



# OCR-D

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

**KONZEPT ZUR VERSTETIGUNG DER OCR-D-  
ERGEBNISSE –  
AKTUALISIERTE FASSUNG OKTOBER 2024**

14.10.2024

OCR-D-Koordinierungsprojekt

# Inhalt

1 Einleitung.....	3
2 Ausgangslage.....	3
3 Ziele der Verstetigung.....	4
4 Anforderungen für die Verstetigung.....	4
4.1 Definition der Software-Produktlevel.....	4
4.2 Robustheit der Software.....	6
4.3 Support der Software über die Projektphase hinaus.....	7
5 Betriebsmodelle.....	8
5.1 Grundsätzliches.....	8
5.2 On-Premise-Deployment.....	8
5.3 OCR-as-a-Service.....	9
5.4 Kombiniertes Betriebsmodell.....	11
6 Organisationsmodelle.....	12
6.1 Vorgehen.....	12
6.2 HathiTrust Digital Library.....	12
6.3 READ-COOP SCE.....	12
6.4 IMPACT-Kompetenzzentrum für Digitalisierung.....	13
6.5 Kitodo. Key to digital objects e. V. ....	13
6.6 Entscheidung.....	14
7 Verstetigung.....	15
7.1 Integration von OCR-D in Kitodo e. V.....	15
7.2 Releasemanagement von Kernmodulen der Software.....	16
7.3 Weiterentwicklung der Software (Entwicklungsfonds).....	16
7.4 Langzeitverfügbarkeit und Langzeitarchivierung.....	17
8 Executive Summary.....	17

## 1 EINLEITUNG

In befristeten Projekten erstellte Infrastrukturen, Workflows und Software implizieren von vornherein die Herausforderung, wie die Ergebnisse nach Projektende verstetigt und im Idealfall weiterentwickelt werden können. Im Fall von OCR-D, der Koordinierten Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR), das zur 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) Infrastrukturen, Software, Workflows, Richtlinien und Dokumentationen erstellt hat, wurde diese Aufgabe schon frühzeitig mitgedacht und in vorhergehenden Entwurfsstufen für entsprechende Konzepte dargelegt. Das vorliegende Konzept stellt dazu den aktuellen Stand am Ende der dritten Förderphase dar.

Zunächst werden die Ausgangslage und die Ziele der Verstetigung erläutert. Darauf folgt eine Darstellung der Anforderungen für die Verstetigung, die die Definition von Software-Produktleveln und Aussagen zur Robustheit der Software sowie die Planungen zum Support der Software über die Projektphase hinaus umfassen. Im daran anschließenden Kapitel werden die Betriebsmodelle lokales bzw. dezentrales On-Premise-Deployment, zentraler Ansatz OCR-as-a-Service und ein kombiniertes Betriebsmodell vorgestellt. Auf dieser Grundlage schließen sich Ausführungen zu verschiedenen, denkbaren Organisationsmodellen an - von HathiTrust, über READ-COOP SCE und IMPACT bis zum Kitodo e. V. Mit der Entscheidung, die Verstetigung prioritär auf dem Weg einer Integration von OCR-D-Ergebnissen in den Kitodo-Verein. zu realisieren, wobei die Aufgabenfelder Releasemanagement, Weiterentwicklung der Software sowie Langzeitverfügbarkeit und Langzeitarchivierung eigens betrachtet werden, schließt dieses Kapitel. Zuletzt fasst eine Executive Summary die wichtigsten Aussagen des Konzepts zusammen.

Das vorliegende Konzept zur Verstetigung der OCR-D-Software basiert auf dem 2023 eingereichten Konzept<sup>1</sup> und berücksichtigt die Rückfragen und Monita dazu aus dem Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme (AWBI) von 2023. Das Konzept zur Verstetigung der OCR-D-Software steht in engem Zusammenhang mit dem Konzept zur VD-Volltexttransformation, das parallel vorgelegt wird.

## 2 AUSGANGSLAGE

OCR-D ist die koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der OCR. Sie verfolgt das Ziel, die Volltexttransformation der VD 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, wurde in der aktuellen dritten Phase (2021–2024) die Software so weiterentwickelt, dass sie produktiv im Massenverfahren eingesetzt werden kann. Vier Implementierungsprojekte entwickelten

---

<sup>1</sup> Konzept zur Volltexttransformation der VD 2023 <https://ocr-d.de/konzepte/VD-Volltexttransformation%202023.pdf>, Anlage [https://ocr-d.de/konzepte/Anlage\\_VD-Volltexttransformation%202020.pdf](https://ocr-d.de/konzepte/Anlage_VD-Volltexttransformation%202020.pdf)

skalierende Lösungen für verschiedene Einsatzszenarien und drei Modulprojekte arbeiteten an der Weiterentwicklung und der Verbesserung einzelner Komponenten. Nach Abschluss dieser Projektphase sollen die OCR-D-Ergebnisse verstetigt werden.

### 3 ZIELE DER VERSTETIGUNG

Basis einer möglichst vollständigen Texterkennung der bereits bilddigitalisierten und der zukünftig noch zu digitalisierenden VD-Bestände sind die OCR-D-Software und die mit ihr verbundenen Spezifikationen, Richtlinien, Dokumentationen, Ground Truth (GT)-Materialien und die in diesem Zusammenhang aufgebauten Kompetenzen. Die Verstetigungsaktivitäten für die Software und für die genannten Strukturen stellen sicher, dass auch nach Projektende diese Projektergebnisse robust, einsatzfähig und zugänglich bleiben. Ein weiteres Ziel der Verstetigung ist die Erhaltung und der bedarfsgerechte Ausbau von Kompetenzen nach Projektende, da nur dadurch a) die Nachhaltigkeit der erreichten Ergebnisse gewährleistet wird, b) Dienstleistungen (vor allem Unterstützungsleistungen beim Einsatz der Software) erbracht werden und c) das Wissen sowie die Fähigkeiten der beteiligten Teams zur Unterstützung der Community bewahrt werden können. Zugleich wird es so möglich, auf zukünftige Entwicklungen zu reagieren und potentielle technologische Verbesserungen aufzunehmen.

Die primäre Zielgruppe für die Erstellung und Bereitstellung der Volltexte sind die Nutzenden der VD-Daten aus Forschung und Lehre sowie Nutzende außerhalb des engeren akademischen Spektrums. Die Zielgruppe für die Verstetigung der OCR-D-Software bilden wissenschaftliche und öffentliche Bibliotheken, die beabsichtigen, ihre digitalen Bestände um eine Texterkennung anzureichern oder ihrem Digitalisierungsworkflow eine Texterkennung hinzuzufügen. Auch für weitere wissenschaftliche und wissenschaftsnahe Einrichtungen aus dem Umfeld der Bibliotheken, wie Archive, Museen (GLAM) oder außeruniversitäre Forschungseinrichtungen wird eine verstetigte, offen zugängliche Volltext-Infrastruktur von Interesse sein. Im Fokus stehen Institutionen, derzeit vor allem Bibliotheken, die Bilddigitalisate ihrer historischen Drucke mittels OCR-D-Software mit Volltexten ergänzen wollen, insbesondere die federführenden und VD-Trägerbibliotheken sowie die VD-Partnerbibliotheken.

### 4 ANFORDERUNGEN FÜR DIE VERSTETIGUNG

#### 4.1 DEFINITION DER SOFTWARE-PRODUKTLEVEL

Die OCR-D-Software ist über mehrere Projektphasen hinweg von verschiedenen Einrichtungen mit Förderung der Deutschen Forschungsgemeinschaft (DFG) weiterentwickelt worden. In enger Abstimmung mit der OCR-D-Community wurden folgende Produktlevel definiert, die aufgrund ihrer Merkmale und des Nutzens, den sie dem Nutzendenkreis bieten, differenziert wurden. Zu beachten ist dabei, dass die Komponenten zum Teil ergänzenden, zum Teil aber auch ausschließenden Charakter haben (wenn zum Beispiel eine andere OCR-Engine eingesetzt werden soll). Im Hinblick auf die Verstetigung der OCR-D-Strukturen geht es hier besonders um die verbindliche Zuweisung von Verantwortlichkeiten, die dazu benötigt wird, den Einsatz einzelner Komponenten aus der Nutzendenperspektive verlässlich zu gestalten. Außerdem wird so erreicht, dass kein unbeherrschbarer

Software- und Technikumfang entsteht, sondern klar abgegrenzte Segmente, die für eine Verstetigung geeignet sind.

Die Komponenten des Produktlevels 1 (a und b) sind essentiell für die OCR-Generierung. Die Komponenten des Produktlevels 2 sind für die OCR-Generierung optional, auch wenn sie einen bedeutenden Mehrwert bieten. Während die Komponenten von Produktlevel 1a während der jetzigen Projektphase vom OCR-D-Koordinierungsprojektteam aktuell gehalten werden, werden die Komponenten der Produktlevel 1b und 2 in Zusammenarbeit mit der OCR-D-Entwickelnden-Community gepflegt. Mit Verstetigung der Software werden die Komponenten des Produktlevels 1a über Entwicklungsaufträge (siehe Kapitel Weiterentwicklung der Software (Entwicklungsfonds)) aktualisiert. Für die Pflege der Produktlevel 1b sowie 2 wird dann die OCR-D-Entwickelnden-Community in der Verantwortung sein.

Zum Produktlevel 1a zählen:

- [spec](#): OCR-D-Standards und -Spezifikationen
- [core](#): OCR-D-Python-Implementierung
- [gt-guidelines](#): OCR-D-GT-Richtlinien
- [ocrd-website](#) / [ocr-d.github.io](#): OCR-D-Website

Zum Produktlevel 1b zählen:

- [ocrd\\_all](#): OCR-D-Installation via venv oder Docker mit
  - [ocrd\\_tesseract](#): 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

Zum Produktlevel 2 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 frei verfügbar (Lizenzen: Apache 2.0 bzw. GPLv3).
- Sie wird nach dem Prinzip „Continuous 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 (Nutzende 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 aktuellen, 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<sup>2</sup> und Entwicklung eines automatisierten Benchmarking aufgrund der von der Community zurückgemeldeten Metriken
- Überarbeitung der Dokumentationen sowie Konzeption und Umsetzung der Betriebs- und Hostinginfrastruktur

Insbesondere wurde in Abstimmung mit dem Kitodo-Verein ein Review der OCR-D-Software durch einen externen Dienstleister (Kitodo-Releasemanagement) durchgeführt, das die Robustheit der Software bestätigte.<sup>3</sup>

<sup>2</sup> Quiver auf Github: <https://github.com/OCR-D/quiver-benchmarks>; Quiver-Dashboard: <https://ocr-d.de/quiver-frontend/#/workflows?view=table>

<sup>3</sup> Die Reviewergebnisse sind schriftlich dokumentiert. Die wenigen und geringfügigen Monita sind abgearbeitet und in einem Whitepaper dokumentiert.

### 4.3 SUPPORT DER SOFTWARE ÜBER DIE PROJEKTPHASE HINAUS

Für die vollumfängliche Nutzung der OCR-D-Software sind kontinuierliche Entwicklungs-, Update- und Wartungsmaßnahmen erforderlich. Diese stellen sicher, dass die Software den sich dynamisch verändernden Nutzeranforderungen ebenso Rechnung trägt wie dem stetig wandelnden IT-Umfeld. Die Beseitigung von Fehlern und die Schließung von Sicherheitslücken sind dabei zentrale Aspekte, um optimale Leistung und Zuverlässigkeit zu gewährleisten und der technologischen Obsoleszenz entgegenzuwirken. Gleichwohl erfordern Änderungen am Softwarecode und den dazu gehörigen Dokumentationen stetige Aktualisierungen und Pflege der OCR-D Benutzerdokumentationen, insbesondere wenn diese Code-Änderungen Auswirkungen auf den Funktionsumfang haben.

Während der Projektphase wurden Standards und Prozesse festgelegt, um sicherzustellen, dass die veröffentlichte Software von hoher Qualität ist und bleibt. Dies beinhaltet die Verfeinerung der bestehenden Testverfahren, die Überprüfung von Code und die Fehlerbehebung vor der Veröffentlichung. Das ermöglicht den Entwickelnden, Fehler zu identifizieren und zu beheben, bevor unter Einbeziehung der Rückmeldungen der OCR-D-Nutzenden eine neue Version veröffentlicht wird, was die Stabilität und Zuverlässigkeit der Software insgesamt erhöht. Eine Software-Versionierung mittels Semantic Versioning kennzeichnet klar die verschiedenen Iterationen der Software. Dies erleichtert die Identifizierung und den Vergleich von verschiedenen Versionen. Derzeit ist für den Core-Code Version 3.0 aktuell.

Um den Support der OCR-D-Ergebnisse gewährleisten zu können, sind u. a. die folgenden Kompetenzen erforderlich:

**Entwicklung:** Entwickelnde kennen die Software auf Code-Ebene. Sie sorgen dafür, dass die Software einwandfrei einsatzbereit ist, aktualisieren sie auf den neuesten Stand, verbessern bereits bestehende Funktionen und entwickeln neue bei Bedarf.

**Technischer Support:** Die Mitarbeitenden für den technischen Support kennen die technische Anwendung der Software. Sie unterstützen bei der Installation, Konfiguration, Fehlerbehebung, Aktualisierung sowie in der Benutzung und pflegen die dazugehörige Dokumentation. Da sie im direkten Kontakt mit den Anwendenden stehen, geben sie Rückmeldungen (zum Beispiel Verbesserungsvorschläge oder Funktionswünsche) direkt in das OCR-D-Koordinierungsteam.

**Release Management:** 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\*in sammelt und organisiert Software-Änderungen, Erweiterungen und Fehler-behebungen, sowohl in der Kernsoftware als auch in einzelnen Prozessoren. Er/Sie arbeitet eng mit den Entwickelnden und der Community zusammen, um den geeigneten Release-Inhalt und -Zeit-punkt zu bestimmen.

**Öffentlichkeitsarbeit:** Die OCR-D-Software bekannt zu machen sowie über Funktionen und Einsatzbereiche zu informieren, ist Aufgabe der Öffentlichkeitsarbeit. 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. Sie fördert damit den Aufbau einer aktiven und engagierten Community. Das ist wichtig, um

die bereits bestehende, breite OCR-D-Community mit ihren umfangreichen und wertvollen Beiträgen zu erhalten und weiterhin eng mit einzubeziehen. Sie trägt maßgeblich dazu bei, Komponenten der Software zu entwickeln und zu pflegen. 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 Prozessierungsworkflows.

Der genaue Personalbedarf und die organisatorische Struktur für die hier genannten Kompetenzen und Aufgaben wird im Abschnitt 4.4.2 des Konzepts zur VD-Volltexttransformation dargelegt.

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

**Integration:** Es ist geplant, dass das Releasemanagement und die notwendigste Weiterentwicklung der OCR-D-Software über die Integration der Software als drittes Modul von Kitodo im Rahmen des Kitodo e. V. zu realisieren (siehe dazu die Abschnitte 6.5 und 6.6 sowie Kapitel 7).

## 5 BETRIEBSMODELLE

### 5.1 GRUNDSÄTZLICHES

Die OCR-D-Software steht kostenfrei und quelloffen als Apache 2.0 bzw. GPLv3 lizenziert zur Verfügung. Sie kann von den Anwendenden oder potentiellen Anbietenden eigenständig und unabhängig heruntergeladen, installiert und genutzt werden (On-Premise-Deployment). Die OCR-D-Software wird zudem zentral bereitgestellt und kann als zentraler Service genutzt werden (OCR-as-a-Service). Auch kombinierte Verfahren kommen in Frage.

### 5.2 ON-PREMISE-DEPLOYMENT

#### 5.2.1 LOKAL-DEZENTRALER ANSATZ

Bei einem On-Premise-Deployment installieren sich die Anwendenden die (neueste Version der) OCR-D-Software lokal auf eigenen Rechnern, um sie dort zu verwenden. Dieser Weg bietet sich an, wenn eine Einrichtung den OCR-as-a-Service (siehe Kapitel OCR-as-a-Service) nicht verwenden möchte oder, wenn eine erneute, spezifischere OCR für bestimmte Dokumente erforderlich ist. Erfahrungsgemäß werden Support bei der Installation, individuelle fachliche Beratung zur Benutzung der Software oder weitergehende Schulungen benötigt werden. Mit diesem Ansatz kann die Nutzendengemeinschaft gestärkt werden und selbst unterstützend mitwirken, da es einen Austausch zum Beispiel zu Usability-Aspekten und nutzungsbezogenen Bedarfen und Einsatzszenarien gibt.



## 5.2.2 WORKFLOW

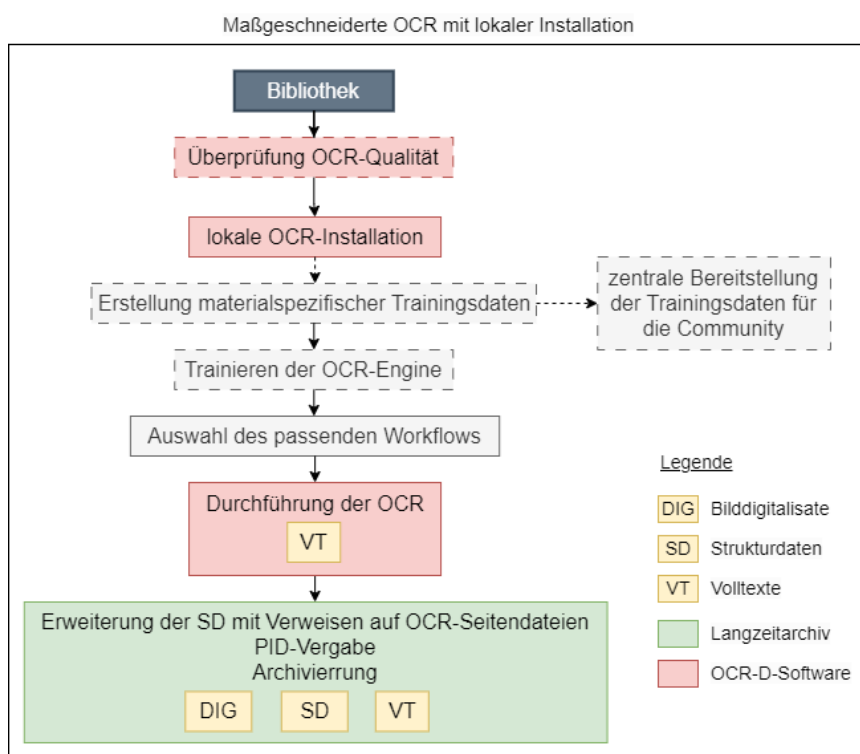


Abbildung 1 Workflow der maßgeschneiderten OCR-Generierung mittels lokaler Installation der OCR-D-Software.

Bei lokalen Installationen erfolgt die OCR-Generierung dezentral (siehe Abbildung 1). 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<sup>4</sup> hervorgehen). Sind bereits OCR-erstellte Volltexte vorhanden, wird deren Qualität von Expert\*innen anhand von Evaluierungsroutinen überprüft. Ist die Erkennung nicht zufriedenstellend (Erkennungsquote/Fehlerrate), können mittels der lokalen Installation materialspezifische Trainingsdaten erstellt werden, womit die OCR-Engine trainiert wird. Nach Auswahl des passenden Workflows wird die OCR-Prozessierung durchgeführt. Bereits die Auswahl passgenauer Prozessoren in OCR-D-Workflows führt in der Regel zu besseren Ergebnissen. Die METS-Dateien werden um Verweise auf die OCR-Seitendateien erweitert und es wird ein Persistent Identifier (PID) vergeben. Es wird empfohlen, dass die generierten Volltexte zusammen mit den Bilddigitalisaten und den METS-Dateien (inklusive der Strukturdaten) zentral (zum Beispiel in OLA-HD) langzeitarchiviert werden.

Damit stehen die erstellten Trainingsdaten der Community zur Verfügung, sodass sie auch bei anderen Korpora verwendet werden können und so das Vorhaben der Volltextgenerierung der gesamten VD-Bestände unterstützt wird.

## 5.3 OCR-AS-A-SERVICE

### 5.3.1 ZENTRALER ANSATZ

Es wird ein zentraler Dienst aufgebaut, der die Nutzung der OCR-D-Software anbietet (OCR-as-a-Service). Die OCR-D-Software und die erforderliche IT-Infrastruktur mit den entsprechenden Rechen-

<sup>4</sup> <https://ola-hd.ocr-d.de/> (zuletzt aufgerufen am 14.10.2024)

und Speicherkapazitäten werden bei einem Dienstleister betrieben und können gegen eine Kostenerstattung als Dienstleistung genutzt werden. Alle Softwarekomponenten werden vom Dienstleister bereitgestellt. Auf diese Weise entstehen keine Aufwände für die lokale Installation und deren Betrieb. Die Software kann so auch leichter aktuell gehalten werden. Außerdem wird dadurch die Effizienz der Bearbeitung erhöht und das Homogenhalten der Daten erleichtert. Im Zuge der vorgeschlagenen 4. Phase ist es vorgesehen, OCR-as-a-Service in den Produktivbetrieb zu überführen, um den bereits bilddigitalisierten VD-Bestand zu volltextdigitalisieren (vgl. Konzept zur Volltexttransformation der VD, Abschnitt 4.2.1).

### 5.3.2 WORKFLOW

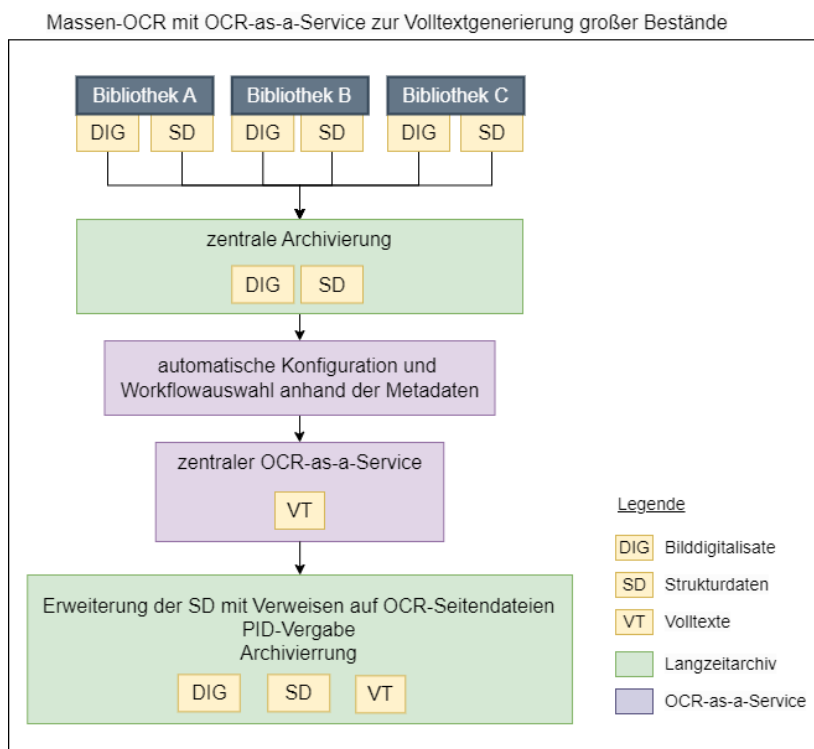


Abbildung 2 Workflow der Massen-OCR mit zentralem OCR-Service zur Volltextgenerierung großer Bestände.

Im Rahmen des OCR-D-Implementierungsprojekts OPERANDI haben die SUB Göttingen und die GWDG das OCR-as-a-Service-Modell entwickelt, das die umfangreichen technischen Ressourcen der HPC-Umgebung nutzt. Ziel ist es, eine stabile Massen-Volltexterfassung mit verbessertem Durchsatz und höherer Ergebnisqualität zu ermöglichen und diese in einer möglichen vierten Phase der OCR-D-Koordinierung im Produktivbetrieb zu überprüfen, um die VD-Bestände effizient volltextdigitalisieren zu können. Darüber hinaus haben die gleichen Einrichtungen im OCR-D-Modulprojekt OLA-HD ein Langzeitarchiv für historische Drucke entwickelt, das mittlerweile auch als Nachweissystem für Ground Truth-Daten dient.

Die OCR-Prozessierung kann sowohl aus bestehenden Digitalisierungsworkflows (z. B. über Kitodo oder Goobi) gestartet werden als auch durch das zugrunde liegende Integrationsskript unabhängig davon genutzt werden. Die Interaktion mit dem OCR-as-a-Service in OPERANDI erfolgt vollständig über eine REST-basierte Web-API. Ein zusätzlicher Startpunkt kann OLA-HD sein, insbesondere wenn eine bereits abgeschlossene Volltextdigitalisierung aufgrund verbesserter Workflows oder Trainingsmaterialien

erneut durchgeführt werden soll. Dies könnte auch dazu beitragen, die Einstiegshürde für weniger technisch versierte Nutzer zu senken und eine suchfilterbasierte Massenverarbeitung zu ermöglichen.

Die wesentlichen Schritte sind: 1) Automatischer Import der Daten, einschließlich des Aufbaus des Workspaces, z. B. durch METS-URLs oder den Upload von Bildern im ZIP-Format; 2) Teilautomatisierte Auswahl oder Nutzung eines bestehenden Workflows, der die Prozessierungspipeline beschreibt, oder ggf. die Beschreibung und den Upload eines neuen Workflows; 3) Überprüfung des Prozessierungsfortschritts; und nach erfolgreichem Abschluss 4) Bereitstellung der OCR-Ergebnisse im PAGE XML- oder ALTO-Format sowie Speicherung dieser Ergebnisse im OCR-Archiv (OLA-HD). Im Falle einer gescheiterten Prozessierung erfolgt der Download der Logs. Die OCR-Ergebnisse im PAGE XML- oder ALTO-Format können in die Digitalisierungsworkflows integriert und in das OLA-HD importiert werden. Abschließend erfolgen 5) Aufräumarbeiten, indem die Logs und der nicht mehr benötigte Workspace gelöscht werden.

Der Index von OLA-HD ermöglicht neben der OCR-Volltextsuche und den üblichen bibliographischen Filteroptionen wie Autor, Erscheinungsjahr oder -ort auch OCR-spezifische Filter, z. B. nach verwendetem Prozessor oder semantischen Eigenschaften, etwa ob die Vorlage Tabellen oder Stempel enthält oder ob es sich um ein- oder mehrspaltige bedruckte Seiten handelt. Dadurch können OCR-Ergebnisse identifiziert werden, die mit einem fehlerhaften Prozessor bearbeitet wurden, um diese erneut mit einem korrigierten oder verbesserten Workflow zu überarbeiten und den Archivinhalt schrittweise zu optimieren. Zudem ist es möglich, Vergleichswerke zu ermitteln, um einen Workspace mit einem neuen oder geänderten Workflow zu prozessieren und die erzielten Verbesserungen oder Verschlechterungen zu überprüfen.

#### 5.4 KOMBINIERTES BETRIEBSMODELL

Der OCR-as-a-Service-Ansatz ist für die Massenverarbeitung optimiert und liefert einen hohen Durchsatz bei guter Erkennungsrate. Im Gegensatz dazu bietet das On-Premise-Deployment eine lokale Integration, bessere individuelle Anpassungsmöglichkeiten und mit Hilfe von gezieltem Training und Ground Truth (GT) eine potenziell höhere Erkennungsrate für sinnvoll ausgewählte Korpora. Das On-Premise-Deployment eignet sich daher für Einrichtungen und Projekte, bei denen eine lokale Integration in bestehende Systeme und Workflows von besonderer Bedeutung ist oder für Situationen, in denen spezifische Anforderungen für wissenschaftliche Auswertungen erfüllt werden müssen. Ein wesentlicher Vorteil des zentralen Ansatzes hingegen ist die Möglichkeit, OCR-Daten samt Metadaten in größerem Umfang für die weitere Nutzung bereitzustellen. Dies ermöglicht neue wissenschaftliche Möglichkeiten, wie Textmining-Analysen und der Aufbau eines zentralen Indexes, wie im VD-Suchraum-Projekt geplant.

Die Betriebsmodelle (dezentral/lokal und zentral über den Service) schließen sich nicht gegenseitig aus. Empfohlen wird, 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, wie zum Beispiel für ein Korpus von Leichenpredigten, die zusätzliche oder variierende Schritte in den OCR-Workflows benötigen, um eine höhere Erkennungsrate zu erreichen. Um Redundanzen zu vermeiden, und um den Bedarf für lokale Installationen zweifelsfrei empirisch nachzuweisen, wird von der Reihenfolge her vorgeschlagen, zunächst die zentrale HPC-gestützte Bearbeitung zu fördern und durchzuführen; darauf

kann die dezentrale Installation bei Bedarf (z. B. bei niedrigen Erkennungsraten/hohen Fehlerraten der zentralen Bearbeitung oder bei nachgewiesenen Bedarfen der wissenschaftlichen Nutzenden) geplant, beantragt und ggf. gefördert werden.

## 6 ORGANISATIONSMODELLE

### 6.1 VORGEHEN

Es gibt verschiedene Organisationsmodelle, die im Rahmen des Koordinationsprojekts evaluiert wurden. Hierzu wurden Gespräche mit Organisationen aus vergleichbaren Themenbereichen geführt. Impulse zur Bewertung der Modelle kamen außerdem durch Beiträge des DFG-Rundgesprächs vom (24./25. August 2023) in Göttingen und weiteren Austausch der Beteiligten im Nachgang der Veranstaltung. Erwogen wurden 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.“.

### 6.2 HATHITRUST DIGITAL LIBRARY

Die HathiTrust Digital Library (HathiTrust) ist ein gemeinsam verwalteter und ko-finanzierter 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 Forschende, Studierende und Bibliotheksnutzende, 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 Umfang der Daten-Kollektion (cost-per-volume). Die Gebühren für gemeinfreie und urheberrechtlich geschützte Werke werden unterschiedlich berechnet.

### 6.3 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 (eG) 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 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.

#### 6.4 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“ 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 Vertretenden 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 Arten von Mitgliedschaften: 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 Mitgliedschaften haben vergünstigtem Zugang zu IMPACT-Veranstaltungen und Konferenzen.

#### 6.5 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 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.<sup>5</sup> Die finanzielle Unterstützung trägt dazu bei, dass der Verein seine Ziele erreichen und die Bemühungen zur Förderung von der Softwaremodule von Kitodo fortsetzen kann. Zwischen Kitodo-Verein und OCR-D-Koordinierung besteht seit 2019 eine Vereinbarung zur Kooperation bei der Volltexterstellung innerhalb von Digitalisierungsworkflows.

## 6.6 ENTSCHEIDUNG

Der Aufbau einer dem HathiTrust oder dem IMPACT-Kompetenzzentrum vergleichbaren Organisation wird als eine zu komplexe Aufgabe angesehen, die einen unverhältnismäßig hohen Aufwand für Planung und Koordination sowie einen unkalkulierbar hohen Personal- und Mitteleinsatz erfordern würde. Die spezifischen OCR-D-Bedürfnisse müssten im Hinblick auf eines der beiden Modelle neu definiert werden, vor allem aber müsste ein solches Modell an die deutschen rechtlichen Rahmenbedingungen angepasst werden. Die Option als interessierte Institutionen selbst Mitglied bei HathiTrust oder IMPACT zu werden, würde hohe Kosten für die einzelne Einrichtung nach sich ziehen (Jahresbeiträge zwischen 6.600 und 13.000 USD, je nach Bibliotheksbudget und Datenmenge). HathiTrust hat bislang keine deutschen, institutionellen Mitglieder,<sup>6</sup> was neben den Kosten vor allem daran liegt, dass die Nutzung digitalisierter Werke und der Zugang zu deren Inhalten von HathiTrust nach den Regeln des US-amerikanischen Urheberrechts reguliert wird, was mit den deutschen Rechtsstandards und der bevorzugten Open Source- bzw. Open Access-Prämisse des OCR-D-Projekts kollidiert. Bei IMPACT würde sich die Option, die Verstetigung der Software durch Mitgliedschaften der beteiligten Institutionen zu realisieren, wegen des notwendigen Bedarfs zu implementieren und weiterzuentwickeln auf die Premium-Mitgliedschaft beschränken. Der jährliche Beitrag dafür würde ebenfalls sehr hohe Jahresbeiträge (6.000 € zzgl. Mehrwertsteuer) erfordern. Wegen der fehlenden Möglichkeiten zu Implementierung und Weiterentwicklung bei der Standardmitgliedschaft kommt diese finanziell günstigere Option nicht in Betracht.

Eine Genossenschaft wie die READ-COOP SCE ist wirtschaftlich ausgerichtet, d. h. sie arbeitet gewinnorientiert. Gewinnorientiert zu wirtschaften, ist nicht mit den Zielen und der Open Source-basierten Ausrichtung des OCR-D-Projekts vereinbar. Das spricht gegen eine Beteiligung der an OCR-D beteiligten Institutionen an READ-COOP SCE. Zudem gibt es aus denselben Gründen kein Interesse bei

<sup>5</sup> Stand Okt. 2024: [https://www.kitodo.org/fileadmin/groups/kitodo/Verein/Kitodo\\_Beitragordnung.pdf](https://www.kitodo.org/fileadmin/groups/kitodo/Verein/Kitodo_Beitragordnung.pdf)

<sup>6</sup> Vgl. <https://www.hathitrust.org/member-libraries/member-list/>

den in Frage kommenden Institutionen aus der OCR-D-Community, sich an der Gründung einer eigenen, vergleichbaren Genossenschaft zu beteiligen.

Durch die Integration in einen bestehenden, etablierten Verein mit angrenzenden bzw. sich überschneidenden, inhaltlichen Zielen kann der Aufwand, neue (Parallel-)Strukturen aufzubauen vermieden werden. Stattdessen werden die bestehenden Strukturen und Ressourcen gestärkt und genutzt. Kitodo e. V. hat in den zurückliegenden Jahren bewiesen, dass sein Geschäftsmodell funktioniert, und konnte bereits eine für den Betrieb seiner Software und Strukturen erforderliche Reichweite aufbauen.

Im DFG-geförderten VD-Rundgespräch 2023 wurde festgestellt, dass ein Großteil der Mitglieder von Kitodo e. V. grundsätzlich auch zur OCR-D-Zielgruppe gehört. Zudem ist die OCR-D-Software Werkzeug für einen Schritt im größeren Digitalisierungs- und Volltext-Workflow. Daher wird die OCR-D-Software allein von den Anwendenden weniger als eigenständiges Einzelprodukt, sondern vielmehr als eine sinnvolle, zunehmend unverzichtbar werdende Komponente im Gesamtdigitalisierungsprozess angesehen. Aus diesen Gründen waren sich die Teilnehmenden im VD-Rundgespräch 2023 einig und es wurde beschlossen, dass eine organisatorische Anbindung von OCR-D an Kitodo e. V. angestrebt werden sollte und die anderen Modelle und Kooperationen nur dann in Erwägung gezogen werden sollen, wenn die Kooperation mit dem Kitodo e. V. nicht zustande kommen kann. Ein weiterer Vorteil der Integration von OCR-D-Ergebnissen in die Vereinsstrukturen des Kitodo-Vereins besteht darin, dass die bestehende Governance-Struktur des Vereins effizient genutzt werden kann. Die Vereinsmitglieder beteiligen sich aktiv an demokratischen Entscheidungsprozessen zur Weiterentwicklung der Software, während sie gleichzeitig gemeinsames technisches und organisatorisches Wissen im Bereich der Digitalisierung von Kulturgut aufbauen. So werden der Sachverstand und die berechtigten Interessen der Community bei der Governance einbezogen. Zudem organisieren und finanzieren sie gemeinsam ein professionelles Release-Management. Viele Mitglieder leisten darüber hinaus freiwillig finanzielle Beiträge zum Entwicklungsfonds, aus dem regelmäßig benötigte Programmierarbeiten finanziert werden, um beide Softwaremodule auf dem neuesten Stand zu halten. Es wurde daher im Laufe der dritten Projektphase geprüft, ob OCR-D in den Kitodo-Verein aufgehen und OCR-D-Software analog zu Kitodo-Software über die Vereinsstrukturen angeboten und gepflegt werden kann. Eine solche Integration schließt die Nutzung der OCR-D-Software mit anderer Digitalisierungssoftware wie Goobi oder Dwork nicht aus. Beide Betriebsmodelle, sowohl der zentrale als auch der dezentrale Ansatz, wurden in Verbindung mit Goobi in Modulprojekten erfolgreich getestet.

## 7 VERSTETIGUNG

### 7.1 INTEGRATION VON OCR-D IN KITODO E. V.

In einem intensiven Austausch wurde von der OCR-D-Koordinierung und den Kitodo-Vereinsstrukturen (vor allem Vorstand und Mitgliederversammlungen) seit vor und nach dem VD-Rundgespräch 2023 mehrfach konkret über den Stand, die Desiderate sowie über notwendige Folgeschritte gesprochen sowie Anforderungen zur Qualität an die OCR-D-Software für die Integration in den Kitodo-Verein erörtert. Konsens über den notwendigen Reifegrad der Software und die Überschaubarkeit von Risiken für den Verein stellen die wichtigsten Voraussetzungen für eine Aufnahme unter das Dach des Vereins dar. Dazu wurde ein Code Review von der für einen Teil des Kitodo-Releasemanagements zuständigen

Firma effective Webwork durchgeführt, das die Qualität der OCR-D-Software bestätigt hat. Der jüngste intensivere Austausch zwischen der OCR-D-Koordinierung und dem Verein fand im Rahmen der Kitodo-Mitgliederversammlung bei der BiblioCon in Hamburg am 5. Juni 2024 statt, bei der über die Reviewergebnisse und die Folgerungen aus dem Review berichtet wurden.

Die angestrebte Integration der OCR-D-Software als neues Kitodo-Modul umfasst dazu festgelegte Kernkomponenten (Produktlevel 1a und 1b), nicht aber diejenigen Softwarekomponenten, die optionale Funktionen bieten (Produktlevel 2) und die in anderen Projekten und von der Community entwickelt werden können.

Des Weiteren wurde zwischen der OCR-D-Koordinierung und dem Kitodo-Vereinsvorstand kommuniziert, dass Kosten für ein eigenes OCR-D-Release Management in Höhe der Aufwandsentschädigung für die bisherigen Release Management-Arbeiten einkalkuliert (siehe Releasemanagement von Kernmodulen der Software) sowie ein Entwicklungsfond in Höhe der Entwicklungsfonds der anderen Module innerhalb des Vereins aufgebaut werden soll (siehe Weiterentwicklung der Software (Entwicklungsfonds)).

Die Partnerinstitutionen in der OCR-D-Koordinierung (HAB Wolfenbüttel, SB zu Berlin, BBAW, SUB Göttingen, GWDG) streben an, den Verein bei der Einwerbung dieser Beträge zu unterstützen. Konkret werden alle Partner als Mitglied in den Kitodo-Verein eintreten bzw. sofern sie schon Mitglied sind, nach Maßgabe der haushälterischen Möglichkeiten ihre Beiträge stabil halten oder erhöhen sowie nach Möglichkeit, das Gewinnen zusätzlicher Mitglieder unterstützen. Diese Übergangsphase sollte durch die vorgeschlagene vierte Phase der OCR-D-Koordination unterstützt werden (siehe Konzept zur Volltexttransformation der VD). Dadurch wird die Integration stabilisiert, und die notwendigen Arbeiten im Bereich des Release-Managements und der Weiterentwicklung, die zukünftig durch Vereinsbeiträge und den Entwicklungsfonds finanziert werden sollen, auch in der Übergangszeit nahtlos fortgeführt und abgesichert.

## 7.2 RELEASEMANAGEMENT VON KERNMODULEN DER SOFTWARE

Die Pflege der notwendigsten Module der OCR-D-Software sowie das Release Management der neuen Versionen wird bei Zustimmung der Vereinsgremien im Rahmen des Kitodo e. V. organisiert und verwaltet. Der Verein kalkuliert auf der Basis seiner Erfahrungen mit den bisherigen Modulen für das Release Management mit einer Entschädigung in Höhe von ca. 10.000 € jährlich. Das Releasemanagement kann in Form einer Beauftragung für eine Firma mit Aufwandsentschädigung in der genannten Höhe oder durch Garantien für den Einsatz eigener Personalressourcen zum selben Zweck organisiert werden.

## 7.3 WEITERENTWICKLUNG DER SOFTWARE (ENTWICKLUNGSFONDS)

Für die Weiterentwicklung der OCR-D-Software ist der Aufbau eines Entwicklungsfonds geplant, der ca. 35.000 € p. a. umfassen soll. Das ist in der Struktur und Höhe vergleichbar mit denen der anderen Module des Vereins. Mit diesen Ressourcen können Aufträge an Firmen und Personen vergeben werden, die konkrete Funktionen für die OCR-D-Software entwickeln.

Die OCR-D-Kooperationspartner\*innen erklären sich dazu bereit, eigene Mittel zum Aufbau dieses Fonds beizusteuern sowie weitere Einrichtungen zur Beteiligung daran zu gewinnen und Mittel einzuwerben.



Bei einem OCR-as-a-Service kann eingeplant werden, einen Anteil des Seitenpreises, der bei der Prozessierung anfällt, dem Entwicklungsfond zuzuführen, um damit das Fortbestehen und die Weiterentwicklung der Software sicherzustellen.

Zudem ist es notwendig, die bereits bestehende breite OCR-D-Entwickelnden-Community mit ihren umfangreichen und wertvollen Beiträgen zu erhalten und weiterhin eng mit einzubeziehen. Sie kann maßgeblich dazu beitragen, Komponenten der Software weiterzuentwickeln, sofern dafür - etwa über den Entwicklungsfonds - Finanzierungs- oder Förderungsmöglichkeiten existieren.

#### 7.4 LANGZEITVERFÜGBARKEIT UND LANGZEITARCHIVIERUNG

Die Langzeitverfügbarkeit der Software wird durch die Aufnahme in den Kitodo e. V. und die damit verbundene Aktualisierung der Software sowie ihrer Dokumentation sichergestellt.

In einem Langzeitarchiv als Bestandteil des zentralen Dienstes (OCR-as-a-Service) können die Bild- und Textdigitalisate, die dazugehörigen METS-Daten (inklusive Strukturdaten) sowie die erzeugten Volltexte gespeichert und die Daten für den von den VD-Trägerbibliotheken geplanten VD-Suchraum bereitgestellt werden. Es wird empfohlen das Langzeitarchiv „OLA-HD Service“ (<https://ola-hd.ocr-d.de/>) zu verwenden, da dieses speziell auf die benötigten Anforderungen hin ausgelegt ist.

## 8 EXECUTIVE SUMMARY

Das Konzept zur Verstetigung der OCR-D-Ergebnisse verfolgt das Ziel einer nachhaltigen Nutzung und Weiterentwicklung der OCR-D-Software, um das Ziel einer möglichst vollständigen Texterkennung der VD-Bestände zu gewährleisten.

Es werden zwei bzw. drei Betriebsmodelle beschrieben, eine zentrale Umsetzung, eine dezentrale sowie eine kombinierte Vorgehensweise. Bei einem On-Premise-Deployment (dezentrale Lösung) installieren sich die Anwendenden die (neueste Version der) OCR-D-Software lokal auf eigenen Rechnern, um sie dort zu verwenden. Bei OCR-as-a-Service wird ein zentraler Dienst aufgebaut, der die Nutzung der OCR-D-Software anbietet. Empfohlen wird, dass beide Modelle angeboten werden. Je nach vorhandener Infrastruktur, finanziellen Mitteln, personeller Ausstattung und Erfahrung, sowie Besonderheiten der Gruppe von Vorlagen kann eine zentrale oder dezentrale Lösung genutzt werden.

Nach eingehender Prüfung verschiedener Organisationsmodelle wurde entschieden, die OCR-D-Software möglichst in den Kitodo e. V. zu integrieren. Dies ermöglicht es auf bestehende, funktionierende Strukturen im Digitalisierungskontext zurückzugreifen und Ressourcen zu bündeln, anstatt (neue) Parallelstrukturen aufzubauen. Durch die Integration in den Kitodo-Verein wird eine langfristige Pflege und Verfügbarkeit der Kernkomponenten sichergestellt. Ein weiterer Vorteil besteht darin, dass die bestehende Governance-Struktur des Vereins effizient genutzt werden kann.

Es wird angestrebt, dass im Rahmen des Kitodo-Vereins das Releasemanagement der notwendigsten Module der OCR-D-Software organisiert und verwaltet wird. Die Kosten dafür werden auf etwa 10.000 € pro Jahr geschätzt. Ein Entwicklungsfonds in Höhe von etwa 35.000 € pro Jahr soll für die Weiterentwicklung der OCR-D-Software eingerichtet werden. Letztlich entscheiden die Mitglieder über

die Strukturänderung bzw. über Änderungen in der Satzung. Die Partnerinstitutionen des derzeitigen OCR-D-Koordinierungsprojekts streben an, den Verein bei der Einwerbung von Beträgen für den Entwicklungsfonds zu unterstützen. Die Langzeitverfügbarkeit der Software wird durch die Aufnahme in den Kitodo-Verein und die damit verbundene Aktualisierung der Software sowie ihrer Dokumentation sichergestellt. Ein Langzeitarchiv wurde eingerichtet, um die Bild- und Textdigitalisate, die dazugehörigen METS-Daten und die erzeugten Volltexte zu speichern. Nach intensivem Austausch, nach abgeschlossener Review der Software und Vorliegen der in diesem Zusammenhang benötigten Version 3 der Software bestehen günstige Voraussetzungen dafür, die Integration in den Verein zu realisieren.