ABSCHLUSSBERICHT

der Koordinierten Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR) im Programm für Wissenschaftliche Literaturversorgungs- und Informationssysteme (LIS)

1 Allgemeine Angaben

DFG-Geschäftszeichen:	BU 1218/38-1 (HAB), GE 1119/12-1 (BBAW), HO 3987/65-1 (SUB), SCHN 743/75-1 (SBPK), WI 5054/2-1 (GWDG)
Projektnummer:	AOBJ: 674752
Titel des Projekts:	Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR) [Phase 3]
Namen d. Antragstellenden:	Prof. Dr. Peter Burschel, Privatdozent Dr. Alexander Geyken, Zeki Mustafa Dogan, Barbara Schneider-Kempf, Prof. Dr. Philipp Wieder
Ehemalige Antragstellende:	Prof. Dr. Wolfram Horstmann

Dienstanschrift/en:
Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22/23, 10117 Berlin; Stiftung Preußischer Kulturbesitz / Staatsbibliothek zu Berlin, Unter den Linden 8, 10117 Berlin; Niedersächsische Staats- und Universitätsbibliothek, Platz der Göttinger Sieben 1, 37073 Göttingen; Gesellschaft für wissenschaftliche Datenverarbeitung mbH, Burckhardtweg 4, 37077 Göttingen; Herzog August Bibliothek, Lessingplatz 1, 38304 Wolfenbüttel

Namen der Mitverantwortlichen: Reinhard Altenhöner, Johannes Mangei

Namen der Kooperationspartner*innen: Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), Staatsbibliothek zu Berlin, PK (SBB), Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB), Herzog August Bibliothek Wolfenbüttel (HAB), Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG).

Kooperierende Modul- und Implementierungspartner*innen: Universitätsbibliothek Braunschweig, Sächsische Landesbibliothek – Staats- und Landesbibliothek Dresden, Universitätsbibliothek Mannheim, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Georg Eckert Institut Braunschweig, HCI und ZPD der Universität Würzburg, Universitäts- und Landesbibliothek Sachsen-Anhalt, Johannes Gutenberg-Universität Mainz, Friedrich-Alexander-Universität Erlangen-Nürnberg[1]

Berichtszeitraum (gesamte Förderdauer): 01.04.2021 – 31.05.2025

Projektstart 01.04.2021 (BBAW, GWDG) (erste Einrichtungen) – Projektende 31.05.2025 GWDG (letzte Einrichtung, inklusive kostenneutrale Verlängerung)

2 Zusammenfassung / Summary

Die „Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition“ (OCR-D) hat seit 2015 in drei Förderphasen Lösungen für die automatische Texterkennung mit einem Fokus auf gedruckten historischen Materialien entwickelt, wie sie in den „Verzeichnissen der im deutschen Sprachraum erschienenen Drucke des 16.–18. Jh.“ (VD16, VD17, VD18, insgesamt: VD) nachgewiesen sind. Übergeordnetes Ziel von OCR-D war es, einerseits ein robustes Verfahren zu entwickeln und zu erproben, eine leistungsfähige, stabile Softwarelösung und einen möglichst effizienten Workflow zu entwickeln, andererseits ein Konzept zur Transformation der VD in eine maschinenlesbare Form (strukturierter Volltext) und ein Konzept zur Verstetigung der Software zu erstellen. Über die ersten beiden Projektphasen liegen eigene Abschlussberichte vor, so dass hier der Bericht über die nun abgeschlossene dritte Projektphase (Phase III) im Mittelpunkt steht. Die beiden Konzepte wurden im Oktober 2024 über die DFG-Geschäftsstelle dem Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme (AWBI) vorgelegt, der sie in seiner Herbstsitzung beraten und bewertet hat. Die Konzepte sowie der Protokollauszug des AWBI sind dem Abschlussbericht beigefügt. Die Robustheit der Software und die Qualität der Dokumentation wurden in einer Software-Review durch einen externen Gutachter auf Vorschlag des Kitodo-Vereins hin überprüft und positiv bewertet. Seine Monita wurden bis zum Projektende Punkt für Punkt ausgeräumt und ebenfalls

[1] https://ocr-d.de/de/phase3

dokumentiert (das Review sowie die Whitelist befinden sich im Anhang dieses Berichts). Zuletzt wurde mit Hausmitteln der beteiligten Einrichtungen die Integration der Software in die Strukturen des Kitodo-Vereins, die Antragserstellung für eine Förderung der Volltexterstellung für die VD sowie die Verzahnung dieser Volltexterstellung mit der Planung für ein gemeinsames Portal der VD ("VD-Portal) vorangetrieben.

3 Arbeits- und Ergebnisbericht

3.1 Ausgangslage und Zielsetzung des Projekts

Die Arbeiten in Phase III hatten zum Ziel, die Ergebnisse der vorangegangenen Projektphasen zu stabilisieren, die OCR-D-Software weiter zu optimieren und ihren Einsatz für die Massendigitalisierung technisch und organisatorisch einzuleiten. Dabei lag ein Schwerpunkt auf der Betreuung der entsprechenden Implementierungs- (IP) und Modulprojekte (MP), ein anderer auf der Sicherstellung nachhaltiger Betreuung und Weiterentwicklung der Software (Verstetigung), ohne dass in diesem Zuge die Funktionsweise des Gesamtworkflows eingeschränkt wird. Zudem sollten die Ergebnisse an einen breiten Kreis von Nutzer*innen vermittelt werden (Verbreitung), um die effiziente Volltext-Digitalisierung, vor allem von VD-Beständen, zu ermöglichen.

3.2 Arbeitsschritte im Berichtszeitraum

Die Arbeitsschritte im Berichtszeitraum waren in drei Arbeitspakete gegliedert: AP1: Koordination und Kommunikation, AP2: Implementierung und AP3: Dissemination.

AP 1: Koordination und Kommunikation

Zur *projektinternen* Kommunikation fanden regelmäßige Besprechungen auf verschiedenen Arbeitsebenen und in unterschiedlichen Formaten statt, u.a. zweiwöchentliche Entwicklertreffen, Koordinations-Reviews, eine regelmäßige Leitungsrunde, bilaterale Gesprächsformate sowie mehrere Workshops. Der Kommunikation *nach außen* dienten offene Calls für Anwender*innen und Interessierte, ein DFG-Rundgespräch mit den VD-Trägerbibliotheken und anderen Teilnehmenden der Community sowie Beiträge zu Tagungen und Veranstaltungen Dritter, wie etwa des Kitodo-Vereins, u.a.m. Die Termine wurden in der Regel per Videokonferenz durchgeführt, besondere Veranstaltungen wie u.a. das DFG-Rundgespräch fanden in Präsenz statt. Der Austausch und die Abstimmung mit den Implementierungs- und Modulprojekten erfolgte sowohl auf Bearbeiter*innen- als auch Leitungsebene mit allen verfügbaren Mitteln. Zwischenstände, Ergebnisse von Sitzungen und Entwürfe für Ausarbeitungen wurden ebenso

wie alle anderen Protokolle im Projektwiki Open Project dokumentiert. Software und Software-Dokumentation wurden (und werden auch zukünftig) auf GitHub publiziert. Drei Treffen mit dem wissenschaftlichen Beirat wurden per Videokonferenz durchgeführt. Zur Unterstützung der IP und MP und zur Förderung des Austauschs fanden insgesamt fünf Workshops statt, darunter ein Kick-Off-Treffen und ein Abschlussworkshop. Diese dienten der Identifikation von Synergien sowie der Vernetzung der Kompetenzzentren.

Der Abschlussworkshop, in dem IP, MP und das Koordinationsprojekt ihre Ergebnisse präsentierten, wurde am 24. Oktober 2024 per Videokonferenz abgehalten[2].

Im Rahmen dieses Arbeitspakets wurde außerdem die Erstellung von Ground-Truth (GT) fortgesetzt: Zur bedarfsgerechten Erweiterung der GT-Daten wurden 25.443 Seiten ergänzt, die die bestehenden 1.200 Seiten im PAGE-XML-Format aus den ersten beiden Projektphasen erweitern.[3] [4]Zusätzlich stellte die SBB weitere 403 Seiten GT bereit. Die OCR-D-GT-Guidelines[5] wurden kontinuierlich aktualisiert und erweitert, wobei eine überarbeitete Fassung in die DFG-Praxisregeln „Digitalisierung“[6] sowie in die Standardisierungsmaßnahmen von AP3 integriert wurde. Im Rahmen der Standardisierungsbemühungen wurden außerdem spezifische Guidelines entwickelt und umgesetzt. Das OCR-D-Struktur-GT-Korpus wurde um die genannten zusätzlichen Seiten erweitert, wobei unter anderem DTA-Segmentierungsdaten in das PAGE-XML-Format gemäß den OCR-D-GT-Guidelines konvertiert wurden.

AP 2: Implementierung

(1) Optimierung und bedarfsgesteuerte Weiterentwicklung des OCR-D-Frameworks

Die zentrale Software-Bibliothek OCR-D/core implementiert alle unterstützten Prozessoren und Komponenten für deren Kombination in Workflows. Während der Projektlaufzeit wurde die Software um eine Netzwerkschnittstelle erweitert, die es ermöglicht, OCR-D-Komponenten auf Basis verfügbarer Hardware zu verteilen und zu skalieren. Eine neue Major Version 3 wurde veröffentlicht, die mit wichtigen Features wie Fehlerbehandlung, seitenweiser Parallelisierung und Caching von Instanzen eine erheblich robustere und performantere Prozessierung als die Vorgängerversionen ermöglicht. Das GitHub Repository OCR-D/ocrd_all[7] bündelt OCR-D/core sowie alle relevanten Prozessoren mit Skripten, um diese in verschiedenen Kombinationen nativ oder als Docker-Container installierbar zu machen. In Phase III wurde die zwar vollständige, aber

[2] Protokoll: https://pad.gwdg.de/NYoMZNpISxWZKqGz8y5pJw#
[3] https://github.com/OCR-D/gt_structure_all
[4] https://github.com/OCR-D/gt_structure_text/
[5] https://ocr-d.de/de/gt-guidelines/trans/
[6] https://zenodo.org/records/7435724
[7] https://github.com/OCR-D/ocrd_all

sehr umfangreiche Installation in einem „fat“ Docker Container um eine leichtgewichtigere Architektur ergänzt, die kleinere und schnellere prozessorenspezifische „slim containers“ anbietet, die über die hinzugekommene ocrd_network-Schnittstelle miteinander kommunizieren und so auch eine effizientere Skalierung über Docker Compose oder Kubernetes erlauben.

Eine Spezifikation für die Integration granularer Qualitätsmetriken wurde erstellt und in Workflows implementiert. Die Standardisierungen im Bereich OCR (Training, Formate, Evaluation) wurden im steten Austausch mit der OCR-D-Community weiter vorangetrieben und dokumentiert. Die OCR-D-Spezifikationen einschließlich der Spezifikationen für PAGE-XML wurden aktualisiert und erweitert, um die Nachhaltigkeit der Software sicherzustellen.

Die GT-Daten wurden aus unterschiedlichen Quellen zusammengetragen und in dem Langzeitarchivierungssystem OLA-HD indexiert[8]. Dieser Prozess gewährleistet die Erhaltung der GT und ermöglicht eine unkomplizierte Suche über die integrierte Suchmaschine. Die GT-Daten umfassen ein Datenvolumen von 30,1 GB. Zusätzlich stehen mehr als 5,6 GB vortrainierte Modelle zur Verfügung, die unmittelbar einsatzbereit sind.

(2) Qualitätssicherung der OCR-D-Software

Repositorien und nicht mehr genutzte Code-Teile wurden in Abstimmung mit der OCR-D-Community archiviert oder entfernt. Überflüssige Kommentare und duplizierter Code wurden bereinigt, um die Lesbarkeit und Wartbarkeit zu optimieren. In einigen Bereichen erfolgte ein Refactoring, um den Code zu vereinfachen. Zudem wurden alte GitHub-Issues und Pull Requests überprüft und priorisiert oder geschlossen.

Im Rahmen der Qualitätssicherung und des Benchmarkings wurde das Quiver-Dashboard[9] entwickelt, das Metriken zur Bewertung der OCR-Leistung in verschiedenen Workflows bereitstellt. Es ermöglicht eine detaillierte Analyse der OCR-Ergebnisse durch den Vergleich von GT-Dokumenten, die eingesetzten OCR-Prozessoren sowie Messwerte wie Character Error Rate (CER), Pages per Minute (PPM) und Verarbeitungszeit.[10] Über Filter- und Anzeigefunktionen können die Daten gezielt analysiert werden. Eine Zeitleiste zeigt die Entwicklung der Metriken über mehrere OCR-D-Core-Releases hinweg, während eine Tabelle aktuelle Werte darstellt. Das Dashboard unterstützt so das Benchmarking und die Auswahl optimaler OCR-Workflows, indem es den Vergleich unterschiedlicher Materialien und Verarbeitungsstrategien ermöglicht. Derzeit enthält das Dashboard 24 Dokumente mit jeweils vier Workflows. Darüber hinaus wurde die

[8] https://ola-hd.ocr-d.de/
[9] https://github.com/OCR-D/quiver-frontend; https://github.com/OCR-D/quiver-back-end
[10] https://ocr-d.de/quiver-frontend/#/workflows

REST-API optimiert, um den Datendurchsatz zu erhöhen und eine automatische Skalierung der Prozessierungsumgebung zu ermöglichen. Die kontinuierliche Weiterentwicklung und Aktualisierung der erfassten Daten bleibt für eine langfristige Nutzung essenziell.

Um die Codequalität der Weiterentwicklungen im Berichtszeitraum auch über das Entwicklungsteam des Koordinierungsprojektes hinaus sicherzustellen, wurden Code-Review-Guidelines in GitHub erstellt, an die sich auch die IP halten müssen. Der zentrale Bestandteil des Entwicklungsworkflows sind Pull-Requests von GitHub. Die IP erstellen Pull-Requests und weisen einem oder mehreren Entwicklern des Koordinierungsprojekts die Rolle als Reviewer zu. Nach einer erfolgreichen Überprüfung kann der neue Code in die Codebasis integriert werden.

Die bis Phase II fokussierten Kommandozeilenschnittstellen für OCR-D Prozessoren boten nur begrenzte Unterstützung für Parallelisierbarkeit, da sie im Kern stetig über einzelne Seiten iterierten. Ein wichtiger Meilenstein war daher die Implementierung des METS-Servers, der den Flaschenhals des schreibenden Zugriffs auf die METS-Datei eines OCR-D-Workspace beseitigt. Weiterhin wurde die API für die Prozessoren so umgestellt, dass das Iterieren über Seiten in die Workflow-Engine von OCR-D/core verlagert wurde und damit eine seitenweise Parallelisierung und Fehlerbehandlung direkt möglich ist. Zudem ermöglicht eine Message Queue in der ocrd_network Schnittstelle nun die beliebige Skalierung der Verarbeitung ganzer Werke oder Teile davon über alle verfügbaren Prozessoren-Instanzen im Netzwerk.

Das bereits in vorangegangenen Phasen bewährte Prinzip, die OCR-D Prozessoren über eine einheitliche ocrd-tool.json-Datei maschinenlesbar zu dokumentieren, wurde in Phase III weiter ausgebaut. So wurden etwa auch die Beschreibungen der Parameter im Workflow Guide automatisch aus der ocrd-tool.json generiert. Zudem können auch Dateiressourcen, die zusätzliche Funktionalität für den Prozessor bieten, beschrieben und über den Ressource Manager von OCR-D/core nachinstalliert werden. Neben Modellen für bspw. die Texterkennung umfasst dies auch abgestimmte Sets von Parametern, die den Prozessor für bestimmte Aufgaben vorkonfigurieren. Im Rahmen der Prozessor-Registry wurden zudem formale Aspekte wie das Vorhandensein einer README-Datei, Informationen über die CI/CD eines Prozessors, sowie Statistiken über die Git-Repositories erfasst, um Konsistenz über die Projekte hinweg zu gewährleisten.

AP 3: Dissemination

(1) Offene Dokumentation als Teil der Dissemination

Die frei zugängliche Dokumentation des Projekts dient neben anderen Maßnahmen zur Verbreitung der Projektergebnisse. Ein Schwerpunkt dieser Dokumentation besteht in der

zweisprachigen Website[11] des Projekts. Die aktualisierte Installationsdokumentation steht dort ebenfalls zur Verfügung.[12] Die weitere, vor allem technische Dokumentation erfolgt auf Github.[13] Webseiten der einzelnen am Projekt beteiligten Einrichtungen unterstützen diese Form der Dissemination zusätzlich.[14]

(2) Veranstaltungen

Zur Verbreitung der Software und der Projektergebnisse dienten außerdem die folgenden regelmäßigen Veranstaltungsformate und Einzelveranstaltungen:

- OCR-D Open Tech Call (seit 2018): Zweiwöchentliche (ab 2025 monatliche) Videokonferenz mit Fokus auf die technische Entwicklung von OCR-D/core, OCR-D/ocrd_all, den Implementierungs- und Modulprojekten.
- OCR-D GT Call (2021-2024): Vierwöchentliche Videokonferenz rund um Fragen der Erstellung und Evaluation von Ground Truth.
- OCR-D in der Praxis (seit 2024): Zweiwöchentliche Videokonferenz mit Fokus auf den praktischen, produktiven Einsatz von OCR-D.
- Kick-Off Workshop, 30.07.2021
- Abschlussworkshop, 24.10.2024
- Treffen mit Projektbeteiligten im Rahmen der DHd AG OCR, virtuell bspw. Mannheim (27.-28.11.2024)

(3) Konzeptionelle Vorbereitung der Massendigitalisierung – Aktualisierung und Weiterentwicklung der Konzepte

Die Erzeugung großer Mengen qualitativ hochwertiger Volltexte eröffnet neue Möglichkeiten für die Arbeit mit digitalisierten Beständen. Das OCR-D-Projekt hat daher in dieser Förderphase die Themen Indexierung und Bereitstellung der Daten schon mit der Antragsstellung aufgenommen und verstärkt an die VD-Bibliotheken herangetragen. Eine grundlegende Maxime war dabei, die historisch gewachsene Separierung der einzelnen VDs (nach Jahrhunderten) zu überwinden und einen gemeinsamen Suchraum anzustreben. Neben den auch für die Planung der Vervolltextung wichtigen Analysearbeiten zu Bestandsgrößen bei einzelnen Bibliotheken und der Planung von

[11] https://ocr-d.de/
[12] https://ocr-d.de/en/setup
[13] https://github.com/OCR-D
[14] Zum Beispiel: https://www.sub.uni-goettingen.de/projekte-forschung/projektdetails/projekt/ocr-d-koordinierungsprojekt/ (SUB) oder https://www.bbaw.de/forschung/ocr-d (BBAW)

prozessierbaren Mengengerüsten bildete die Vorbereitung der Weiterverarbeitung einen Schwerpunkt der Aktivitäten.

Auf der Basis entsprechender Umfragen[15] und konzeptioneller Vorbereitungen veranstaltete das Koordinierungsprojekt im Sommer 2023 ein DFG-gefördertes Rundgespräch, bei dem Verantwortlichen aus den VD-Trägerbibliotheken, aus der OCR-D-Koordinierung sowie weitere Personen der Community beteiligt waren. Dabei wurden zum einen die erzielten Projektergebnisse unter quantitativen und qualitativen Gesichtspunkten diskutiert, zum anderen mögliche Szenarien für die Vervolltextung diskutiert. Der Vorschlag einer konzertierten Maßnahme, in der in einem zentralisierten Prozess bei der GWDG alle VD-Bilddigitalisate ohne vorhandene Volltexte in einem GPU-unterstützten Massenverfahren verarbeitet werden, fand Zustimmung unter der Voraussetzung, dass für spezielle Materialien, die medientypologisch und gattungssystematisch abgrenzbare Teilgruppen darstellen, bei Bedarf gesonderte, auch dezentral angelegte Volltext-Erstellungsprojekte ermöglicht werden. Bei der Veranstaltung wurde außerdem die zur Vermeidung von Redundanzen notwendige Abstimmung und Kooperation zwischen den OCR-D-verwendenden Einrichtungen und den an der Google-Kampagne beteiligten Einrichtungen, hier der Bayerischen Staatsbibliothek (BSB), im Grundsatz vereinbart. Auf dieser Basis hat die GWDG weitere Berechnungen angestellt, die die Umsetzung im Rahmen eines auf vier Jahre angelegten Projekts vorsehen.

Die Überlegungen des OCR-D-Projekts wurden als aktualisierte Konzepte dem AWBI im Herbst 2024 vorgelegt und in weiten Teilen bestätigt. Der Hinweis auf die enge Verknüpfung zwischen der Vervolltextung einerseits und der Planung des VD-Portals andererseits durch den AWBI wird aufgegriffen. Nach mehreren diskussionsintensiven Zwischenstufen hat inzwischen parallel zu den Aktivitäten des OCR-D-Konsortiums die Gruppe der verantwortlichen VD-Trägerbibliotheken die Arbeit an einem DFG-Antrag zum Aufbau einer Portallösung aufgenommen, die auf Basis der DDB-Infrastruktur entstehen soll. Diese Gruppe ist personell mit dem OCR-D-Projekt verknüpft, sodass sichergestellt ist, dass ein eng aufeinander abgestimmter und passender Antrag zur Vervolltextung aus dem Kreis des OCR-D-Konsortiums zeitlich und inhaltlich an die Planungen zum VD-Portal anschließt. Gemeinsames Ziel der VD-Bibliotheken und des OCR-D-Konsortiums ist es, in miteinander engverwobenen, parallelen Arbeitssträngen bis Mai 2025 zur Beantragung sowohl der VD-Vervolltextung als auch der Indexierung und Bereitstellung in einem VD-Portal zu kommen.

[15] https://ocr-d.de/de/umfrage.html

(4) Pflege und Weiterentwicklung der Praxisregeln

Im Berichtszeitraum wurde der zum Ende der vorhergehenden Förderphase (Phase II) unterbreitete Vorschlag für die Überarbeitung der Praxisregeln wieder aufgenommen und mit Spezifikationen aus den Bereichen Modell-Training und Metadaten ausgebaut sowie bei den o.g. Veranstaltungen zur Diskussion gestellt. Aus den Veranstaltungen wurden außerdem praxisrelevante Vorgaben und Forderungen aus der Wissenschaft aufgenommen. Da es schon zum Ende der vorhergehenden Projektphase wegen gravierender Veränderungen in der digitalen Transformation zu einem modifizierten Bedarf auch hinsichtlich der Praxisregeln gekommen war, müssen die dabei erzielten Ergebnisse nun in anderer Form eingebracht werden: Auf einem im April 2021 durchgeführten DFG-Rundgespräch zum Thema „Selbstorganisation der Praxisregeln Digitalisierung“ wurde ein neues Format der Praxisregeln beschlossen [16] . Die Fachgemeinschaften werden aufgefordert, Qualitätskriterien und Standards in Selbstorganisation weiterzuentwickeln. Dazu wird die OCR-D-Koordinierung die im Projekt entwickelten Vorschläge einbringen.

(5) Verstetigung der OCR-D-Software

Ein zentrales Ergebnis des Sunsetting- und Maintenance-Konzeptes ist die klare Strukturierung der OCR-D-Software in abgestufte Produktlevel, um Wartung und Weiterentwicklung nachhaltig zu gestalten. In Abstimmung mit der Community wurden Produktlevel definiert, die den Einsatz und die Verantwortlichkeiten klar strukturieren. Essenzielle Komponenten (Produktlevel 1a) werden aktuell vom Koordinierungsprojekt gepflegt und sollen langfristig über Entwicklungsaufträge verstetigt werden, während optionale Komponenten (Produktlevel 1b und 2) in der Verantwortung der Community bleiben. Dadurch wird eine nachhaltige Nutzung und Weiterentwicklung der Software gewährleistet. Eine detaillierte Beschreibung dieses Modells ist im Konzept zur Verstetigung der OCR-D-Ergebnisse (siehe Anlage 5.3, S. 5 Kapitel 4, Anforderungen der Verstetigung, 4.1. Definition der Software-Produktlevel) zu finden.

(6) Erarbeitung eines Konzepts für eine Betriebs- und Weiterführungsinfrastuktur

Voraussetzung für die mittel- und langfristig erfolgreiche Weiterführung- und Nutzung des OCR-D-Frameworks ist das Bestehen einer aktiven Anwendergemeinschaft, die über eine stabile, auch rechtlich belastbare Organisationsstruktur verfügt. Das Projekt hat in einem mehrstufigen Analyseprozess verschiedene formal-organisatorische Optionen geprüft und letztlich die Anbindung an den Kitodo-Verein als optimalen Lösungsweg identifiziert. Hier besteht ein

[16] https://www.dfg.de/resource/blob/175474/e39db8a25bd6f06e2a609853962e3995/rundgespraech-praxisregeln-data.pdf

erprobtes Ökosystem für die gemeinschaftliche Entwicklung und den laufenden Betrieb eines Produktions- und Präsentationsframeworks für die Retrodigitalisierung. OCR-D als Auswertungsumgebung für den Transfer digitaler Images in maschinenlesbare Texte schließt hier nahtlos an. In einer Reihe von Gesprächen mit dem Vorstand von Kitodo e.V. wurden Optionen der Zusammenarbeit und insbesondere Erwartungen an die Qualität des OCR-D-Produkts diskutiert. Neben der fachlichen Qualität der Software und der mit ihr erzielbaren Ergebnisse steht dabei vor allem die Betreib- und Wartbarkeit der Software sowie die Existenz einer tragfähigen Community im Fokus, die sicherstellt, dass OCR-D als Framework für die Vervolltextung aktiv weiterentwickelt und vor allem genutzt wird. Damit ist sichergestellt, dass die OCR-D-Community eine kritische Masse bildet, die auch ressourcenseitig ausreicht, die Weiterführung der Software abzusichern.

Vorgesehen ist, den Mitgliedern des Kitodo-Vereins die Aufnahme des OCR-D-Frameworks bzw. der im Projekt definierten Kernsoftware in die technische und organisatorische Obhut des Vereins vorzuschlagen. Vorbereitend wurde hierzu eine Begutachtung des erreichten Reifegrades der Software vorgenommen (siehe auch Anlage 5.4), ferner sind die Mechanismen zum Releasemanagement und zur Pflege der einzelnen Softwarekomponenten dokumentiert und Zuständigkeiten verteilt worden. Darüber hinaus erwartet der Kitodo-Verein ein konkret belegtes Commitment einer Reihe von Einrichtungen jenseits des OCR-D-Konsortiums sowie eine erkennbare Leistung des Konsortiums, um bereits in der Anlaufphase der Überführung in den Verein keine Risiken für den Verein zu produzieren.

(7) Öffentlich zugängliche Projektergebnisse

Wie oben bereits dargelegt wird die Öffentlichkeit seit Projektbeginn mittels Website, GitHub, Präsenz- und Online-Veranstaltungen informiert. Außerdem wurden die Projektergebnisse auch durch die folgenden Veröffentlichungen und Vorträge öffentlich zugänglich gemacht:

Publikationen mit wissenschaftlicher Qualitätssicherung:

- Antonacopoulos, Apostolos; Baierer, Konstantin; Clausner, Christian; Gerber, Mike; Neudecker, Clemens; Pletschacher, Stefan: A survey of OCR evaluation tools and metrics. In: HIP '21: Proceedings of the 6th International Workshop on Historical Document Imaging and Processing. 31.10.2021 (OA). Online: <https://doi.org/10.1145/3476887.3476888>
- Baierer, Konstantin; Büttner, Andreas; Engl, Elisabeth; Hinrichsen, Lena; Reul, Christian: OCR-D & OCR4all: Two Complementary Approaches for Improved OCR of Historical Sources. In: Proceedings of the 6th International Workshop on Computational History (HistoInformatics 2021) co-located with ACM/IEEE Joint Conference on Digital Libraries 2021 (JCDL 2021), 01.10.2021 (OA). Online: <https://ceur-ws.org/Vol-2981/>
- Baierer, Konstantin; Gerber, Mike; Labusch, Kai; Neudecker, Clemens; Rezanezhad, Vahid: Document Layout Analysis with Deep Learning and Heuristics. In: HIP '23: Proceedings of the 7th

International Workshop on Historical Document Imaging and Processing. 25.08.2023 (OA). Online: <https://doi.org/10.1145/3604951.3605513>

Weitere Publikationen und öffentlich gemachte Ergebnisse:

- Baierer, Konstantin; Büttner, Andreas; Engl, Elisabeth; Reul, Christian: Vom Bild zum Text – praktische OCR für die DH. OCR-D und OCR4all, TEI-Konvertierung. Vortrag auf der vDHd2021, 05.05.2021 (OA). Online: <https://dhd-ag-ocr.github.io/slides/OCR@vDHd-Z1.pdf>
- Baierer, Konstantin; Büttner, Andreas; Engl, Elisabeth; Kamlah, Jan: Vom Bild zum Text – praktische OCR für die DH. Evaluation, Transkription, Training. Vortrag auf der vDHd2021, 12.05.2021 (OA). Online: <https://dhd-ag-ocr.github.io/slides/OCR@vDHd-Z2.pdf>
- Engl, Elisabeth; Fink, Robert; Sachunsky, Robert; Schäfer, Robin: Vom Bild zum Text – praktische OCR für die DH. Postcorrection, Hackathon. Vortrag auf der vDHd2021, 19.05.2021 (OA). Online: <https://dhd-ag-ocr.github.io/slides/OCR@vDHd-Z3.pdf>
- Engl, Elisabeth: OCR-D: Von Prototypen zu Digitalisierungsprojekten. Vortrag auf dem Bibliothekartag 2021, Bremen, 16.06.2021 (OA). Online: <urn:nbn:de:0290-opus4-175935>
- Sachunsky, Robert; Würzner, Kay-Michael: Kollaborative Erstellung von Trainingsmaterialien für OCR – Ein Werkstattbericht. Vortrag auf dem Bibliothekartag 2021, Bremen, 17.06.2021 (OA). Online: <https://wrznr.github.io/bibliothekartag-2021/#1>
- Hertling, Anke; Klaes, Sebastian: OCR on demand: Der Ansatz eines User-generated Content. Vortrag auf dem Bibliothekartag 2021, Bremen, 17.06.2021 (OA). Online: <urn:nbn:de:0290-opus4-176621>
- Hartwig, Uwe: Open Source OCR-Systeme im Umfeld von OCR-D. 06.07.2021 (OA). Online: <https://doi.org/10.5281/zenodo.5076012>
- Engl, Elisabeth: Vom Bild zum Text – praktische OCR für die DH. Abschlussveranstaltung. Vortrag auf der vDHd2021, 15.09.2021 (OA). Online: <https://dhd-ag-ocr.github.io/slides/OCR@vDHd-Abschluss.pdf>
- Hinrichsen, Lena: Community Building und Community Management in OCR-D. Vortrag auf der #vBIB21, 01.12.2021 (OA). Online: <urn:nbn:de:0290-opus4-178109>
- Baierer, Konstantin; Boenig, Matthias; Engl, Elisabeth; Geestmann, Mareen; Hinrichsen, Lena; Neudecker, Clemens; Pestov, Paul; Weidling, Michelle: Dokument, Transkription, Forschungsdatum. Technische und kulturelle Überlegungen für interdisziplinäre Transkriptionspraxis. Kulturen des digitalen Gedächtnisses. 8. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum (DHd2022), 10.03.2022
- Neudecker, Clemens: OCR-D. DFG-funded initiative for Optical Character Recognition Development. DARIAH Workshop: Harmonizing Workflows in HTR/OCR Publication Pipelines of Textual Heritage, 15.02.2023
- Hinrichsen, Lena; Baierer, Konstantin; Neudecker, Clemens; Boenig, Matthias; Weidling, Michelle: OCR-D für die Massendigitalisierung. Projektstand und Ausblick. 111. BiblioCon, 23-26.05.2023
- Boenig, Matthias; Hartwig, Uwe; Weidling, Michelle; Baierer, Konstantin; Neudecker, Clemens: Moderne Standards bei der Erstellung und Evaluation von Volltextdaten. 111. BiblioCon, 23-26.05.2023