Beschreibung des Vorhabens - Projektanträge im Bereich „Wissenschaftliche Literaturversorgungs- und Informationssysteme“ (LIS)

LIS-Förderprogramm oder Ausschreibung:
Digitalisierung und Erschließung

VD-Volltext: Zentrale OCR-D-basierte Erschließung der Volltexte der digitalisierten Drucke aus dem deutschen Sprachraum des 16. (VD 16), 17. (VD 17) und 18. Jahrhunderts (VD 18)

Prof. Dr. Peter Burschel, Herzog August Bibliothek Wolfenbüttel (HAB)
Zeki Mustafa Dogan, Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)
Prof. Dr. Philipp Wieder, Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
Prof. Dr. Achim Bonte, Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (SBB)

Beschreibung des Vorhabens

Kapitel 1-3, insgesamt maximal 15 Seiten

1 Ausgangslage

1.1 Ausgangslage

Die Deutsche Forschungsgemeinschaft (DFG) hat in den vergangenen Jahren die Erschließung der Drucke aus dem deutschen Sprachraum des 16. (VD 16), 17. (VD 17) und 18. Jahrhunderts (VD 18) umfassend gefördert. Dadurch ist ein Korpus von mehreren hunderttausend Titeln entstanden, das auf hochwertigen bibliographischen und strukturierten Metadaten basiert und in drei zentralen Katalogen erschlossen ist. Diese Daten bilden heute die Referenzgrundlage für Recherche, Forschung und digitale Nachnutzung. Weitere von der DFG geförderte Anstrengungen dienten der Bilddigitalisierung eines Großteils der Titel dieses Korpus:

	Teilnehmende Institutionen	**A: VD-Aufnahmen / Manifestationen**	**B: Geschätzte Gesamtzahl Manifestationen (Abdeckungsgrad digitalis. Dez. 2025)**	**C: Exemplarnachweise in VD-Aufnahmen / Items**	**D: VD-Aufnahmen mit min. einem Image-Digitalisat (Manifestation)**	**E: VD-Aufnahmen mit min. einem Volltext (Manifestation)**
VD 16	345	108.680	150.000 (72 %)	483.894	73.708	mind. 45.200
VD 17	64	314.011	460.000[1] (68 %)	850.000	226.136	ca. 110.000
VD 18	30	332.926	600.000 (49 %)	[332.926][2]	307.336	ca. 75.000

Tabelle 1: Mengengerüst VD-Korpus: Manifestationen (Zahlen ermittelt im Rahmen des Projekts VD-Portal)

Aus dem Mengengerüst (Tabelle 1) geht hervor, dass für das hier beantragte Projekt die vorhandenen Manifestationen mit Bilddigitalisat und ohne Volltext relevant sind (d. h. Spalte D abzüglich Spalte E: gerundet 28.000 plus 110.000 plus 230.000 Drucke) also insgesamt 368.000 Drucke. Bei einer durchschnittlichen Anzahl von 132 Seiten (konsolidierter Durchschnitt über alle VD) ergeben sich 48.576.000 zu prozessierende Seiten.
Die antragstellenden Einrichtungen haben als eigene Vorarbeiten zusammen mit ihren Projektpartnern in der von der DFG initiierten *Koordinierten Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR-D)*[3] die technischen und organisatorischen Voraussetzungen für eine nachhaltige und stabile Infrastruktur im Bereich der automatischen Volltexterkennung geschaffen und dabei ihre Leistungsfähigkeit in diesem Arbeitsfeld nachgewiesen. HAB, SBB und SUB haben damit und außerdem mit ihrem langjährigen Engagement als VD-Trägerbibliotheken ebenso ihre Kompetenz und Eignung für das beantragte Projekt unter Beweis gestellt wie die GWDG als zentrales Rechenzentrum für die

[1] Die anfängliche Schätzung von 260.000 Manifestationen wurde bereits deutlich übertroffen, hier handelt es sich um eine aktualisierte Schätzung.

[2] Für das VD 18 wurde in der Regel nur ein Item pro Manifestation (die Vorlage für die redigierte Aufnahme) autoptisch erfasst.

[3] Dazu u.a. https://ocr-d.de/ sowie Herrmann/Stäcker (2017), Engl (2019), Engl (2020), Baierer et al. (2019a), Baierer et al (2019b),Baierer et al. (2019c), Baierer et al. (2019d), Baierer et al. (2020a), Baierer et al. (2020b),Baierer et al. (2020c), Baierer et al. (2021).

Georg-August-Universität Göttingen und die Max-Planck-Gesellschaft (u. a. mit Erfahrungen im Einsatz performanter, parallelisierter und skalierbarer Hochleistungsrechner).
Die HAB hat langjährige Erfahrungen in der Organisation und Koordination von größeren Koordinationsprojekten, zuletzt u.a. bei OCR-D. Die SUB Göttingen hat über viele Jahre fundierte Expertise in der Massendigitalisierung sowie in der Entwicklung und dem Betrieb von OCR-Verfahren aufgebaut, insbesondere durch das Göttinger Digitalisierungszentrum (GDZ)[4], über das heute mehr als 80.000 Titel mit rund 14 Millionen Seiten weltweit langfristig zugänglich gemacht werden. In mehreren DFG- und EU-geförderten Projekten (u. a. Massendigitalisierung Mathematik, IMPACT[5], OCR-D[6]) hat die SUB OCR-Verfahren für Millionen von Seiten etabliert und offene, nachnutzbare OCR-Workflows und -Services entwickelt. Zuletzt haben die SUB und die GWDG mit OPERANDI[7] ein Implementierungspaket der OCR-D-Software für die Massendigitalisierung sowie mit OLA-HD[8] ein OCR-Langzeitarchiv aufgebaut, die in diesem Vorhaben zum Einsatz kommen sollen.

In den vergangenen Jahren entstanden so neue quelloffene Werkzeuge und standardisierte Workflows zur Volltextdigitalisierung. Dazu wurden Standards entwickelt, die es erstmals ermöglichen, Volltexte für historische Drucke in großem Umfang automatisiert, reproduzierbar und qualitätsgesichert im Sinne von Open Science ohne kommerzielle Software herzustellen. Damit ist es nun möglich, den nächsten Schritt zu gehen und den bereits lange eingeforderten flächendeckenden, qualitativ belastbaren Volltextbestand der VD-Titel zu schaffen. Bislang sind die bereits existierenden Volltexte zum Teil fragmentiert, heterogen in der Qualität und oft nicht nach OCR-D-Standards aufbereitet. Wegen dieser Defizite und wegen eines bisher fehlenden gemeinsamen Nachweis- und Zugangsportals sind die bisherigen Erschließungsleistungen noch nicht wie von der Forschung gewünscht nutzbar.
Vor diesem Hintergrund soll die identifizierte Lücke nun in enger Abstimmung mit dem bei der DFG beantragten Projekt „VD-Portal“ der Deutschen Digitalen Bibliothek sowie den weiteren daran beteiligten Institutionen geschlossen werden. Bereits im Vorfeld wurden hierfür Datenflüsse konzipiert, die insbesondere die Integration der im hier beantragten Projekt entstehenden Volltexte in das VD-Portal ermöglichen und beiden Anträgen zugrunde liegen. Diese Abstimmung ist so angelegt, dass sie Synergien nutzt, ohne wechselseitige Abhängigkeiten zu begründen; beide Projekte bleiben in Zielstellung, Umsetzung und Zeitplan eigenständig und unabhängig anschlussfähig. Damit ist gewährleistet, dass Ergebnisse dieses Projektes auch in anderen Kontexten nachgenutzt werden können. Mögliche Abweichungen in der Verfügbarkeit oder Nutzung externer Portalinfrastrukturen sind im Antrag berücksichtigt und durch alternative Datenflüsse und Bereitstellungsszenarien abgesichert[9].
Die an beiden Vorhaben beteiligten Einrichtungen[10] stellen durch kontinuierliche Koordination sicher, dass inhaltliche Überschneidungen produktiv gemacht, zugleich aber Abhängigkeiten und redundante Entwicklungen vermieden werden. In dieser Konstellation eröffnet sich eine

[4] https://gdz.sub.uni-goettingen.de/
[5] https://www.digitisation.eu/
[6] https://ocr-d.de/
[7] https://www.sub.uni-goettingen.de/projekte-forschung/projektdetails/projekt/operandi-ocr-d-performance-optimisation-and-integration/
[8] https://ola-hd.ocr-d.de/
[9] Vgl. Anlage zu den Risiken.
[10] HAB Wolfenbüttel, SB Berlin, SUB Göttingen.

besondere Gelegenheit, die oben angesprochenen Lücken nachhaltig und strukturell tragfähig zu schließen – unter anderem aus den folgenden Gründen:

- Datenmengen und Metadaten der zu bearbeitenden Titel sind klar definiert und zugänglich.
- Werkzeuge, Standards und Infrastrukturen liegen aus den zurückliegenden OCR-D-Projektphasen vor und sind erprobt.
- Die OCR-D-Projekte OPERANDI (Hochleistungs-OCR-Pipeline) und OLA-HD (Plattform zur Ablage und Bereitstellung von OCR-D-Ergebnissen) stellen Infrastrukturen bereit, auf denen die OCR-Workflows in Hochleistungsumgebungen ausgeführt, die Ergebnisse zugänglich gemacht und nachhaltig archiviert werden können.
- Damit werden die Voraussetzungen geschaffen, Metadaten, Digitalisate und perspektivisch auch Volltexte der VD-Bestände zentral zusammenzuführen und gemeinsam nutzbar zu machen, wie z. B. in dem separat beantragten Projekt VD-Portal vorgesehen.

Vor diesem Hintergrund soll die Aufgabe, in einem absehbaren Zeitraum und in belastbarer Qualität Volltexte zu liefern und diese zentral zugänglich zu machen, in Angriff genommen werden. Dadurch werden nicht nur die Potenziale der bisherigen DFG-Investitionen besser ausgeschöpft, sondern auch Forschung, Text- und Data-Mining sowie wissenschaftliche Editionen und Recherchen nachhaltig unterstützt.
Zur Einordnung des Handlungsbedarfs verweisen wir auf den Stand der Katalogisierung, Digitalisierung und OCR-Erschließung für VD 16, VD 17 und VD 18 im VD-Rahmenbericht für 2023/2024.[11]

2 Ziele und Arbeitsprogramm

2.1 Voraussichtliche Gesamtdauer des Projekts

36 Monate, ein etwaiger Bedarf für einen Fortsetzungsantrag kann sich im Lauf des beantragten Projekts ergeben.

2.2 Ziele

Das Projekt verfolgt das übergeordnete Ziel, die zentralisierte Volltexterschließung und Bereitstellung der Volltexte für die digitalisierten Drucke aus dem deutschen Sprachraum des 16., 17. und 18. Jahrhunderts (VD 16, VD 17, VD 18) sicherzustellen. Das Vorgehen basiert auf dem Konzept von OCR-D zur Volltexttransformation der VD[12] und folgt den darauf bezogenen Empfehlungen des Ausschusses für Wissenschaftliche Bibliotheken und Informationssysteme (AWBI). Damit wird eine deutlich breitere und belastbare Volltextbasis geschaffen, die

- Durchsuchbarkeit und Möglichkeiten der wissenschaftlichen Auswertung der Werke wesentlich verbessert
- den Zugang zu diesen historischen Quellen erleichtert
- die Nachnutzung in Forschung, Bibliotheken und anderen Infrastrukturen unterstützt und
- die bisherigen Investitionen der DFG in Digitalisierung und OCR-D weiter in Wert setzt

[11] Vgl. Anlage VD-Rahmenbericht für 2023/2024
[12] OCR-D: Konzept zur Volltexttransformation der VD

Zur Erreichung dieses Gesamtziels setzt das Projekt zwei strategische Schwerpunkte:

1. Zentrale OCR-Erschließung

- Einsatz hochperformanter, parallelisierter Volltexterkennung auf Basis der OCR-D-Workflows und -Standards
- Nutzung der produktiven Infrastrukturen OPERANDI und OLA-HD zur effizienten, skalierbaren Verarbeitung
- Einbindung der VD-Partnerbibliotheken in standardisierte Workflows, um heterogene Ausgangsbedingungen zu berücksichtigen

2. Integrierte Bereitstellung

- Bereitstellung der generierten Volltexte über OLA-HD als zentrale Infrastruktur, mit Such-, Download- und Reprozessierungsfunktionen
- Bereitstellung der Schnittstellen für eine interoperable Anbindung, z.B. mit dem geplanten VD-Portal der Deutschen Digitalen Bibliothek, sodass Metadaten, Digitalisate und Volltexte perspektivisch aus einer Hand präsentiert werden können

Zusammen leisten diese vorgesehenen Arbeiten einen entscheidenden Beitrag zur Stärkung der nationalen Informationsinfrastruktur und schaffen eine dauerhafte Grundlage für die Forschung im Bereich historischer Drucke.
Um die unterschiedlichen Ausgangsbedingungen der VD-Bibliotheken angemessen zu berücksichtigen und gleichzeitig den Aufbau einer für den Projektablauf notwendigen, zentralen Volltextinfrastruktur sicherzustellen, adressiert das Projekt vier Szenarien. Diese umfassen das Spektrum von bereits bilddigitalisierten Beständen bis hin zu künftigen Digitalisaten und bilden den Rahmen für Priorisierung, Workflow-Design und technische Umsetzung.

Szenario A – Vollständig bilddigitalisierte Bestände ohne Volltexte (hier beantragt)

- Bilddigitalisate liegen vor, standardisierte Schnittstellen (OAI-PMH für Metadaten, IIIF für Bildzugriff) sind vorhanden
- Automatisierte und skalierbare Verarbeitung ist unmittelbar möglich
- Diese Bestände werden vorrangig bearbeitet, um schnell eine kritische Masse an Volltexten bereitzustellen

Szenario B – Bilddigitalisierte Bestände ohne Volltexte, mit unvollständigen oder fehlenden Schnittstellen (hier beantragt)

- Metadaten sind vorhanden, Bilddatenzugriff ist nur eingeschränkt möglich
- Entwicklung ergänzender Verfahren (z. B. alternative Datenwege, abgestimmte statische Dumps) und Anpassung der Workflows)
- Mittelfristige Integration in den zentralen Volltextprozess

Szenario C – Vorhandene, aber qualitativ heterogene Volltexte[13]

- OCR-Volltexte sind bereits vorhanden, jedoch in unterschiedlicher Qualität und Konformität
- Stichprobenartige Prüfung und ggf. Reprozessierung zur Vereinheitlichung und Qualitätsverbesserung
- Nachrangige Bearbeitung, um Ressourcen auf die Volltexterschließung bisher unerschlossener Bestände zu konzentrieren

Szenario D – Künftig zu digitalisierende Bestände (ggf. einschließlich weitere in Planung befindliche VD-bezogene Digitalisierungsvorhaben)[14]

- Für Bestände, die erst in Zukunft katalogisiert oder (bild-)digitalisiert werden, wird ein Konzept zur späteren Volltexterschließung entwickelt
- Aufbau eines dezentralen OCR-as-a-Service-Verfahrens (Self-Service-Schnittstellen, definierte Upload-Pakete, standardisierte Workflows), um Bibliotheken zu befähigen, neu digitalisierte Werke eigenständig oder koordiniert Volltext-erschließen zu können

Mehrwert der Szenarien

Diese vier Szenarien decken alle relevanten Eventualitäten ab und sorgen dafür, dass

- bestehende Bestände schnell erschlossen
- heterogene Datenquellen integriert
- die Qualität der Daten vereinheitlicht
- Redundanzen vermieden werden und
- zukünftige Digitalisate planbar eingebunden werden können

Damit entsteht eine skalierbare, zukunftssichere Infrastruktur. Die beantragte Projektlaufzeit von 36 Monaten ermöglicht es, die Szenarien A und B umzusetzen und die notwendigen Vorbereitungen für die Szenarien C und D zu treffen sowie diese bei vorhandenen Kapazitäten und geeigneter Datenlage testweise umzusetzen.

2.3 Arbeitsprogramm und Umsetzung

Das Projekt VD-Volltext konzentriert sich auf die großvolumige, standardisierte Volltexterschließung der in den VD-Projekten digitalisierten Drucke des 16., 17. und 18. Jahrhunderts. Es setzt auf den in der OCR-D-Förderinitiative entwickelten, quelloffenen

[13] Die im Rahmen der Google-Kooperation der BSB München und ihrer Partnereinrichtungen erstellten Volltexte sowie die Volltexte der SBB Berlin, ULB Halle und der SLUB Dresden werden von der hier geplanten Bearbeitung in den Szenarien A und B bewusst ausgeschlossen, um redundante Aufwände zu vermeiden. Im Rahmen des Projekts „VD-Portal“ ist vorgesehen, die VD-relevanten Daten dieser Einrichtungen separat in das VD-Portal einzubringen.

[14] Beispielsweise befindet sich ein umfangreicher von der Forschung initiierter Antrag zur Digitalisierung von Periodika des VD-Zeitraums in Vorbereitung ("VD-Periodika"), an dem 15 Institutionen beteiligt sein werden. Geplant sind ca. 900.000 bilddigitalisierte Seiten, die z.T. noch im Antragszeitraum prozessiert werden könnten. Darüber hinaus plant die SBB in Zusammenarbeit mit Partnereinrichtungen aus Polen ein Projekt, um VD-relevanten Bestand polnischer Bibliotheken zu digitalisieren und mit OCR zu erschließen ("VD Polen") und arbeitet in einem Projekt an der Erschließung der Bestände der Bestände in den Bibliotheken der Evangelischen Kirche in Mitteldeutschland ("EKM-Projekt").

Verfahren und Infrastrukturen auf und überführt diese in einen produktiven Betrieb. Dabei werden insbesondere die Systeme OPERANDI und OLA-HD genutzt und projektspezifisch erweitert.[15]
Die zu erfüllenden Aufgaben sind in fünf Arbeitspakete gegliedert:

- AP 1: Projektkoordination und Schnittstellenmanagement
- AP 2: Datenmanagement und technische Vorbereitung
- AP 3: Volltextdigitalisierung der VD-Bestände
- AP 4: Bereitstellung und Integration
- AP 5: Qualitätssicherung und Optimierung

AP 1 und 2 legen die organisatorische und technische Grundlage. AP 3 und 4 bilden den operativen Kern des Projekts: Sie erzeugen, speichern und verknüpfen die Volltexte. AP 5 gewährleistet die Qualitätssicherung und Dokumentation. Die enge Verzahnung dieser Arbeitspakete sichert einen durchgängigen Workflow von der Metadatenerfassung bis zur Bereitstellung der Volltexte.

1. Die beantragte Laufzeit von zunächst 36 Monaten ist erforderlich, um die großvolumige OCR-Verarbeitung, Qualitätssicherung und Bereitstellung der Volltexte aus den Szenarien A und B für die nationalen Infrastrukturen (Quellsysteme und Portale) nachhaltig umzusetzen. Sie ergibt sich aus den teilweise komplementären Faktoren Verarbeitungsvolumen, Kosten für die Bereitstellung der technischen Kapazität und der organisatorischen Kapazität auf Seiten der teilnehmenden Institutionen und des Projektes. Die Zahl der zu bearbeitenden Seiten beläuft sich auf 48.576.000 Seiten.[16] Deren serielle Bearbeitung würde gemäß der durchgeführten Benchmarks 184,8 Jahre benötigen, was durch die in OPERANDI realisierte Parallelisierung im Rahmen des vorliegenden Antrags entsprechend verkürzt wird. Dabei wurden die Faktoren (i) Kosten für die HPC-Infrastruktur, die mit steigender Parallelisierung entsprechend wachsen, (ii) organisatorische Nutzung, insbesondere bzgl. Anlaufzeiten und Unterbrechungen, sowie (iii) Qualitätssicherung (s. AP 5) gegen die Projektlaufzeit abgewogen.
2. Iterative Qualitätssicherung und Reprozessierung: Neben der reinen OCR-Verarbeitung ist eine mehrstufige Qualitätssicherung vorgesehen. Ergebnisse werden stichprobenartig geprüft, mit Fehlerbildern annotiert und bei Bedarf erneut verarbeitet. Dieses iterative Vorgehen gewährleistet, dass die Qualität der Volltexte über den gesamten Projektzeitraum und für alle vorgesehenen Vorlagen möglichst optimal sichergestellt wird.
3. Koordination und Integration in Infrastrukturen: Die enge Abstimmung mit VD-Bibliotheken, GWDG und Präsentationsstrukturen wie dem separat beantragten VD-Portal erfordern sukzessive Implementierungs-, Test- und Integrationsphasen. Der Aufbau standardisierter Schnittstellen (OAI-PMH, IIIF), die Harmonisierung heterogener Datenquellen und die Rückführung der Ergebnisse in OLA-HD und VD-Portal sind komplexe Arbeitsschritte, die nur in mehrjährigen Zyklen sicher umsetzbar sind.

Die Weiterentwicklung der Systeme und Workflows erfolgt parallel zur OCR-Prozessierung. Durch die Nutzung der modernen CI/CD-Infrastruktur und der HPC-Umgebung der GWDG können Entwicklungen, Tests und Deployments parallel zur laufenden Produktion erfolgen, ohne die Prozessierung zu unterbrechen. Diese Parallelität von Entwicklung und Produktion gewährleistet eine effiziente Nutzung der Projektlaufzeit und reduziert das Risiko von Stillstandszeiten.

[15] Dazu Abschlussbericht, siehe https://ocr-d.de/Abschlussbericht_OCR-D_%C3%B6ffentlich.pdf
[16] Siehe Tabelle 1 oben, S. 1.

Die Qualitätssicherung (im Folgenden QS) (AP 5) wirkt als Querschnittsaufgabe über den gesamten Projektverlauf und beeinflusst insbesondere die Arbeitspakete zur Datenverarbeitung und Integration (AP 3 und 4). Durch die kontinuierliche Auswertung und Rückkopplung der QS-Ergebnisse können erkannte Fehler frühzeitig korrigiert und Workflows angepasst werden, etwa durch Verbesserung der Layouterkennung durch Einsatz der an der SBB entwickelten Software *eynollah*.[17] Reprozessierungen werden automatisiert ausgelöst, sodass die Produktionspipeline nicht unterbrochen wird. Etwaige Verzögerungen durch Qualitätsanalysen werden durch die Parallelisierung von QS und Verarbeitung sowie durch geplante Zeitpuffer kompensiert. Auf diese Weise trägt die QS zur Gesamtstabilität und zur stetigen Verbesserung der Infrastruktur bei, ohne den Zeitplan wesentlich zu beeinflussen.
Die Laufzeit von 36 Monaten erlaubt es, die Projektphasen Pilotierung, Produktion, Konsolidierung und Integration für die Szenarien A und B systematisch aufzubauen und durchzuführen. Sie entspricht der Größenordnung einer nationalen Infrastrukturmaßnahme und reflektiert die in OPERANDI gewonnenen Erfahrungswerte[18].

2.3.1 Technisches Konzept und Systemarchitektur

Die technische Architektur des Projekts folgt einem modularen, interoperablen Ansatz. Daten werden über einen OCR-Service (vgl. AP 2) konsolidiert, der OAI-PMH-, IIIF- und METS-Schnittstellen standardisiert bereitstellt. Diese Plattform dient zugleich als Austauschschicht zwischen den VD-Bibliotheken, OPERANDI, OLA-HD und dem VD-Portal. Sie ermöglicht sowohl den Import der bibliographischen Daten und der Adressen von METS-Dateien bzw. iiif-Manifesten aus den VD-Quellen inkl. VD-Portal als auch die Rückgabe der erzeugten Volltexte, Adressen und Metadaten an die VD-Bibliotheken und das VD-Portal, um eine durchgängige Interoperabilität sicherzustellen.
Die Volltexterzeugung erfolgt über OPERANDI, das als skalierbare Ausführungsumgebung für OCR-D-Workflows auf der HPC-Infrastruktur der GWDG betrieben wird. Im Projekt wird OPERANDI für die großvolumige Verarbeitung heterogener VD-Digitalisate angepasst und mit automatisierten Prüf- und Monitoringverfahren ausgestattet.
Die Bereitstellung der Ergebnisse erfolgt über OLA-HD, das als technische Infrastruktur und Backend für Speicherung, Versionierung und Nachnutzung dient. OLA-HD bietet zudem eine Weboberfläche, über die Partnerbibliotheken ihre OCR-Ergebnisse prüfen und reprozessieren können. Über standardisierte APIs wird OLA-HD mit dem VD-Portal verbunden, um die Volltexte nahtlos in die bibliographische Suche zu integrieren. Abb. 1 veranschaulicht den grundlegenden Datenfluss zwischen den beteiligten Komponenten. Die Metadaten und Digitalisate werden von den VD-Bibliotheken und dem VD-Portal über standardisierte Schnittstellen in den OCR-Service überführt. Von dort erfolgt die Übergabe an OPERANDI zur OCR-Verarbeitung und an OLA-HD zur Speicherung und Bereitstellung der Volltexte. Das VD-Portal übernimmt die Präsentation der Ergebnisse und gewährleistet den übergreifenden Zugang.
Die Qualitätssicherung erfolgt integriert in den OPERANDI-Workflows und ergänzend durch stichprobenartige manuelle Kontrollen (vgl. AP 5). Damit wird eine reproduzierbare und transparente Datenqualität über alle Verarbeitungsstufen hinweg gewährleistet.

[17] Vgl. Baierer et al. (2023) und https://github.com/qurator-spk/eynollah
[18] vgl. OPERANDI-Abschlussbericht (Anhang).

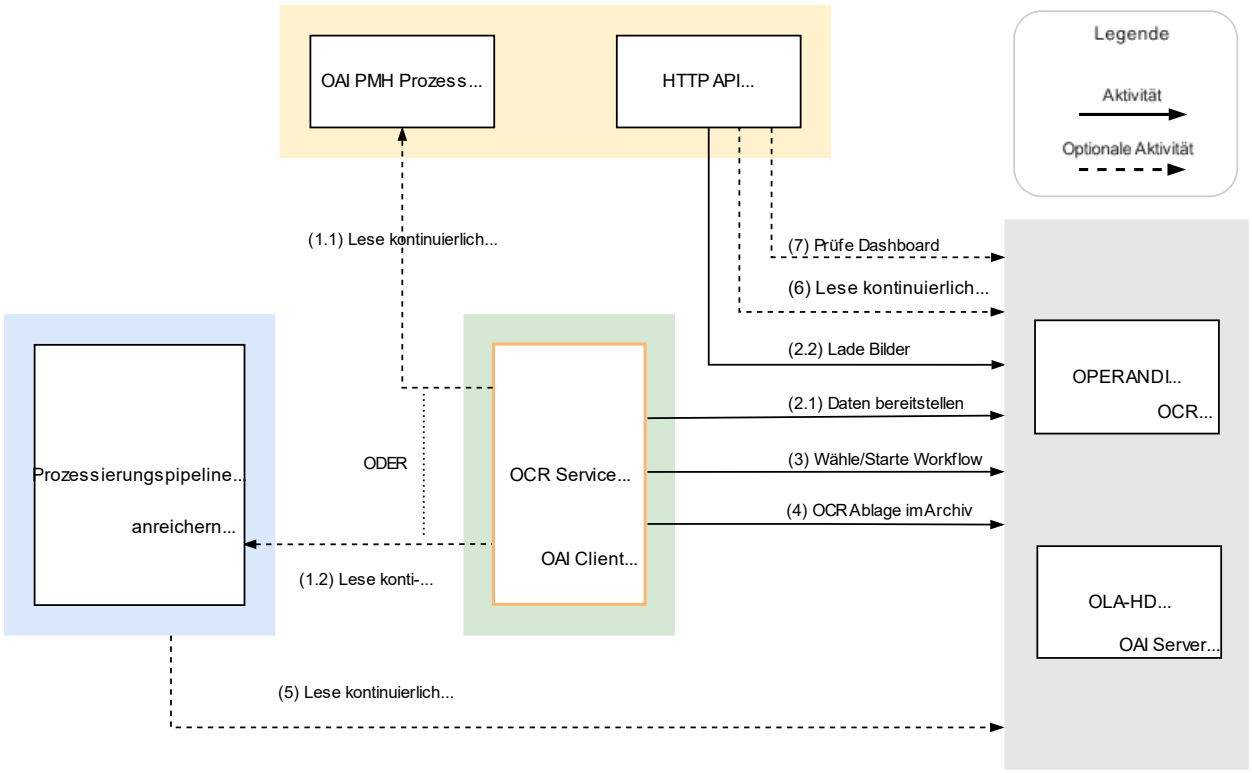

Abb. 1: Datenbeschaffung und Verarbeitung

2.3.2 Arbeitspakete

AP 1: Projektmanagement und Kommunikation (SUB 18 PM, HAB 18 PM)

Dieses Arbeitspaket umfasst die Gesamtsteuerung und organisatorische Koordination des Projekts. Dazu gehören die Abstimmung mit den beteiligten VD-Bibliotheken, dem VD-Portal-Projekt (Bewilligung vorausgesetzt) sowie die Einhaltung der in den Arbeitspaketen definierten Schnittstellen und Standards. Regelmäßige Projekttreffen dienen der Synchronisierung der Arbeitspakete und der Qualitätssicherung des Gesamtprozesses. Die Projektkoordination verantwortet zudem das Berichtswesen, die Kommunikation mit der DFG und die Dissemination der Projektergebnisse.

Die Projektkoordination erfolgt auf verschiedenen Ebenen: Die HAB übernimmt dabei die Gesamtkoordination mit strategischer Verantwortung, einschließlich der Kommunikation mit Fördergebern, des Stakeholder-Managements sowie der übergreifenden Planung. Zudem begleitet bzw. koordiniert die HAB die notwendigen Abstimmungsprozesse zur Datenbeschaffung und -bereitstellung in enger Zusammenarbeit mit dem separat beantragten Projekt VD-Portal und den VD-Bibliotheken. Dazu sind monatliche Online-Treffen mit den genannten Einrichtungen vorgesehen. Im Bewilligungsfall beginnen diese organisatorischen und koordinatorischen Arbeiten aus Eigenmitteln der HAB vor der technischen Bearbeitung in den weiteren AP. Die technische Koordination liegt bei der SUB, die ein agiles Projektmanagement nach Scrum-Prinzipien einsetzt und insbesondere die Steuerung der technischen Implementierung und Prozessierung übernimmt. Dabei übernimmt der Certified Product Owner (PO) der SUB die fachliche Leitung und detaillierte Ausarbeitung des Requirements Engineerings für technische Fragestellungen, wie Schnittstellendefinitionen, Datenkonvertierungsprozesse, die Spezifikation

technischer Workflows sowie das fortlaufende Refinement der Szenarien im Projektverlauf in enger Abstimmung mit den AP 2 und 5.

Teilaufgaben

- A 1.1: Projektorganisation, Einrichten der Workflows (HAB)
- A 1.2: Etablierung kontinuierlicher Kommunikationsformate und Organisation von drei Workshops (HAB)
- A 1.3: Stakeholder-Kommunikation und externe Sichtbarkeit (SUB, HAB):
- A 1.4: Technische Koordination (agil, nach Scrum) (SUB)
- A.1.5: Technisches Requirementsengineering (SUB)
- A 1.6: Fortschrittsüberwachung und stichprobenartige Qualitätskontrolle (HAB)

Meilensteine

- MS 1 (Monat 3) – Projektorganisation und Kommunikationsstrukturen vollständig etabliert
- MS 2 (Monat 6) – Kick-Off-Workshop „Datenbereitstellung & Datenrückfluss“ durchgeführt
- MS 3 (Monat 30) – Workshop „Qualitätssicherungsmethoden“ zur Vorbereitung von Szenario C
- MS 4 (Monat 36) – Abschlussworkshop „Fehleranalyse & materialspezifische Workflows“ durchgeführt, Abschlussbericht erstellt und Projektergebnisse überführt

AP 2: Datenmanagement und technische Vorbereitung (HAB 18 PM, SUB 18 PM)

Ziel dieses Arbeitspakets ist die Einrichtung eines zentralen OCR-Services als Austausch- und Harmonisierungsschicht sowie die Etablierung der Abläufe zwischen den beteiligten Systemen und Institutionen.
Auf dieser Grundlage werden für jedes Werk vollständige Verarbeitungspakete erzeugt, die sowohl die Identifikation des Titels (z. B. VD-Nummer), die Seitenstruktur und Reihenfolge als auch die Referenzen auf die zugehörigen Bilddateien enthalten. Diese Pakete bilden die Eingabe für die OCR-Verarbeitung in AP 3.
Die Datenakquise erfolgt nach einem vierstufigen Verfahren, das vorhandene Ressourcen nutzt und den unterschiedlichen technischen Voraussetzungen der VD-Bibliotheken Rechnung trägt:

1. Nutzung vorhandener Daten des VD-Portal-Projekts: Zunächst wird geprüft, ob die erforderlichen bibliographischen und strukturellen Metadaten bereits im Rahmen des VD-Portal-Projekts verfügbar sind. Ist dies der Fall, werden die Daten direkt übernommen, um Doppelverarbeitung zu vermeiden und eine konsistente Datengrundlage sicherzustellen
2. Standardverfahren – Harvesting über Schnittstellen: Automatisiertes Harvesting der Daten über die vorhandenen Schnittstellen der VD-Bibliotheken, insbesondere OAI-PMH, IIIF oder andere Endpunkte, die den Zugriff auf METS-Dateien oder Bildmanifeste ermöglichen
3. Selbstbereitstellung durch Bibliotheken: Alternativ können die Bibliotheken ihre Metadaten und Strukturdaten über eine definierte API direkt an OLA-HD übermitteln. Dazu reicht die Angabe einer METS-Sammlung oder IIIF-Manifest-URI für eine Kollektion oder einen Titel, die zugehörigen Bilddateien werden anschließend automatisiert eingespielt
4. Physischer Datentransfer: In begründeten Ausnahmefällen (z. B. sehr große Datenmengen, fehlende technische Kapazitäten oder Know-how) können Datenträger mit den erforderlichen Digitalisaten bereitgestellt werden. Der Import erfolgt in diesem Fall über standardisierte

Skripte, sodass auch offline übermittelte Daten vollständig in den OCR-Service integriert werden

Der Arbeitsschwerpunkt der HAB liegt in der Kommunikation mit den VD-Bibliotheken und in der Organisation des Datenaustauschs. Dazu gehört auch die Koordination, falls es beim Harvesting zu Abbrüchen oder Unregelmäßigkeiten kommt. Die HAB übernimmt somit vor allem die organisatorischen und kommunikativen Aufgaben. Die SUB Göttingen bringt sowohl die technische Expertise bezüglich der Spezifikation der Schnittstellen und Datenkonvertierungen ein als auch ihr umfassendes Know-how zu bibliothekarischen Datenanforderungen, Metadatenstandards und Interoperabilitätsfragen. Sie unterstützt die Konzeption der Datenflüsse im Hinblick auf langfristige Nachnutzbarkeit und Integration in bestehende Bibliotheksinfrastrukturen.

Teilaufgaben

- A 2.1 Aufbau der des OCR-Service und Definition der Datenaustauschprozesse (SUB), Kommunikation mit datenliefernden Einrichtungen (HAB)
- A 2.2 Prüfung und Integration der vom VD-Portal bereitgestellten Daten (SUB, HAB)
- A 2.3 Implementierung und Test der Harvesting-Komponente für OAI-PMH, IIIF und METS (SUB)
- A 2.4 Entwicklung und Dokumentation der OLA-HD-API für den Upload durch Bibliotheken (SUB, HAB) sowie Vermittlung an datenliefernde Einrichtungen (HAB)
- A 2.5 Implementierung des Skript-basierten Imports für physische Datenträger (SUB), Organisation: Übermittlung Datenträger (HAB)
- A 2.6 Bildung verarbeitbarer OCR-Pakete: Metadaten, Strukturdaten, Bildreferenzen (SUB)
- A 2.7 Einrichtung von Logging, Monitoring und Validierungsroutinen für alle Importpfade (SUB)

Meilensteine

- MS 5 (Monat 6): OCR-Service betriebsbereit, Harvesting-Schnittstellen aktiv
- MS 6 (Monat 9): Datenübernahme aus VD-Portal erfolgreich getestet
- MS 7 (Monat 12): Vollständiger Datenimport für erste VD-Kollektionen abgeschlossen
- MS 8 (Monat 15): API-basierter Upload und automatischer Bildimport über OLA-HD produktiv

AP 3: Volltextdigitalisierung der VD-Bestände (SUB 18 PM, SBB 6 PM, GWDG 18 PM)

Dieses Arbeitspaket bildet den zentralen Kern der Volltexterschließung der VD-Bestände. Es fokussiert die automatisierte Texterkennung der bereits digitalisierten Werke auf Basis der im OCR-D-Programm entwickelten Werkzeuge, Workflows und Erkennungsmodelle. Die Verarbeitung erfolgt großvolumig und standardisiert auf der HPC-Infrastruktur der GWDG.
Im Vordergrund steht die robuste, skalierbare OCR-Prozessierung der aus AP 2 bereitgestellten Datenpakete (METS/MODS mit Bildreferenzen). Eine gezielte Entwicklung oder Feinjustierung von OCR-Modellen für Sondermaterialien ist nicht Gegenstand dieses Arbeitspakets, sondern die massenhafte OCR-Prozessierung. Entsprechende Erweiterungen können bei Bedarf in Zusammenarbeit von VD-Bibliotheken geplant und nach Evaluierung im Rahmen von AP 5 in die produktiven Workflows integriert werden.

Die Umsetzung erfolgt in einem iterativen, agilen Prozess. SUB und GWDG stimmen kontinuierlich geeignete OCR-Workflows ab, konfigurieren und evaluieren sie stichprobenartig. Die SBB bringt ihre Erfahrungen aus der produktiven Anwendung der OCR-D Software auf die eigenen VD-Bestände ein, sodass die Erreichung einer optimalen Ergebnisqualität gewährleistet ist. Die Qualitätssicherung ist Bestandteil des Workflows: automatisierte OCR-Metriken und Rückmeldungen aus AP 5 fließen unmittelbar in die Auswahl und Optimierung der Workflows ein. Die enge Abstimmung mit AP 2 gewährleistet die standardisierte Übergabe der Ergebnisse an OLA-HD (AP 4). Organisatorische Unterstützung erfolgt durch AP 1 (Projektmanagement).

Teilaufgaben

- A 3.1 Durchführung der OCR-Prozessierung und Fortschrittsüberwachung
  - Harvesting, Einrichtung von Workspaces, Workflow-Ausführung, Ergebnisarchivierung
  - Monitoring und Fehlerbehandlung auf der HPC-Infrastruktur
- A 3.2 Stichprobenartige Qualitätskontrolle gemeinsam mit AP 5 und Dokumentation der Ergebnisse (SBB)
- A 3.4 Pflege, Qualitätssicherung und Versionsmanagement der OCR-D-Softwarekomponenten (SBB)
- A 3.5 Administration und Pflege der OPERANDI- und HPC-Komponenten
- A 3.6 Sicherstellung reibungsloser Datenflüsse zu AP 2 (OCR-Service) und AP 4 (OLA-HD)

Meilensteine

- MS 9 (Monat 3): Start der OCR-Prozessierung für Szenario A (Digitalisate ohne Volltexte, mit IIIF/OAI)
- MS 10 (Monat 12): Integration Szenario B (Digitalisate ohne IIIF, aber mit OAI) und Beginn der stichprobenartigen QS
- MS 11 (Monat 15): Stabilisierung und Dokumentation der eingesetzten OCR-D-Versionen (SBB) für den produktiven Masseneinsatz
- MS 12 (Monat 24): 70 % der Zielmenge verarbeitet; Vorbereitung Szenario C (heterogene Volltexte)
- MS 13 (Monat 30): Abgleich der OCR-D-Softwarepflege mit aktuellen Standardisierungs- und QS-Ergebnissen aus AP 5
- MS 14 (Monat 36): Szenarien A und B abgeschlossen; Konzeptentwicklung Szenario D (zukünftige Digitalisate), Integration mit OLA-HD, VD-Portal und AP 5 (QS-Rückkopplung)

Perspektivisch für etwaige Beantragung weiterer Projektlaufzeit:

- Szenario D produktiv; Vollständige Integration mit OLA-HD, VD-Portal und AP 5 (QS-Rückkopplung) abgeschlossen

AP 4: Bereitstellung und Integration (SUB 18 PM, GWDG 18 PM)

Dieses Arbeitspaket umfasst die Ablage, Verwaltung und Bereitstellung der erzeugten Volltexte. OLA-HD wird als Backend-Infrastruktur ausgebaut und bietet Funktionen zur Speicherung, Versionierung und Nachnutzung der Daten. Ein zentrales Dashboard dient als Steuerungs- und Monitoring-Werkzeug für den gesamten Verarbeitungsprozess. Es visualisiert den Fortschritt der OCR-Prozessierung, zeigt Qualitätsmetriken an und ermöglicht den VD-Bibliotheken die Einsicht

in den Status ihrer Titel sowie die Anforderung von Reprozessierungen. Über standardisierte APIs werden die Volltexte an das VD-Portal übermittelt und dort mit den bibliographischen Metadaten verknüpft. Auf diese Weise bleibt das VD-Portal die Präsentations- und Discovery-Ebene, während OLA-HD die technische Grundlage für Verwaltung, Bereitstellung und interne Prozesssteuerung bildet.
Die SUB Göttingen arbeitet in enger Kooperation mit der GWDG an der Implementierung und fortlaufenden Optimierung der zentralen Speicher- und Schnittstellenmechanismen. Schwerpunktmäßig übernimmt die SUB Göttingen die Implementierung der Schnittstellen für den OCR-Service sowie die erforderlichen Entwicklungsarbeiten im Frontend-Bereich. Die GWDG ist hingegen für die Entwicklungsarbeiten im Zusammenhang mit den infrastrukturellen Anpassungen und der technischen Integration der Systeme verantwortlich.

Teilaufgaben

- A 4.1 Erweiterung von OLA-HD für Massendaten und Versionierung (SUB, GWDG)
- A 4.2 Entwicklung des Dashboards für Monitoring und Nachvollziehbarkeit (SUB)
- A 4.3 Implementierung der Schnittstellen OPERANDI → OLA-HD → VD-Portal (SUB, GWDG)
- A 4.4 Integration von Statistik- und Reporting-Funktionen (GWDG)

Meilensteine

- MS 15 (Monat 18): Dashboard v1 verfügbar
- MS 16 (Monat 36): API-Integration mit VD-Portal produktiv, Vollständige Integration und Bereitstellung der Szenarien A und B abgeschlossen

Perspektivisch für etwaige Beantragung weiterer Projektlaufzeit:

- Vollständige Integration und Bereitstellung der Daten nach den Szenarien C und D abgeschlossen

AP 5: Qualitätssicherung (QS) und -optimierung (SBB 18 PM)

Dieses Arbeitspaket stellt sicher, dass die erzeugten OCR-Ergebnisse den Anforderungen an Belastbarkeit, Nachvollziehbarkeit und Nachnutzbarkeit genügen. Es umfasst die Entwicklung, Implementierung und Integration von Metriken und Werkzeugen zur automatisierten und manuellen QS als auch die fortlaufende Optimierung der Workflows auf Basis der im Projekt gewonnenen Erkenntnisse.
Die QS erfolgt iterativ und datengetrieben. Grundlage ist ein Konzept für eine automatisierbare, skalierbare QS, das sowohl Ground-Truth-basierte als auch Ground-Truth-freie Verfahren berücksichtigt.[19] Während eine auf Ground-Truth-gestützte Bewertung nur stichprobenartig durchgeführt werden kann, werden ergänzend Konfidenzwerte der OCR-Engines herangezogen, um die Qualität großer Datenmengen effizient einzuschätzen.
Jüngere Untersuchungen zeigen, dass zwischen OCR-Konfidenzen und tatsächlicher Zeichen- oder Wortfehlerquote unter geeigneten Bedingungen eine belastbare Korrelation bestehen

[19] Vgl. Baierer et al. (2025) und Antonacopoulos et al. (2021)

kann.[20] Daher werden im Projekt Korrelationstests auf repräsentativen Stichproben durchgeführt, um die verwendeten OCR-Modelle und Workflows zu kalibrieren. Dabei kann auf Erfahrungen aus der OCR-D-Förderinitiative zurückgegriffen werden, in der insbesondere die SBB umfangreiche Evaluierungen der OCR-D-Software und ihrer Metriken vorgenommen hat.[21]
Erst auf dieser Grundlage werden Konfidenzmetriken für großvolumige Auswertungen eingesetzt. Sie ermöglichen eine kontinuierliche Überwachung der Erkennungsqualität, ohne die Ground-Truth-basierten Prüfungen vollständig zu ersetzen. Das Zusammenspiel beider Verfahren erlaubt eine skalierbare und zugleich verlässliche Qualitätsbewertung: Ground-Truth-Stichproben sichern die empirische Belastbarkeit, während automatisierte Konfidenzmetriken eine laufende Beobachtung und adaptive Optimierung der OCR-Workflows ermöglichen.
Die QS ist integraler Bestandteil der Produktions-Workflows. Sie erfolgt in enger Abstimmung mit der Synchronisierung und Überwachung der Arbeitspakete sowie der Qualitätskontrolle des Gesamtprozesses (AP1). OPERANDI erzeugt Konfidenzmessungen im Rahmen der OCR-Verarbeitung und Fehlerprotokolle im Falle von Abbrüchen, die in AP 5 ausgewertet und für Optimierung und Reprozessierung genutzt werden. Die Ergebnisse werden über standardisierte Schnittstellen in OLA-HD bereitgestellt und dienen den VD-Bibliotheken zur Nachverfolgung und Dokumentation der OCR-Qualität.
Ergänzend werden Dokumentation, Schulungen und Workshops angeboten, um die Projektpartner bei der Interpretation der Ergebnisse und der Anwendung geeigneter Workflows zu unterstützen. Neue wissenschaftliche Verfahren zur OCR-Evaluation werden kontinuierlich beobachtet und – sofern relevant – in die bestehenden Methoden integriert.

Teilaufgaben

- A 5.1 Entwicklung und Implementierung geeigneter Metriken für skalierbare Qualitätssicherung (SBB)
- A 5.2 Kombination von Ground-Truth-basierten und Ground-Truth-freien Verfahren (OCR-Konfidenzen, lexikalische Verfahren) (SBB)
- A 5.3 Aggregation und Visualisierung der Qualitätsdaten im Dashboard von OLA-HD (SBB)
- A 5.4 Durchführung von Fehleranalysen, Ableitung von Optimierungsempfehlungen (SBB)
- A 5.5 Organisation von Online-Workshops und Austauschformaten zu QS-Methoden und materialspezifischen Workflows (SBB)
- A 5.6 Pflege und Weiterentwicklung der Dokumentation gemäß den DFG-Praxisregeln Digitalisierung (SBB)

Meilensteine

- MS 17 (Monat 9): QS-Werkzeuge implementiert und in OPERANDI integriert
- MS 18 (Monat 12): Aggregation und Visualisierung der Qualitätsdaten über OLA-HD produktiv
- MS 19 (Monat 18): Erweiterte Verfahren zur automatisierten Evaluation implementiert
- MS 20 (Monat 36): Qualität, Optimierung und Nachhaltigkeit dokumentiert

[20] Vgl. Cuper et al. (2023)
[21] Seit dem Abschluss der Phase 3 des OCR-D-Koordinierungsprojekts wird die OCR-D-Software produktiv für die Verarbeitung mehrerer Millionen Seiten historischer Drucke eingesetzt; die Erfahrungen aus der Evaluation dieser Verarbeitung wurde bei der Computational Humanities Research Conference vorgestellt, vgl. Bubula et al. 2025.

Perspektivisch für etwaige Beantragung weiterer Projektlaufzeit:

- Nachnutzbarer Werkstattbericht und Praxisrichtlinien zu Qualität, Optimierung und Nachhaltigkeit publiziert

2.3.3 Rollen im Projekt

- **Projektkoordination:** Projektmanagement, Beiträge zur Qualitätssicherung, Konzeption und Organisation von Workshops, Stakeholder-Management, zentrale Informationsverteilung, Abstimmung mit Partnereinrichtungen
- **Technische*r Product Owner*in:** Technische Koordination, Scrum-Management, Requirements Engineering, Abstimmung Schnittstellen & Workflows
- **Datenmanagement:** Organisation der Datenprozesse, Monitoring, Fortschrittskontrolle, Abstimmung mit VD-Bibliotheken
- **Metadaten/Datenhandling:** Verarbeitung & Integration digitalisierter Daten, Schnittstellenspezifikation, Verwaltung Metadaten, Unterstützung QS
- **HPC-Entwickler*in:** Technische Ausführung der OCR-Workflows auf HPC, Integration der OPERANDI-Komponenten, OLA-HD Speicherinfrastruktur, Unterstützung bei Infrastrukturproblemen, Monitoring automatisierter Prozesse
- **OCR-Service-Entwickler*in:** Implementierung von OCR Service & OLA-HD Schnittstellen, Orchestrierung der Workflows, Integration von QS-Maßnahmen, Weiterentwicklung des Dashboards
- **QS-Expert*in:** Implementierung Metriken, QS-Prozesse, Workshops, Evaluation OCR-Workflows

2.4 Relevanz von Geschlecht und/oder Diversität im Forschungsvorhaben

Das Projektvorhaben hat sich zum Ziel gesetzt, allen Beteiligten - unabhängig von Geschlecht und anderen persönlichen Eigenschaften - gleiche Chancen zu eröffnen. Daher wird auf eine ausgewogene Zusammensetzung der Projektarbeitsgruppe geachtet. Gleichstellung und Diversität werden bei der Personalauswahl systematisch berücksichtigt.
Auch wenn das Vorhaben keine eigene Fragestellung zu Geschlecht oder Diversität verfolgt, so gibt es doch projektbezogene Aspekte, die in diesem Zusammenhang von Interesse sind. Erschlossene Werke werden maschinenlesbar und dadurch Menschen mit Sehbehinderungen wesentlich einfacher zugänglich, indem sie z. B. mittels eines Screenreaders gelesen werden können. Barrieren bei der Durchsuchung und Weiterverarbeitung der Texte werden reduziert. Sie können einfacher auf verschiedenen Geräten dargestellt und visuell angepasst werden.

3 Projekt- und themenbezogenes Literaturverzeichnis

(1) Antonacopoulos et al. (2021): 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. Online: < https://doi.org/10.1145/3476887.3476888>

(2) Bubula et al. (2025): Bubula, Michał, Baierer, Konstantin, Jörg Lehmann, Clemens Neudecker, Vahid Rezanezhad, und Doris Škarić: How Scalable is Quality Assessment of Text Recognition? A Combination of

Ground Truth and Confidence Scores. In: Anthology of Computers and the Humanities 3 (2025): S. 1286–1310. https://doi.org/10.63744/GR59c1iXu6Wj

(3) Baierer et al. (2023): 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. Online: < https://doi.org/10.1145/3604951.3605513>

(4) Baierer et al. (2021): 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. Online: <http://ceur-ws.org/Vol-2981/>

(5) Baierer et al. (2020a): Baierer, Konstantin; Boenig, Matthias; Engl, Elisabeth; Hartmann, Volker; Neudecker, Clemens: Volltexte – die Zukunft alter Drucke. Bericht zum Abschlussworkshop des OCR-D-Projekts, in: o-bib 7 (2), 05.05.2020, S. 1–4. Online: <https://doi.org/10.5282/o-bib/5600>

(6) Baierer et al. (2020b): Baierer, Konstantin; Boenig, Matthias; Engl, Elisabeth; Hartmann, Volker; Neudecker, Clemens: Volltexttransformation frühneuzeitlicher Drucke – Ergebnisse und Perspektiven des OCR-D-Projekts, in: DHd 2020: Spielräme - Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts, Paderborn 05.03.2020, S. 244–247. Online: <https://doi.org/10.5281/zenodo.3666690>

(7) Baierer et al. (2020c): Baierer, Konstantin; Boenig, Matthias; Engl, Elisabeth; Neudecker, Clemens; Hartmann, Volker: Volltexte für die Frühe Neuzeit. Der Beitrag des OCR-D-Projekts zur Volltexterkennung frühneuzeitlicher Drucke, in: Zeitschrift für Historische Forschung 47 (2), 2020, S. 223–250

Baierer et al. (2019a): Baierer, Konstantin; Dong, Rui; Neudecker, Clemens: okralact – a multi-engine Open Source OCR training system, in: Proceedings of the 5th International Workshop on Historical Document Imaging and Processing, Sydney 20.09.2019, S. 25–30. Online: <https://dl.acm.org/doi/10.1145/3352631.3352638>

Baierer et al. (2019b): Neudecker, Clemens; Baierer, Konstantin; Federbusch, Maria; Würzner, Kay-Michael; Boenig, Matthias; Herrmann, Elisa; Hartmann, Volker: OCR-D: An end-to-end open source OCR framework for historical documents, in: EuropeanaTech Insight (13), 31.07.2019. Online: <https://pro.europeana.eu/page/issue-13-ocr#ocr-d-an-end-to-end-open-source-ocr-framework-for-historical-documents>

Baierer et al. (2019c): Boenig, Matthias; Baierer, Konstantin; Hartmann, Volker; Federbusch, Maria; Neudecker, Clemens: Labelling OCR Ground Truth for Usage in Repositories, in: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, Brüssel 09.05.2019, S. 3–8. Online: <https://dl.acm.org/doi/10.1145/3322905.3322916>

Baierer et al. (2019d): Neudecker, Clemens; Baierer, Konstantin; Federbusch, Maria; Würzner, Kay-Michael; Boenig, Matthias; Herrmann, Elisa; Hartmann, Volker: OCR-D: An end-to-end open-source OCR framework for historical documents, in: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, Brüssel 09.05.2019, S. 53–58. Online: <https://dl.acm.org/doi/10.1145/3322905.3322917>

Cuper et al. (2023): Cuper, Mirjam; Dongen, Corine van; Koster, Tineke: Unraveling Confidence: Examining Confidence Scores as Proxy for OCR Quality. In: Proceedings of the 17th International Conference on Document Analysis and Recognition (ICDAR). Cham: Springer Nature Switzerland, 2023, S. 104–120. https://doi.org/10.1007/978-3-031-41734-4_7

(8) Engl (2020): Engl, Elisabeth: OCR-D kompakt: Ergebnisse und Stand der Forschung in der Förderinitiative, in: Bibliothek Forschung und Praxis 44 (2), 29.07.2020, S. 218–230. Online: <https://doi.org/10.1515/bfp-2020-0024>

(9) Engl (2019): Engl, Elisabeth: Das Projekt OCR-D – Ein Fortschrittsbericht zur Volltextdigitalisierung frühneuzeitlicher Drucke, in: Medium Buch 1, 2019, S. 233–235

(10) Herrmann/Stäcker (2017): Herrmann, Elisa; Stäcker, Thomas; : OCR-D – Koordinierte Förderinitiative zur Weiterentwicklung von OCR-Verfahren, in: Bibliotheksdienst 52 (1), 05.12.2017. Online: <https://doi.org/10.1515/bd-2018-0007>

Begleitinformationen zum Projektkontext

4.1 Allgemeine ethische Aspekte

Im Rahmen der Durchführung des geplanten Projekts sind in ethischer Hinsicht keine Risiken, Belastungen oder andere negative Auswirkungen für Personen bzw. Personengruppen zu erwarten.

4.2 Reflexion zu ökologischen Nachhaltigkeitsaspekten in der Planung und Durchführung des Vorhabens

Um den ökologischen Fußabdruck des Projekts gering zu halten, vermeidet die Projektgruppe Dienstreisen (insbesondere per Flugzeug), wenn sie nicht dringend erforderlich sind. Regelmäßige Projektbesprechungen finden bevorzugt virtuell statt, so dass trotz der geographischen Verteilung der Beteiligten nur geringe ökologische Kosten entstehen. Auch Software- und Datenstrategien sind möglichst ressourcenschonend angelegt. So werden zur Volltextdigitalisierung (AP 3) Infrastrukturen und Services der GWDG verwendet, welche ein hochmodernes Rechenzentrum nachhaltig betreibt (u.a. Nutzung von Energie aus regenerativen Quellen, Nachnutzung der Abwärme und Warmwasserkühlung) . Auch werden die Daten effizient gespeichert (vornehmlich AP 4), um zusätzlichen Speicher- und Energieverbrauch zu vermeiden. Es kommen offene Standards und Open-Source-Lösungen zum Einsatz, die modular aufgebaut und wiederverwendbar sind. Damit wird eine langfristige Nachnutzbarkeit der digitalen Projektinfrastruktur sichergestellt und unnötige Doppelentwicklungen werden vermieden.

4.3 Maßnahmen zur Erfüllung der Förderbedingungen und Umgang mit den Projektergebnissen

Die GWDG, die HAB, die SBB und SUB sind in der Lage und verbindlich dazu bereit, die Projektergebnisse zu verstetigen und deren Nachhaltigkeit zu sichern. Sie verpflichten sich, alle im Merkblatt 12_19 beschriebenen Anforderungen einzuhalten. Ebenso verpflichten sie sich eine Software- und Benutzungsdokumentation anzufertigen. Alle durch das Projekt zustande gekommenen Ergebnisse werden der Fachöffentlichkeit bekannt gemacht und zur kostenlosen Nachnutzung für die Wissenschaft zur Verfügung gestellt. Die umfassende Dokumentation und die produzierten Quellcodes werden in GitHub bzw. GitLab mit der Bereitstellung der Projektergebnisse offengelegt (als Open Source). Die Projektpartner haben auch weiterhin den Anspruch, verlässlichen Partner in der OCR-D-Community zu bleiben und stabile Strukturen vorzuhalten, wie sie es seit zehn Jahren mit und ohne DFG-Förderung sind.

4.4 Erklärungen zur Erfüllung der Förderbedingungen

Die Antragsteller versichern, dass die Voraussetzungen für die Förderung vorliegen und die finanziellen Eigenleistungen eingehalten werden. Die Ergebnisse des Vorhabens, insbesondere die zu erarbeitenden Richtlinien und Standards, werden entsprechend nachhaltig bereitgestellt.

5 Personen/Kooperationen/Finanzierung

5.1 Angaben zur Dienststellung

Für jede Antragstellerin und jeden Antragsteller, unter Angabe von Name, Vorname, Dienststellung (bei befristetem Arbeitsvertrag Angaben zur Laufzeit und ggf. zum Zuwendungsgeber)

Prof. Dr. Peter Burschel, Direktor HAB

Zeki Mustafa Dogan, Abteilungsleiter „Digitale Bibliothek“, SUB

Prof. Dr. Philipp Wieder, Stellvertretender Leiter GWDG

Prof. Dr. Achim Bonte, Generaldirektor SBB

5.2 Zusammensetzung der Projektarbeitsgruppe

Dr. Johannes Mangei, Stv. Direktor, Abteilungsleitung Neuere Medien, Digitale Bibliothek, HAB

Dr. Hartmut Beyer, Abteilungsleitung Alte Drucke, HAB

Ingo Pfennigstorf, Gruppenleitung Software und Service-Entwicklung, SUB

Kristine Schima-Voigt, Stv. Gruppenleitung Software und Service-Entwicklung, SUB

Greta Richter, Wissensmanagement, Abteilung Digitale Bibliothek, SUB

Jörg-Holger Panzer, Senior-Entwickler für OCR-Technologien, Datenprozessierung und Bibliotheksdatenmanagement, SUB

Paul Pestov, Senior-Entwickler für Frontendentwicklung, SUB

Rolf Röper, Gruppenleiter Retrodigitalisierung, SUB

Johannes Biermann, Experte für High-Performance Computing in den Digital Humanities, GWDG

Björn Braunschweig, Service Owner und Senior Developer Archivierung, GWDG

Dr. Christian Köhler, Experte High-Performance Computing, GWDG

Ralph Krimmel, Experte verteilter Servicebetrieb, GWDG

Clemens Neudecker, Referatsleiter Data Science, Abteilung Informations- und Datenmanagement der Staatsbibliothek zu Berlin, SBB

Konstantin Baierer, Wissenschaftlicher Mitarbeiter, Referat Data Science, Abteilung Informations- und Datenmanagement, SBB

5.3 Institutionen oder Wissenschaftler*innen in Deutschland, mit denen für dieses Vorhaben eine konkrete Vereinbarung besteht

-

5.4 Institutionen oder Wissenschaftler*innen im Ausland, mit denen für dieses Vorhaben eine konkrete Vereinbarung besteht

-

5.5 Institutionen, Wissenschaftler*innen, mit denen in den letzten drei Jahren gemeinsame Projekte durchgeführt wurden

SBB: Aus den Kooperationsbeziehungen, die die SBB im Rahmen ihrer vielfältigen Projektaktivitäten unterhält, seien lediglich die folgenden Partnereinrichtungen der vergangenen drei Jahre auf den Feldern von Retrodigitalisierung und Volltextgenerierung via OCR herausgegriffen:

Berlin-Brandenburgischen Akademie der Wissenschaften | Deutsches Forschungszentrum für Künstliche Intelligenz | Europeana Foundation | Forschungsbibliothek Gotha | Fraunhofer Institut für Offene Kommunikationssysteme | Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen | Herzog August Bibliothek Wolfenbüttel | Karlsruher Institut für Technologie | Leopold-Franzens-Universität Innsbruck: Forschungszentrum Digitalisierung und Elektronische Archivierung | Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden | Universitäts- und Landesbibliothek Sachsen-Anhalt | Universitätsbibliothek Bielefeld | Universitätsbibliothek Braunschweig | Universitätsbibliothek Leipzig

SUB, HAB, GWDG: Listen der Kooperationspartner im Anhang

5.6 Projektrelevante Zusammenarbeit mit erwerbswirtschaftlichen Unternehmen

Trifft für dieses Vorhaben nicht zu.

5.7 Projektrelevante Beteiligungen an erwerbswirtschaftlichen Unternehmen

Trifft für dieses Vorhaben nicht zu.

5.6 Weitere Antragstellungen

Der Antrag auf Finanzierung dieses Vorhabens wurde bei keiner anderen Stelle eingereicht. Wenn die Antragsteller einen solchen Antrag stellen, werden sie der Deutschen Forschungsgemeinschaft unverzüglich benachrichtigen. Der DFG-Vertrauensdozent der Georg-August-Universität Göttingen, Prof. Dr. Martin Suhm, wurde von dieser Antragstellung unterrichtet.

5.9 **Eigenleistung** (Ist-Stand der Personalkosten zzgl. 3% Personalkostensteigerung)

SUB:

Name	**Anzahl Personenmonate**	**Personal-kategorie**	**Arbeitspakete**	**Summe der Kosten**
Pfennigstorf	2	E13	1	14.700 €
Schima-Voigt	4,5	E13	1	33.075 €
Richter	2	E13	1	14.700 €
Röper	14	E8	2	70.905 €
Panzer	9	E13	3	66.150 €
Pestov	9	E13	4	66.150 €
Zwischensumme				265.680 €
Eigenleistung der SUB inkl. 3% Personalkostensteigerung:				**273.620 €**

HAB:

Name	Anzahl Personenmonate	Personalkategorie	Arbeitspakete	Summe der Kosten
Mangei	4,5	A16	1 und 2	38.220 €
Beyer	3	A14	1 und 2	15.810 €
Baumgarten	1	E14	1 und 2	8.530 €
Busse-Hagen	1	E13	1 und 2	8.050 €
N.N.	36	WHK, 86h/Monat BA	1 und 2	61.200 €
Zwischensumme				131.810 €
Eigenleistung der HAB inkl. 3% Personalkostensteigerung:				**135.760 €**

GWDG:

Name **Anzahl**

	Personenmonate	Personalkategorie	Arbeitspakete	Summe der Kosten
Biermann	4	E13	2 und 4	32.373 €
Braun-schweig	3	E13	4	26.442 €
Köhler	2	E13	3	17.628 €
Krimmel	2	E14	3 und 4	18.740 €
Philipp Wieder	1,5	E15	3 und 4	17.250 €
N.N.	30	WHK, 40h/Monat	3 und 4	22.304 €
Zwischen-summe				134.737 €
Inkl. 3% Personal-kosten-steigerung				138.779 €
Infrastruktur	./.	./.	alle	62.000
Zwischensumme				196.737 €
Eigenleistung der GWDG:				**200.779 €**

SBB:

Name	Anzahl Personenmonate	Personalkategorie	Arbeitspakete	Summe der Kosten
Clemens Neudecker	6	E13	3 und 5	44.100 €
Konstantin Baierer	6	E13	3 und 5	44.100 €
Zwischensumme				88.200 €
Eigenleistung der SBB inkl. 3% Personalkostensteigerung:				**90.840 €**

Beitrag zur Weiterentwicklung und Verstetigung von OCR-D

Bis heute engagieren sich die beteiligten Einrichtungen aktiv in der Community-Arbeit rund um OCR-D. Dazu gehört insbesondere die Fortführung des Austauschformats „OCR-D Forum“ sowie die regelmäßigen Online-Angebote „Open Tech-Call“ und „OCR-D in der Praxis“, die den fachlichen Dialog und Wissenstransfer zwischen Anwender*innen, Entwickler*innen und Infrastrukturpartner*innen fördern. Darüber hinaus wird die OCR-D-Software im Rahmen der

verfügbaren Ressourcen weiter gepflegt und stabilisiert, um eine reibungslose Übergabe an den Kitodo-Verein sicherzustellen und die Nachhaltigkeit der bisherigen Entwicklungen zu gewährleisten.

Apparative Ausstattung

Für Entwicklung und Tests wird in diesem Vorhaben neben virtuellen Maschinen und Speicher auch Rechenzeit auf HPC-Ressourcen benötigt. Die GWDG ist als Betreiber solcher Infrastrukturen in der Lage, sämtliche gemäß gegenwärtiger Planung benötigte Ressourcen kostenfrei zur Verfügung zu stellen. Dies gilt auch für einen Zeitraum bis zu einem Jahr nach Ende der Förderung. Diese Ressourcen sind zudem über das Campusnetzwerk und das Netz des DFN zugänglich und zudem in die administrativen und technischen Prozesse des Rechenzentrums integriert. Sofern Bedarf an entsprechender Expertise zur Integration der Ressourcen in das Vorhaben besteht, welche von den ProjektmitarbeiterInnen nicht erfüllt werden kann, stellt die GWDG diese ebenfalls in Eigenleistung zur Verfügung. Weitere, hier nicht explizit aufgeführte Punkte wie Softwarelizenzen, unter 5.1.2 nicht aufgeführte Sachmittel oder Räume für Workshops werden den Partnern nach Bedarf zur Verfügung gestellt.

Personelle Ausstattung

Alle unter 5.2 angegebenen Projektmitarbeitenden werden aus eigenen Mitteln finanziert. Der Personalaufwand für Tätigkeiten, die dem regulären Geschäftsgang der Bibliotheken und Rechenzentren entsprechen, wird in Eigenleistung der SUB und GWDG bestritten. Der Personalaufwand für die regelhaften Tätigkeiten der administrativen Leitung und die Zusammenarbeit der Projektpartner wird in Eigenleistung bestritten.

6 Beantragte Module/Mittel

6.1 Mittel für Personal

Insgesamt werden 1.234.800,00 € für Personal beantragt. Dabei werden die Personalmittelsätze der DFG für das Jahr 2026 für wissenschaftliche Mitarbeiter*innen (DFG-Vordruck 60.12 – 01/26) zugrunde gelegt.
Für die Laufzeit von 36 Monaten wird eine verkleinerte Koordinierungsstruktur für das OCR-D-Projekt vorgeschlagen, die eine effektive und nachhaltige Umsetzung der Projektziele gewährleistet. Die Teamstruktur setzt sich aus folgenden Rollen und Personalkapazitäten zusammen:

AP	Rolle	Personal-kategorie	Aufgaben im AP	FTE	Partner	PM pro AP	Gesamtvergütung in €
AP 1: Projekt-koordination & Kommunikation	Projekt-koordination	Postdoc oder vergl.	Gesamt-steuerung, organisato-rische Koordina-tion	1,0	HAB	18	132.300
	Technische*r Product Owner*in	Postdoc oder vergl.	Technische Koordina-tion, Scrum-Manage-	1,0	SUB	18	132.300

			ment, Requirements Engineering, Abstimmung Schnittstellen & Workflows				
AP 2: Datenmanagement & technische Vorbereitung	Datenmanagement	Postdoc oder vergl.	Mitwirkung bei der Datenakquise	1,0	HAB	18	132.300
	Metadaten/ Datenhandling	Postdoc oder vergl.	Verarbeitung & Integration digitalisierter Daten, Schnittstellenspezifikation, Verwaltung Metadaten, Unterstützung QS	1,0	SUB	18	132.300
AP 3: Volltextdigitalisierung der VD-Bestände	HPC-Entwickler*in	Postdoc oder vergl.	Monitoring und Fehlerbehandlung auf der HPC-Infrastruktur, Administration & Pflege der OPERANDI- & HPC-Komponenten	1,0	GWDG	18	132.300
	OCR-Service-Entwickler*in	Postdoc oder vergl.	Konfiguration & Evaluation der Workflows	1,0	SUB	18	132.300
	QS-Expert*in	Post doc oder vergl.	Stichprobenartige Qualitätskontrolle, Ergebnisdokumentation	1,0	SBB	6	44.100
AP 4: Bereitstellung & Integration	HPC-Entwickler*in	Postdoc oder vergl.	Schnittstellen-Implementierung, Integration von Statistik- & Reporting-Funktionen	1,0	GWDG	18	132.300

	OCR—Service-Entwickler*in	Postdoc oder vergl.	Implementierung des OCR Service & OLA-HD Schnittstellen, Orchestrierung der Workflows, Integration von QS-Maßnahmen, Dashboard-Weiterentwicklung	1,0	SUB	18	132.300
AP 5: Qualitätssicherung	QS-Expert*in	Postdoc oder vergl.	QS-Metriken & Verfahren; Qualitätsaggregation & Dashboard; Fehleranalyse & Wissenstransfer	1,0	SBB	18	132.300

6.2 Sachmittel

Mittel für Reisen

Neben den regelmäßigen Online-Besprechungen sollen drei Präsenzveranstaltungen zu Projektstart, im Projektverlauf und als Abschlussveranstaltung (Kick-off, Midterm-Review und Abschluss-Workshop, AP 1) zum Projekterfolg beitragen. Dabei sollen – abgesehen von der gastgebenden Einrichtung – pro beteiligter Institution drei Personen anreisen und je einmal übernachten können. Für die Durchführung eines Kick-off-Workshops in Berlin, des Midterm-Reviews in Wolfenbüttel und des Abschlussworkshops in Göttingen werden daher die folgenden Sachmittel für Reisen beantragt:

AP	Benennung des Bedarfs	GWDG	HAB	SBB	SUB
1	Kick-off	900[22]	900	-	900
1	Workshop 2	900	-	900	900
1	Workshop 3	-	900	900	-
Summe		1800	1800	1800	1800

[22] 300 Euro Kostenerstattung für je drei interne Teilnehmende für Bahnfahrt 2. Klasse und 1 Übernachtung. Gastgebende Einrichtung ohne Kostenerstattung (-).

Mittel für Workshops

Außer den Teilnehmenden aus dem Kreis der antragstellenden Einrichtungen sollen bei den drei Präsenzveranstaltungen jeweils auch externe Expert*innen eingeladen werden. Deren Fahrt- und Übernachtungskosten sollen aus Projektmitteln übernommen werden. Die HAB übernimmt die Abrechnung der Kosten mit den externen Teilnehmenden und beantragt im Rahmen dieses Antrags daher die folgenden Sachmittel für Workshops:

AP	Benennung des Bedarfs	GWDG	HAB	SBB	SUB
1	Kick-off	-	2.400[23]	-	-
1	Workshop 2	-	2.400	-	-
1	Workshop 3	-	2.400	-	-
Summe		-	7.200	-	-

Mittel für Volltexterschließung mittels High-Performance Computing

Wie in den Kapiteln 1 und 2.3 dargelegt, sollen im Rahmen von Szenario A und Szenario B 48.576.000 Seiten digitalisiert werden. Die Kosten pro Seite belaufen sich hierbei auf 0,0026 Euro pro Seite (auf vier Nachkommstellen gerundet). Dies bildet die Vollkosten für die optimierte, parallelisierte Volltexterschließung mittels der OCR-D Workflows ab und inkludiert sämtliche Kostenfaktoren, wie z.B. Nutzung der HPC-Ressourcen, anteilige Lizenzkosten, Energiekosten, etc. In der Summe werden für diesen Posten **Sachmittel in Höhe von 126.975 Euro** beantragt.

Zusammenfassung der beantragten Mittel nach Modulen

Kategorie	GWDG	HAB	SBB	SUB	Zeilensumme
Personalmittel	264.600	264.600	176.400	529.200	1.234.800
Sachmittel	128.775	9.000	1.800	1.800	141.375
Summe Antragsteller	393.375	273.600	178.200	531.000	1.376.175
Gesamtsumme					**1.376.175**

Anlagen:

1. VD-Rahmenbericht 2023/2024
2. OPERANDI-Abschlussbericht
3. Gannt-Diagramm zu Meilensteinen von VD-Volltext
4. Kooperationspartner*innen der letzten drei Jahre (HAB, GWDG, SUB)
5. Risikoanalyse
6. CV Prof. Dr. Achim Bonte
7. CV Prof. Dr. Peter Burschel
8. CV Zeki Mustafa Dogan
9. CV Prof. Dr. Philipp Wieder
10. Formular 12.141 zur Einhaltung von Zusagen SUB Göttingen
11. Formular 12.141 zur Einhaltung von Zusagen GWDG

[23] 300 Euro Kostenerstattung für acht externe Teilnehmende für Bahnfahrt 2. Klasse und 1 Übernachtung. HAB rechnet die Reise- und Übernachtungskosten mit den externen Teilnehmenden ab.