und nach der Prozessierung der übrigen, von den teilnehmenden Bibliotheken bevorzugten Digitalisate in einer weiteren Ausschreibungsphase bearbeitet werden. Zum anderen könnte als Voraussetzung für Anträge zur Volltexterkennung festgelegt werden, dass ein kleiner Prozentsatz der bearbeiteten Titel aus "fremden" Digitalisaten bestehen muss. Hierfür wären zunächst alle Digitalisate zu identifizieren, die von den ursprünglichen Erstellern voraussichtlich nicht prozessiert werden können. Aus diesem Kreis können den Antragstellern dann "fremde" Digitalisate zugewiesen werden. Um den organisatorischen Aufwand möglichst gering zu halten, sollten die Digitalisate der einzelnen Bibliotheken jeweils gebündelt zugewiesen werden. In beiden Fällen ist den Bibliotheken deutlich zu kommunizieren, dass die Bilddigitalisate in den VD, mit Ausnahme der Google-Digitalisate der BSB, alle nach den Praxisregeln der DFG angefertigt wurden und daher eine ähnlich hohe Qualität aufweisen (sollten). Der Aufwand in der Bearbeitung "fremder" Digitalisate dürfte wegen des nötigen Austauschs (zumindest der Master-tiff und der Volltexte) mit den ursprünglichen Erstellern höher sein als bei der Prozessierung eigener Digitalisate. Dies könnte mehr für den zweiten Lösungsvorschlag sprechen, da der Mehraufwand so über einen längeren Zeitraum und mehr Projekte verteilt wird. Im Fall des ersten Lösungsvorschlags müsste der Mehraufwand in den Finanzierungsmodellen entsprechend berücksichtigt werden.

3.3.3. KOSTENABSCHÄTZUNG UND ZEITPLAN

Die Kosten der Prozessierung sind erst im Lauf der 3. OCR-D-Phase zur Implementierung der Software in bestandshaltendenen und -verarbeitenden Einrichtungen ermittelbar. Die Gesamtkosten sind zudem abhängig von der Anzahl der in den verschiedenen Szenarien zu prozessierenden Digitalisate, die erst dann bestimmt bzw. aktualisiert werden sollte, wenn der Beginn der Volltexttransformation absehbar ist. Daraus ist auch ein Zeitplan für die Prozessierung zu erstellen. Da die zahlreichen Google-Digitalisate der BSB bereits im Rahmen von GoogleBooks volltextdigitalisiert und der Nutzung frei zugänglich gemacht wurden, sind diese zumindest vorerst aus den OCR-Arbeiten in den VD auszuschließen. Nach der Prozessierung (eines Großteils) der übrigen VD-Titel kann geprüft werden, inwieweit eine erneute Volltextdigitalisierung der Google-Digitalisate durch eine signifikante Verbesserung der Erkennungsrate gerechtfertigt ist.