OCR-D

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

KONZEPT ZUR VERSTETIGUNG DER OCR-D-ERGEBNISSE

02.10.2023

Koordinierungsprojekt

Inhalt

1 Abstract 1

2 Ausgangslage 1

3 Ziele der Verstetigung 2

4 Anforderung für die Verstetigung 2

4.1 Definition der Software-Produktlevel 2
4.2 Robustheit der Software 3
4.3 Full-Support der Software 4

5 Betriebsmodelle und Ressourcenbedarf: Konsequenzen für die Verstetigung 6

5.1 Betriebsmodelle 6
5.2 Ressourcenbedarf 8

6 Organisationsmodelle 9

6.1 HathiTrust Digital Library 9
6.2 READ-COOP SCE 9
6.3 IMPACT-Kompetenzzentrum für Digitalisierung 10
6.4 Kitodo. Key to digital objects e. V. 10

7 Fazit 11

1 ABSTRACT

Leitendes Ziel der OCR-D-Initiative ist es, die in den Verzeichnissen der im deutschen Sprachbereich erschienenen Drucke des 16. bis 18. Jahrhunderts (VD 16, VD 17, VD 18, insgesamt: VD) erschlossenen historischen Drucke mithilfe der OCR-D-Software mit Volltexten auszustatten. Ziel der Verstetigung ist 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, um damit eine langfristige, zuverlässige Nutzung der OCR-D-Software und qualitativ zufriedenstellende OCR-Ergebnisse zu ermöglichen. Damit die OCR-D-Software als Produkt im Zuge der Verstetigung ausreichend unterstützt wird, benötigt sie regelmäßige Entwicklung, Updates und Wartung sowie die Bereitstellung dafür notwendiger Ressourcen. Der Ressourcenbedarf und die Kosten der Volltextdigitalisierung pro Seite sind davon abhängig, ob die OCR-D-Software zentral bereitgestellt und genutzt wird (Software-as-a-Service) oder in den (VD-)Bibliotheken lokal installiert und betrieben wird (On-Premise-Deployment). Empfohlen wird, dass beide Betriebsmodelle angeboten und seitens der DFG unterstützt werden. Je nach vorhandener Infrastruktur, finanziellen Mitteln, personeller Ausstattung und Erfahrung, sowie je nach speziellen Herausforderungen (einfachere oder komplexere Textstruktur, Layout, Typen- und Sprachenmischung), kann dann eine zentrale oder dezentrale Lösung genutzt werden. Angesichts der Unmöglichkeit, alle potenziellen Herausforderungen vor dem OCR-Prozess zweifelsfrei zu identifizieren, soll erwogen werden, eine umfassende Massenvolltextgenerierung des gesamten VD-Bestands – abzüglich bereits vorhandener Volltexte – durchzuführen und die Ergebnisse aktiv als integralen Bestandteil der Qualitätssicherung sowie zur gezielten Identifizierung von Problemfällen zu nutzen. Organisatorisch wird nach der Evaluation verschiedener Modelle und der Diskussion im VD-Rundgespräch eine enge Zusammenarbeit mit dem Kitodo-Verein empfohlen. Aktuell wird beidseitig geprüft, ob OCR-D in dem Kitodo-Verein aufgehen und die OCR-D-Software analog zu den bisher vorhandenen Kitodo-Software-Modulen über die Vereinsstrukturen gepflegt werden kann. Wenn mindestens ein Dienstleister einen zentralen OCR-Generierungsdienst einrichtet und OCR-D-Software-as-a-Service anbietet, kann dieser Dienstleister einen Teil des Ressourcenbedarfs für die Verstetigung wie z.B. Beratung und Support übernehmen. Der hier vorliegende Bericht stellt einen konsolidierten Zwischenstand dar, der im laufenden Koordinierungsprojekt der OCR-D-Förderphase III unter Einbeziehung der Ergebnisse aus den Modul- und Implementierungsprojekten noch aktualisiert und präzisiert wird.

2 AUSGANGSLAGE

OCR-D ist die koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR). Sie verfolgt das Ziel, die Volltexttransformation der in den Verzeichnissen der im deutschen Sprachbereich erschienenen Drucke des 16. bis 18. Jahrhunderts (VD 16, VD 17, VD 18, insgesamt: VD) nachgewiesenen Drucke technisch und konzeptionell vorzubereiten. Die einzelnen Prozessschritte bei der Durchführung der OCR sind in der OCR-D-Software als eigenständige Softwareprogramme (sogenannte Prozessoren) umgesetzt. Dadurch können die Workflows optimal für die jeweiligen Drucke bzw. Gruppen von Drucken zusammengestellt werden. Nachdem in den ersten beiden Projektphasen (2015–2020) das Funktionsmodell und die interoperable Software entwickelt wurden, wird in der aktuellen dritten Phase (2021–2024) die Software so weiterentwickelt, sodass sie produktiv im Massenverfahren eingesetzt werden kann. Vier Implementierungsprojekte entwickeln skalierende Lösungen für verschiedene Einsatzszenarien und drei Modulprojekte arbeiten an die

Weiterentwicklung und der Verbesserung einzelner Komponenten. Nach Abschluss dieser Projektphase sollen die OCR-D-Ergebnisse verstetigt werden.

3 ZIELE DER VERSTETIGUNG

Basis der zu realisierenden möglichst vollständigen Texterkennung der bereits digitalisierten oder noch zu digitalisierenden VD-Bestände sind die OCR-D-Software und die mit ihr verbundenen Spezifikationen, Richtlinien, Ground-Truth-Materialien und Kompetenzen. Die Verstetigungskativitäten stellen sicher, dass auch nach Projektende die Projektergebnisse robust, einsatzfähig und zugänglich bleiben. Ein weiteres Ziel der Verstetigung ist die Erhaltung und der bedarfsgerechte Aufbau von Kompetenzen nach Projektende, da dadurch die Nachhaltigkeit der erreichten Ergebnisse besser gewährleistet wird, Dienstleistungen, vor allem Unterstützungsleistungen beim Einsatz der Software, erbracht werden können und das Wissen sowie die Fähigkeiten des Teams bewahrt werden. Dadurch wird es auch möglich, auf die zukünftigen Entwicklungen zu reagieren und die potentiellen technologischen Verbesserungen aufzunehmen.

Die OCR-D-Software ist quelloffen (mit den Lizenzen Apache 2.0 bzw. GPLv3) und steht auf gängigen Plattformen (GitHub, DockerHub, PyPI) frei zur Verfügung. Die Zielgruppe sind grundlegend wissenschaftliche und öffentliche Bibliotheken, die beabsichtigen, ihre digitalen Bestände um eine Texterkennung anzureichern oder ihrem Digitalisierungsworkflow eine Texterkennung hinzufügen wollen. Im Fokus stehen Bibliotheken, die ihre historischen Drucke mit Hilfe der OCR-D-Software mit Volltexten ausstatten wollen, insbesondere die federführenden VD-Trägerbibliotheken sowie die VD-Partnerbibliotheken.

4 ANFORDERUNG FÜR DIE VERSTETIGUNG

4.1 Definition der Software-Produktlevel

Die OCR-D-Software ist über mehrere Projektphasen hinweg von verschiedenen Einrichtungen weiterentwickelt worden. In enger Abstimmung mit der OCR-D-Community wurden folgende Produktlevel definiert, die aufgrund ihrer Merkmale und des Nutzens, den sie dem Nutzer*innenkreis bieten, differenziert wurden. Zu beachten ist dabei, dass die Komponenten zum Teil ergänzenden Charakter haben, zum Teil aber auch ausschließenden (wenn zum Beispiel eine andere OCR-Engine eingesetzt werden soll). Aus der Perspektive der Verstetigung geht es hier besonders um die verbindliche Zuweisung von Verantwortlichkeiten, die hilft, den Einsatz einzelner Komponenten aus der Nutzendenperspektive verlässlich zu gestalten.

Produktlevel 1a enthält für die OCR-Generierung essentielle Komponenten, die zurzeit vom OCR-D-Koordinierungsprojektteam aktuell gehalten werden. Dazu zählen:

- spec: OCR-D-Standards und -Spezifikationen
- core: OCR-D-Python Implementierung
- gt-guidelines: OCR-D-Ground-Truth-Richtlinien
- ocrd-website / ocr-d.github.io: OCR-D-Website

Produktlevel 1b enthält für die OCR-Generierung essentielle Komponenten, die aktuell in Zusammenarbeit mit OCR-D-Community gepflegt werden. Dazu zählen:

- ocrd_all: OCR-D-Installation via venv oder Docker mit
  - ocrd_tesserocr: Tesseract-OCR-Engine mit OCR-D-Erweiterungen
  - ocrd_cis: OCRopus-OCR-Engine mit OCR-D-Erweiterungen
  - ocrd_fileformat: u.a. PAGE-zu-ALTO-XML-Umwandlung
  - olahd: OCR-D-Archivierung
  - eynollah: OCR-D-Integration Wrapper für Eynollah-Layoutanalyse
  - ocrd_segment: OCR-D-Segmentierung/ Layoutanalyse
  - dinglehopper: OCR-D-Integration für dinglehopper-OCR-Evaluation

Produktlevel 2 enthält Komponenten, die für die OCR-Generierung optional sind, auch wenn sie einen bedeutenden Mehrwert bieten. Diese werden aktuell vom OCR-D-Koordinationsprojektteam zusammen mit der OCR-D-Community nach besten Kräften aktuell gehalten. Dazu zählen:

- ocrd_all: OCR-D-Installation via venv oder Docker mit
  - ocrd_calamari: Integration der Calamari OCR-Engine
  - ocrd_kraken: Integration der Kraken OCR-Engine
  - ocrd_browse: Grafische Benutzeroberfläche zum Betrachten von OCR-D Workspaces
  - ocrd_typegroups_classifier: Schriftartenerkennung
  - ocrd_wrap: Integration für multiple Bildbearbeitungswerkzeuge
  - ocrd_docstruct: Erzeugung von METS-Strukturdaten aus OCR
  - ocrd_doxa: Adaptive Binarisierung mit DoxaPy
  - ocrd_olena: Diverse Binarisierungsalgorithmen via OLENA
  - ocrd_detectron2: Layouterkennung über die detectron2 Segmentierungsmodelle
  - ocrd_nmalign: Abgleich von Zeichenketten für Mapping heterogener OCR-Ergebnisse
  - ocrd_neat: Erzeugen von Transkriptionsvorlagen für GT-Erstellung
  - ocrd_page2tei/mets_mods2tei: Erzeugen von TEI aus OCR-D Workspaces
  - ocrd_pagetopdf: Erzeugen von PDF mit Volltext und Faksimile aus OCR-D Workspaces
  - ocrd_repair_inconsistencies: Bereinigen inkonsistenter OCR innerhalb PAGE-XML

4.2 Robustheit der Software

Das Ziel ist es, dass die OCR-D-Software langfristig, zuverlässig und zufriedenstellend eingesetzt werden kann. Sie erfüllt daher folgende Kriterien:

- Sie ist skalierbar, das heißt, sie ist in der Lage, mit großen Datenmengen umzugehen, denn die Performance der Software wird laufend optimiert und Parallelisierung wird ermöglicht.
- Sie ist über öffentlich zugängliche Repositorien verfügbar.
- Sie wird nach dem Prinzip „Continuos Integration and Delivery“ (CI/CD) entwickelt und betrieben, das heißt, es sind CI/CD-Pipelines implementiert, die die Code-Qualität sicherstellen.

Ein erfolgreicher Übergang von der Projektphase zur operativen Phase setzt die Sicherstellung der Robustheit der OCR-D-Software voraus. Kriterien hierfür sind:

- Modularität und Erweiterbarkeit (neue Funktionen und Module können leicht hinzugefügt werden ohne die Stabilität des bestehenden Systems zu gefährden)

- Zuverlässigkeit (auch unter unvorhergesehenen Bedingungen funktioniert die Software stabil)
- Effizienz (die Software nutzt Ressourcen optimal und arbeitet bei hoher Beanspruchung stabil)
- Benutzbarkeit (Benutzer*innen können intuitiv und fehlerfrei mit der Software interagieren)
- Wartbarkeit (ermöglicht schnelle und effektive Änderungen und Fehlerbehebungen)
- Lauffähigkeit (die Software kann einfach betrieben, verwaltet und überwacht werden)

In der noch laufenden dritten Phase von OCR-D wurden verschiedene Maßnahmen zur Erhöhung der Robustheit der OCR-D-Software umgesetzt. Darunter:

- Aufräumen und Bereinigung des Codes
- Fehlerbehebungen und Optimierungen
- Aktualisierungen und Updates der einzelnen Software-Komponenten und Module
- Testing und Monitoring
- Qualitätssicherung durch das Quiver-Dashboard und Entwicklung eines automatisierten Benchmarking aufgrund der von der Community zurückgemeldeten Metriken
- Überarbeitung der Dokumentationen sowie Konzeption und Umsetzung der Betriebs- und Hostinginfrastruktur

4.3 Full-Support der Software

Damit die OCR-D-Software als Produkt vollwertig eingesetzt werden kann, benötigt sie regelmäßige Entwicklung, Updates und Wartung. Damit wird sichergestellt, dass sie den sich ändernden Anforderungen der Nutzer*innen gerecht wird. Es werden Fehler behoben und Sicherheitslücken geschlossen. Dadurch wird eine möglichst optimale Leistung und Zuverlässigkeit aufrechterhalten und die Software wird vor technologischer Veralterung (Obsoleszenz) geschützt. Neben dem Code der Software selbst und der dazugehörigen Code-Dokumentationen müssen auch die OCR-D-Nutzer*innen-Dokumentationen laufend aktualisiert und gepflegt werden. Die Gesamtheit dieser Aktivitäten bezeichnen wir als Full-Support.

Während der Projektphase wurden Standards und Prozesse festgelegt, um sicherzustellen, dass die veröffentlichte Software von hoher Qualität ist. Dies beinhaltet die Verfeinerung der bestehenden Testverfahren, die Überprüfung von Code und die Fehlerbehebung vor der Veröffentlichung. Es ermöglicht den Entwickler*innen, Fehler zu identifizieren und zu beheben, bevor unter Einbeziehung der Rückmeldungen der OCR-D-Nutzendenschaf eine neue Version veröffentlicht wird, was die Stabilität und Zuverlässigkeit der Software erhöht. Eine Software-Versionierung mittels Semantic Versioning kennzeichnet klar die verschiedenen Iterationen der Software. Dies erleichtert die Identifizierung und den Vergleich von verschiedenen Versionen.

Um den Full-Support der OCR-D-Ergebnisse gewährleisten zu können, sind mehrere Rollen erforderlich:

Der/Die **Entwickler*in** kennt die Software auf Code-Ebene. Er/Sie sorgt dafür, dass die Software einwandfrei einsatzbereit ist, aktualisiert sie auf den neuesten Stand, verbessert bereits bestehende Funktionen und entwickelt neue bei Bedarf.

Der/Die **DevOps-Entwickler*in** betreut die Hostingumgebung, Continuous Integration, Deployment und Monitoring.

Der/Die **Data Scientist** evaluiert und beschafft Ground-Truth-Daten und verwendet diese, um Trainingsdatensätze für maschinelles Lernen oder Evaluationen vorzubereiten. Er/Sie verwendet diese Daten, um Modelle zu trainieren und sie so zu konfigurieren, dass sie bestimmte Aufgaben wie Klassifikation oder Clustering ausführen. Er/Sie verwendet die Ground-Truth-Daten, um die Leistung von Modellen zu evaluieren und zu optimieren und übernimmt die fachliche Beratung zu geeigneten Modellen und Workflows.

Der/Die **Mitarbeiter*in für den technischen Support** kennt die technische Anwendung der Software. Er/Sie unterstützt bei der Installation, Konfiguration, Fehlerbehebung, Aktualisierung sowie in der Benutzung und pflegt die dazugehörige Dokumentation. Da er/sie im direkten Kontakt mit den Anwender*innen steht, gibt er/sie Rückmeldungen (zum Beispiel Verbesserungsvorschläge oder Funktionswünsche) direkt ins Team.

Ein zentrales Release Management stellt sicher, dass die Softwareentwicklung koordiniert voranschreitet und die Veröffentlichung organisiert und regelmäßig erfolgt. Die OCR-D Software lebt von der Beteiligung der Community und das zentrale Release-Management erleichtert die Zusammenarbeit, indem es klare Prozesse für die Einreichung von Beiträgen, die Überprüfung von Code und die Integration neuer Funktionen festlegt. Ein/e **Release Manager** sammelt und organisiert Software--Änderungen, Erweiterungen und Fehler-behebungen, sowohl in der Kernsoftware als auch in einzelnen Prozessoren ein. Er/Sie arbeitet eng mit den Entwickler*innen und der Community zusammen, um den geeigneten Release-Inhalt und -Zeit-punkt zu bestimmen.

Der/die **Product Owner** analysiert und sammelt die Anforderungen zum Produkt, formuliert diese konkret und nachvollziehbar aus und priorisiert sie anhand der aktuellen Bedarfe und dem von ihm/ ihr aufgestellten Zeitplan. Dafür steht er/sie mit den Stakeholdern im ständigen Austausch. Zudem arbeitet er/sie eng mit den anderen Rollen im Team zusammen, um sicherzustellen, dass die Entwickler*innen an den aktuell relevanten Anforderungen arbeiten sowie, dass der technische und fachliche Support und der Release Manager über die aktuellen Änderungen informiert ist.

Die OCR-D-Software bekannt zu machen sowie über Funktionen und Einsatzbereiche zu informieren, ist Aufgabe der/die **Mitarbeiter*in für Öffentlichkeitsarbeit**. Er/Sie ist verantwortlich für den öffentlichen Auftritt, fördert und fordert die soziale Vernetzung, stellt in Artikeln und Vorträgen die Software und die Arbeit des Teams vor und übernimmt die Kommunikation innerhalb und außerhalb der Community. Er/ sie fördert damit den Aufbau einer aktiven und engagierten Community.

Neben diesen Rollen ist es essentiell, die bereits bestehende, breite Community mit ihren umfangreichen und wertvollen Beiträgen zu erhalten und weiterhin eng mit einzubeziehen. Sie trägt maßgeblich dazu bei, Komponenten der Software zu entwickeln. Da sie die OCR-D-Software aktiv verwendet, kennt sie sich gut mit den spezifischen Einsatzmöglichkeiten bei verschiedenen Materialien aus und kann daher wertvolle Rückmeldungen geben. Zudem umfasst sie Expertise, die genutzt werden kann, um anderen Nutzer*innen fachlich zur Seite zu stehen, unter anderem bei der Auswahl der Workflows.

Es wird eine agile Organisation des Teams empfohlen. Dabei werden agile Methoden verwendet und eine iterative und kollaborative Arbeitsweise gefördert. Die Mitarbeiter*innen bilden zusammen mit den Vertretern der Community ein Team resp. Teams, um gemeinsam an verschiedenen Aspekten des Produkts zu arbeiten. Wissen wird miteinander geteilt, und die Kommunikation untereinander erfolgt direkt.

5 BETRIEBSMODELLE UND RESSOURCENBEDARF: KONSEQUENZEN FÜR DIE VERSTETIGUNG

Grundsätzlich kann die OCR-D-Software zentral bereitgestellt und genutzt (Software-as-a-Service) oder in den Bibliotheken lokal installiert und betrieben (On-Premise-Deployment) werden. Ein Ergebnis des VD-Rundgesprächs ist, dass eine Kombination beider Betriebsmodelle benötigt wird. Das jeweilige Betriebsmodell und die Nutzungshäufigkeit der Betriebsmodelle haben Auswirkungen auf den Ressourcenbedarf und die Kosten der Volltextdigitalisierung pro Seite.

5.1 Betriebsmodelle

Die Software steht kostenfrei und quelloffen als Apache 2.0 bzw. GPLv3 lizenziert zur Verfügung, sie kann von den Anwender*innen oder potentiellen Anbieter*innen eigenständig und unabhängig heruntergeladen, installiert und genutzt werden. Eine kommerzielle Nutzung oder ein kommerzielles Dienstleistungsangebot ist ebenfalls möglich. Das wird auch langfristig und unabhängig vom Betriebsmodell so bleiben.

5.1.1 On-Premise-Deployment

Bei einem On-Premise-Deployment installieren sich die Anwender*innen die (neueste Version der) Software lokal auf eigenen Rechnern, um sie dort zu verwenden. Erfahrungsgemäß werden Support bei der Installation, individuelle fachliche Beratung zur Benutzung der Software oder weitergehende Schulungen benötigt werden. Mit diesem Ansatz kann die Nutzer*innengemeinschaft gestärkt werden und unterstützend mitwirken, da es einen Austausch zum Beispiel zu Usability-Aspekten und nutzungsbezogenen Bedarfen / Einsatzszenarien gibt. Die lokalen Installationen zielen auf eine dezentrale OCR-Generierung. Texte, Digitalisate und Strukturdaten werden lokal oder kooperativ zentral gespeichert (zum Beispiel über die Schnittstellen und die Oberflächen wie sie aus dem Modulprojekt OLA-HD[1] hervorgehen).

5.1.2 Software-as-a-Service

Ein anderer Ansatz besteht darin, einen zentralen Dienst bzw. einen Software-as-a-Service aufzubauen. Die OCR-D Software und erforderliche IT-Infrastruktur mit den entsprechenden Rechen- und Speicherkapazitäten wird bei einem Dienstleister betrieben und kann gegen Kostenerstattung als Dienstleistung genutzt werden. Auf diese Weise entstehen keine Aufwende für lokale Installation und Betrieb. Die Software kann so leichter aktuell gehalten werden. Bibliotheken und andere Nutzer*innen senden entweder ihre Daten direkt an den zentralen Service oder die Daten werden über maschinelle Schnittstellen geharvestet. Sie erhalten die OCR-Ergebnisse zur lokalen Speicherung zurück. Die OCR-Texte können darüber hinaus zusammen mit den Digitalisaten und Strukturdaten direkt zentral (bspw. OLA-HD) abgespeichert, indiziert, präsentiert und archiviert werden. Solch ein zentraler Dienst zur Durchführung der OCR kann nur von Einrichtungen angeboten werden, die die notwendige Infrastruktur bereitstellen können. Bei einer Durchführung in einer HPC-Umgebung (High-Performance-Computing) kann zudem die Volltextdigitalisierung durch Parallelisierung enorm beschleunigt werden, sowie es beim

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

Implementierungsprojekt OPERANDI erprobt wird. Durch die Parallelisierung der Workflows wurde bei den aktuellen Tests ein Durchsatz von ca. 50 Seiten pro Minute erreicht. Somit können bei einem Dauerbetrieb theoretisch 140.000.000 Seiten in 5,33 Jahren volltextdigitalisiert werden. Praktisch erfordert dies weitere Zeitaufwende für die Datenlogistik, Qualitätskontrolle, Abrechnung, Wartung der Technik usw. Im Vergleich zu rechnerisch 133 Jahren, die bei einem sequentiellen Ansatz in einem Prozess notwendig wären, stellt dies einen sehr deutlichen Zeitgewinn dar.

Für die OCR-Generierung würde eine Abrechnung vertraglich geregelt werden und basierend auf Seiten bzw. Seitenpaketen (z.B. pro 10.000 Seiten) erfolgen. Hierbei sind die Seitenpreise abhängig von der Gesamtmenge der Seiten. Je mehr Bibliotheken (mit einer größeren Zahl von Digitalisaten) beabsichtigen, die OCR über einen zentralen Dienst erzeugen zu lassen, desto günstiger kann die Durchführung angeboten werden. Die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) hat als Projektpartnerin eine Beispielberechnung auf Basis der aktuell vorliegenden Zahlen vorgenommen: Bei mindestens 140.000.000 Seiten als Gesamtmenge würde eine Seite 0,00796 € netto kosten (0,00948 € brutto). Bei mindestens 20.000.000 Seiten würde eine Seite 0,02960 € netto kosten (0,03522 € brutto). Um sicherzustellen, dass die Kosten für den Dienst gedeckt werden können, muss der Seitenpreis bei geringeren Prozessierungszahlen erhöht werden. In diesen Preisen sind alle anfallenden Kosten (Personal- und Sachkosten für den Betrieb) auf aktueller Grundlage berücksichtigt. Ein Ergebnis des VD-Rundgesprächs zur Zukunft der Produktion und Bereitstellung von Volltexten für die nationalbibliographischen Verzeichnisse VD 16, VD 17 und VD 18 im August 2023 in Göttingen war, im Laufe dieser Projektphase Umfragen durchzuführen, um die Präferenzen der Bibliotheken zu ermitteln. Es ist zudem essentiell, dass sich der Dienstleister auf eine funktionsfähige, aktuelle Software verlassen kann und bei Bedarf konkrete Ansprechpartner*innen fest stehen. Eine mögliche Förderung der DFG auf Antrag zur Verwendung eines zentralen Software-as-a-Service würde das Ziel Volltextdigitalisierung der VD-Bestände innerhalb von 3 bis 5 Jahren ermöglichen.

5.1.3 Kombi-Betriebsmodell

Die Betriebsmodelle (dezentral/lokal und zentral) schließen sich nicht gegenseitig aus. Sinnvoll erscheint, dass beide Modelle angeboten werden. Je nach vorhandener Infrastruktur, finanziellen Mitteln, personeller Ausstattung und Erfahrung, sowie Besonderheiten der Gruppe von Vorlagen (Schwierigkeitsgrad, Komplexität von Layout, Typen- und Sprachmischung etc.), kann eine zentrale oder dezentrale Lösung genutzt werden. Dies gilt insbesondere für spezielle Materialarten, die zusätzliche oder variierende Schritte in den OCR-Workflows benötigen, um eine höhere Erkennungsrate zu erreichen. Hier ist denkbar, dass einzelne Einrichtungen entsprechende optimierte Prozesse für eine Materialgruppe aufbauen und einrichtungsübergreifend prozessieren. In einer Voranalyse erscheint es nicht möglich, alle potentiellen Herausforderungen vor dem OCR-Prozess zu identifizieren und eine exakte, zeitliche und Prozessierungskosten berücksichtigende Prognose zu erstellen; aktuelle Tests im Rahmen von OCR-D zeigten dennoch, dass in einem Massenvolltextungsverfahren deutlich über 90% der vorhandenen Digitalisate mit guten Ergebnissen - ausschließlich bereits vorhandener Volltexte – prozessiert werden können, um unmittelbar in die Bereitstellung zu gehen. Für den Teil, in dem (zum Beispiel aufgrund der fehlenden vorgeschalteten Layoutanalyse oder gezielter Anlernschritte) inakzeptabel schlechte Ergebnisse erzielt werden, können diese 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. Durch dieses Vorgehen wird die zentrale Speicherung und Indizierung und die stufenweise Bereitstellung von Volltexten für wissenschaftliche Nutzung ermöglicht, wie sie im Kontext des VD-Suchraums bzw. von *VD*

Zukunft im VD-Rundgespräch diskutiert wurde. Gleichzeitig steht schnell ein großer Korpus mit relevantem Material für einen entsprechenden Index zur Verfügung.

Sowohl für den zentralen als auch dezentralen Ansatz werden die o.g. Rollen in der Betreuung der Software und entsprechende Personalmittel benötigt. Lediglich die Rolle des technischen Supports ist vorrangig für lokale Installationen (dezentraler Ansatz) notwendig. OCR-D empfiehlt für eine Förderung beide Betriebsmodelle zu berücksichtigen.

Die Umsetzungsplanung zur technischen und organisatorischen Durchführung einer koordinierten Volltexttransformation der VD und der dafür angestellten Vorüberlegungen wird im separat vorgelegten *Konzept zur Volltexttransformation der VD 2023* dargestellt.

5.2 Ressourcenbedarf

5.2.1 Personal

Für die oben genannten Rollen fallen Personalkosten an. Es bedarf dabei nicht für jede Rolle eines Vollzeitäquivalents, je nach Expertise können einige Rollen in einer Person zusammen gefasst werden, zum Beispiel Product Owner und Release Manager sowie technischer Support und Entwicklung. Zur fachlichen Beratung kann bei Bedarf an die Community verwiesen werden (Ausführung siehe Kap. *Ein Full-Support der Software*). Die Besetzung anteiliger Stellen mit mehr als einer Person stellt demgegenüber eine höher Ausfallsicherheit sicher und bietet Vertretungsmöglichkeiten sowie weitere Vorteile.

Insgesamt kann davon ausgegangen werden, dass sich nach der Ausreifung der Software im Betrieb und der zunehmenden Erfahrung und auch abhängig von der Einsatzbreite der Software und der sie einsetzenden Einrichtungen der Aufwand reduziert. Voraussichtlich werden innovative Weiterentwicklungsschritte und neue Komponenten projektbezogen entstehen und - nach Durchlaufen von Tests - integriert.

Für die Ausübung aller genannten Aufgaben ist nach heutigem Stand die Ausstattung mit zwei Vollzeitäquivalenten, voraussichtlich in der Eingruppierung E13, optimal. Eine geringere Ausstattung erzeugt Lücken und steigert das Risiko ausbleibender Akzeptanz der Lösung.

5.2.2 Ausstattung

Zusätzlich zur Personalausstattung fallen Büroarbeitsplätze an (Sachkosten wie Raumkosten, Geschäftskosten, Telekommunikationskosten und IT-Kosten) und Gemeinkosten (Kosten für Zentrale Services, Verwaltung usw.) an.

Für die fortlaufende Auslieferung der Entwicklung (Continuos Integration and Delivery) wird die bestehende, kostenlose Infrastruktur von GitHub verwendet. Dadurch entstehen für diesen Aufwand keine zusätzlichen Sachkosten.

6 ORGANISATIONSMODELLE

Es gibt verschiedene Organisationsmodelle, die im Rahmen des Koordinationsprojekts evaluiert wurden. Hierzu wurden Gespräche mit Organisationen aus vergleichbaren Themenbereichen geführt. Hervorzuheben ist hierbei die „HathiTrust Digital Library“, die Europäische Genossenschaft „READ-COOP SCE“, das „Kompetenzzentrum für Digitalisierung IMPACT“ sowie der Verein „Kitodo. Key to digital objects e. V.“. Im Ergebnis des DFG-Rundgesprächs wurde beschlossen, die Verhandlungen mit dem Kitodo-Verein zu intensivieren und die anderen Modelle und Kooperationen nur dann in Erwägung zu ziehen, wenn die Kooperation mit dem Kitodo e.V. nicht zustande kommen kann.

6.1 HathiTrust Digital Library

Die HathiTrust Digital Library (HathiTrust) ist ein gemeinsam verwalteter und kofinanzierter Zusammenschluss von Hochschul- und Forschungsbibliotheken. Die Hauptziele von HathiTrust umfassen die digitale Archivierung und den Schutz von gedruckten und digitalen Werke vor dem Verfall, die Bereitstellung von Online-Zugängen zu diesen Werken für Forscher*innen, Studenten*innen und Bibliotheksnutzer*innen, die Umwandlung gedruckter Werke mithilfe von OCR in durchsuchbare und maschinenlesbare Formate, die Förderung der Zusammenarbeit und Partnerschaften zwischen Bildungseinrichtungen zur gemeinsamen Finanzierung und zum Aufbau digitaler Sammlungen sowie die Gewährleistung eines barrierefreien Zugangs für Menschen mit Behinderungen zu den digitalisierten Ressourcen.

Die University of Michigan ist derzeit der administrative und rechtliche Träger von HathiTrust, und alle Mitarbeiter von HathiTrust sind Mitarbeiter der University of Michigan. Die Universität schließt Vereinbarungen und Verträge für die Aktivitäten von HathiTrust ab und stellt auch die Infrastruktur für die digitalen Inhalte ihrer Mitglieder bereit.

Die Mitglieder von HathiTrust beteiligen sich an den Kosten für den Betrieb der Dienstleistungen und Programme. Die Jahresbeiträge für die meisten Mitglieder liegen zwischen 6.600 und 13.000 USD, je nach Bibliotheksbudget (basierend auf Gesamtausgaben p. a.) und Begebenheiten der Kollektion (cost-per-volume). Die Gebühren für gemeinfreie und urheberrechtlich geschützte Werke werden unterschiedlich berechnet.

6.2 READ-COOP SCE

READ-COOP SCE ist eine europäische Genossenschaft mit beschränkter Haftung und wurde am 1. Juli 2019 gegründet, um Transkribus - eine Plattform zur Layoutanalyse und Texterkennung von Handschriften - zu betreiben und weiterzuentwickeln. Die Plattform entstand im Rahmen der EU-Projekte tranScriptorium und READ (Recognition and Enrichment of Archival Documents). Sie wurde von der Universität Innsbruck bzw. der Gruppe Digitalisierung und elektronische Archivierung entwickelt.

Während ein Verein grundsätzlich ausschließlich ideelle Zwecke verfolgt, ist eine Genossenschaft wirtschaftlich orientiert. Bei einer eingetragenen Genossenschaft arbeiten die Mitglieder gleichberechtigt zusammen, um gemeinsame wirtschaftliche und soziale Ziele zu erreichen. Für eine Gründung sind mindestens drei natürliche oder juristische Gründungsmitglieder notwendig. Die Genossenschaftsmitglieder stellen gemeinsam die finanziellen Ressourcen und die Arbeitskraft bereit. Das gemeinsame Kapital wird zum Zweck der Mitglieder eingesetzt.

Im Gegensatz zur eingetragenen Genossenschaft (eG) ist bei der europäischen Genossenschaft (SCE) ein Mindestkapital von 30.000 € sowie fünf (natürliche oder juristische) Gründungsmitglieder (bzw. mindestens zwei juristische Personen oder Personengesellschaften) Voraussetzung.

6.3 IMPACT-Kompetenzzentrum für Digitalisierung

Das IMPACT-Kompetenzzentrum für Digitalisierung ist eine gemeinnützige Organisation, die sich aus öffentlichen und privaten Einrichtungen zusammensetzt und sich zum Ziel gesetzt hat, die Digitalisierung historischer gedruckter Texte "besser, schneller und billiger"[2] zu machen. Es stellt Werkzeuge, Dienstleistungen und Möglichkeiten zur Verfügung, um den Stand der Technik in den Bereichen Dokumentenanalyse, Sprachtechnologie und Verarbeitung historischer Texte weiter voranzutreiben. IMPACT wird von einem Executive Board von Vertreter*innen der Institutionen (Premium-Mitgliedern) geleitet und hat seinen Sitz in den Räumlichkeiten der Fundación General de la Universidad de Alicante. Die Leitung des IMPACT-Zentrums liegt in den Händen einer Managerin, einer Direktorin, eines wissenschaftlichen und technologischen Direktors und der Vorsitzenden des Verwaltungsrats.

IMPACT hat zwei Mitgliedschaftsarten. Für Institutionen, die die IMPACT-Initiative implementieren und weiterentwickeln möchten, wird die Premium-Mitgliedschaft angeboten. Die Premium-Mitglieder werden Teil des IMPACT-Vorstands und erhalten Zugang zu allen Angeboten der IMPACT-Initiative. Der jährliche Beitrag für Premium-Mitglieder beträgt derzeit 6.000 € (für öffentliche Institutionen) oder 10.000 € (für private Institutionen) zzgl. Mehrwertsteuer. Die Standard-Mitgliedschaft richtet sich an Institutionen und Einzelpersonen, die Unterstützung bei ihren Digitalisierungsprogrammen suchen und aktiv in das Netzwerk eingebunden sein möchten. Als Standard-Mitglied gibt es die Möglichkeit, bewährte Verfahren zu teilen und auf den Inhalt zuzugreifen, den die IMPACT-Initiative anbietet. Der jährliche Beitrag für Standard-Mitglieder beträgt 600 € (für öffentliche Institutionen) oder 1.000 € (für private Institutionen) zzgl. Mehrwertsteuer. Alle Mitgliedschaftsarten haben vergünstigtem Zugang zu IMPACT-Veranstaltungen und Konferenzen.

6.4 Kitodo. Key to digital objects e. V.

„Kitodo. Key to digital objects“ e. V. fördert im Besonderen den Einsatz und die Weiterentwicklung der offenen und freien Digitalisierungssoftware Kitodo in allen Bereichen der Erstellung, Erschließung, Zugänglichmachung und Archivierung von Digitalisaten. Dies erfolgt im Rahmen eines gemeinsamen Wissenstransfers und einer langfristig angelegten Kooperation der Beteiligten. Ziel ist dabei auch, softwarebasierte Digitalisierungsplattformen besser an anwenderspezifische Anforderungen anzupassen und die Unabhängigkeit der Kultureinrichtungen gegenüber kommerziellen Softwareanbietern zu stärken.

Kitodo ist ein eingetragener, gemeinnütziger Verein. Der Hauptzweck ist nicht kommerziell. Einnahmen des Vereins dürfen nur für den gemeinnützigen Zweck des Vereins verwendet werden. Die Mitglieder haben die Möglichkeit, ihren Beitrag auf verschiedene Arten zu gestalten, um die Arbeit und Entwicklung von Kitodo e.V. zu unterstützen. Als persönliches Mitglied ist Jahresbeitrag von 200 € zu leisten. Korporative Mitglieder wählen zwischen einem Jahresbeitrag von 200 €, 500 € oder 1.000 €, je nach ihrem gewünschten Unterstützungsgrad. Darüber hinaus wird die Möglichkeit angeboten, zum Kitodo-Entwicklungsfonds beizutragen, um die kontinuierliche Verbesserung der Software und

[2] https://www.digitisation.eu/

Dienstleistungen sicherzustellen. Mitglieder können sich für eine jährliche Unterstützung entscheiden, indem sie einen Jahresbeitrag von 3.000 € (Juniorförderung), 5.000 € (Standardförderung) oder 10.000 € (Premiumförderung) wählen. Die finanzielle Unterstützung trägt dazu bei, dass der Verein seine Ziele erreichen und die Bemühungen zur Förderung von Kitodo fortsetzen kann. Zwischen Kitodo-Verein und OCR-D-Koordinierung besteht seit 2019 eine Absichtserklärung/Vereinbarung zur Kooperation bei der Volltexterstellung innerhalb von Digitalisierungsworkflows.

Im VD-Rundgespräch 2023 waren sich die Teilnehmenden einig, dass eine organisatorische Anbindung von OCR-D an Kitodo e.V. angestrebt werden sollte. Es wurde festgestellt, dass ein Großteil der Mitglieder von Kitodo auch zur OCR-D-Zielgruppe gehört. Es wird daher im Laufe der aktuellen dritten Projektphase geprüft, statt einen weiteren Verein für die gleiche Zielgruppe mit vergleichbaren Aufgaben zu gründen, ob OCR-D in den Kitodo-Verein aufgehen und OCR-D-Software analog zu Kitodo-Software über die Vereinsstrukturen angeboten und gepflegt werden kann. Über mögliche Folgeschritte, Qualitätsanforderungen an die OCR-D-Software als Voraussetzung für eine Aufnahme unter das Dach des Kitodo-Vereins sind konkrete Gespräche aufgenommen worden. Konkretisiert werden kann die Zusammenarbeit und Integration dann, wenn die Implementierungsprojekte ausentwickelt sind und mit einem deutlich höheren Prozessierungsdurchsatz valide Zahlen zur Qualität der Ergebnisse vorliegen.

7 FAZIT

Für die Verstetigung der OCR-D-Ergebnisse sind die folgenden Handlungsfelder ausschlaggebend:

1. Gewährleistung regelmäßiger Entwicklung, Updates und Wartung der OCR-D-Software und Bereitstellung der notwendigen Ressourcen für die Software-Pflege.
2. Berücksichtigung des Ressourcenbedarfs und der Kosten der Volltextdigitalisierung in Abhängigkeit von Betriebsmodell (Software-as-a-Service oder On-Premise-Deployment).
3. Volltexttransformation des VD-Bestands für die Nutzung der Ergebnisse für Qualitätssicherung und Identifizierung von Problemfällen und für den sukzessiven Aufbau des VD-Suchraums im Kontext von VD-Zukunft.
4. Planung einer engen Zusammenarbeit mit dem Kitodo-Verein und Integration von OCR-D-Software in das Portfolio des Kitodo-Vereins.
5. Sicherstellung der Bereitstellung und Nutzung der OCR-D-Ergebnisse (Software, Daten und Expertise) nach Projektabschluss für das Erreichen des Leitzieles.