HERZOG
AUGUST
BIBLIO
THEK

Herzog August Bibliothek Wolfenbüttel
Forschungs- und Studienstätte für europäische Kulturgeschichte

Prof. Dr. Peter Burschel
Direktor

Lessingplatz 1
D-38304 Wolfenbüttel

Telefon (05331) 808-100
Fax (05331) 808-134
E-Mail direktor@hab.de

Georg-August-Universität Göttingen
Seminar für Mittlere und Neuere Geschichte
Professur für Kulturgeschichte des Mittelalters und der Frühen Neuzeit

Heinrich-Düker-Weg 14
D-37073 Göttingen

Absichtserklärung zur OCR-D-Koordinierung

Antragsteller:

- Herzog August Bibliothek Wolfenbüttel (HAB)
- Berlin-Brandenburgische Akademie der Wissenschaften in Berlin (BBAW)
- Staatsbibliothek zu Berlin - Preußischer Kulturbesitz (SBB)
- Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)
- Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)

Vorarbeiten:

Die beteiligten Antragsteller sind größtenteils bereits seit der ersten Projektphase im Rahmen des Koordinierungsprojekts oder eines Modulprojekts in der OCR-D-Förderinitiative engagiert. In der ersten Projektphase haben die Antragsteller u.a. das OCR-D-Funktionsmodell erarbeitet, das als Basis für Entwicklung und Anpassung der OCR-D-Prozessoren in der zweiten Phase diente. Um die Integration der einzelnen Prozessoren zu ermöglichen, wurden Spezifikationen entwickelt und in der Referenzimplementierung core umgesetzt. Die Antragsteller verfügen über umfassende Kenntnisse der Ergebnisse der Modulprojekte, die für die nun folgende Phase von zentraler Bedeutung sind. Dieses Wissen reicht vom erzielten Entwicklungsstand der einzelnen Modulprojekte selbst (funktional und softwaretechnisch) bis hin zu den Anforderungen bei der Anbindung von Modulen und Ergebnissen. Unter den Partnern, aber auch mit den Modulprojekten, besteht eine gute, aktive Arbeitsbeziehung, die – verbunden mit den laufenden Tests, dem Austausch über die Ergebnisse sowie deren Einarbeitung in die Software – eine optimale Grundlage für eine erfolgreiche Implementierungsphase bietet. Viele Anforderungen, die als elementare Voraussetzungen für eine erfolgreiche Implementierungsphase formuliert wurden, hat das Koordinierungsprojekt im Verlauf der ersten und zweiten Projektphase bereits umgesetzt. Dazu gehören bspw. die Konsolidierung und Überarbeitung aller vorhandenen Dokumentation zur OCR-D-Software und den einzelnen Prozessoren in klar nach Anforderungen von Anwendern, Entwicklern sowie Empfehlungen für Workflows strukturierte Anleitungen, die auf der OCR-D Webseite veröffentlicht sind.

Projektziele:

Bereits in der Vorbereitung der Pilotierung zeigt sich, dass die Implementierungen intensiver Abstimmungen und Koordinierung bedürfen, um die in der zweiten Projektphase prototypisch entwickelte OCR-D-Software erfolgreich in Form generischer Implementierungspakete für einen breiten Anwenderkreis nutzbar machen zu können. Entstehende Fragestellungen sowie Probleme müssen gesammelt, analysiert, bewertet und Handlungsvorschläge erarbeitet werden, die ggf. Änderungen an der Software nach sich ziehen. Generalisierbare Erfahrungen sind zu dokumentieren, Handbücher für die Einsatzpraxis sind abzustimmen. Gleichzeitig sind – neben den Erkenntnissen zur Einsatzoptimierung – Hinweise zu funktionalen Defiziten, aber auch zu Features, die verändert oder neu geschaffen werden müssen, zu sammeln. Diese sind auszuwerten und als change requests zu formulieren. Die Software der Modulprojekte muss mit angemessenem Aufwand und mit Blick auf die vorhandenen Ressourcen gepflegt und gewartet werden.

Eine weitere Aufgabe liegt im Aufbau einer Community von Nutzer*innen und Entwickler*innen, die all diejenigen umfasst, die OCR-D selbst einsetzen oder mit den Ergebnissen und ihrer Qualitätssicherung umgehen, so u.a. die DHd-Arbeitsgruppe OCR oder die Anwenderschaft der thematisch verwandten Projekte OCR4all und TRANSKRIBUS.

Das Koordinationsgremium wird die folgenden Schwerpunkte bearbeiten:

Projektkoordination und Öffentlichkeitsarbeit *(Schwerpunkt HAB)*

- Organisation der Kommunikation sowohl innerhalb des Koordinierungsprojekts als auch mit den Implementierungsprojekten sowie ggf. fortgeführten Modulprojekten (Bedarfsfall)
- Vorbereitung und Durchführung von Implementierungs-Workshops. Diese dienen dem intensiven Austausch der an der Förderinitiative beteiligten Projekte über den jeweiligen Projektstand und die Herausforderungen der einzelnen Implementierungsprojekte. Außerdem sollen mögliche Synergieeffekte frühzeitig erkannt und sinnvoll genutzt werden.
- Projektmanagement zur Sicherung des erfolgreichen Projektverlaufs
- Dissemination der Projektergebnisse an die interessierte Fachöffentlichkeit. U.a. in Vorträgen, Publikationen und auf der Projektwebsite wird über den aktuellen Stand des Projekts berichtet und dieses dokumentiert. Die technische Entwicklungsarbeit kann transparent auf GitHub und im gitter-Chat mitverfolgt werden.
- Kontinuierlicher Austausch mit den VD, um deren geplante Volltexttransformation weiter vorzubereiten. Das im Rahmen der zweiten Projektphase entwickelte erste Konzept zur Volltextdigitalisierung der VD kann mit den Erkenntnissen und Ergebnissen der Implementierungsphase aktualisiert und weiter präzisiert werden.

Betreuung der Implementierung *(Schwerpunkt SBB, GWDG und SUB)*

- Prüfung und Weiterentwicklung der aufgaben- und nutzungsbezogenen Mensch-Maschine-schnittstellen der OCR-D-Software. Die Entwicklung und Optimierung von Schnittstellen für die OCR-D-Software ist auf die Hauptaufgabe der Massenvolltextdigitalisierung ausgerichtet. Angestrebt wird ein hoher Automatisierungsgrad, der nur in Ausnahmefällen für wenige Prozessschritte des Workflows manueller Interventionen über ein User Interface bedarf.
- Zusammenführung von Erfahrungen der Anwender zu einer zentralen Dokumentation der OCR-D-Software. Die Dokumentation ist dabei auf die Bedürfnisse der verschiedenen Anwendergruppen auszurichten.
- Betreuung der technischen Integration und praxisbezogene Dokumentation einer einheitlichen Trainingsinfrastruktur, in der verschiedene Einrichtungen zentral neue Modelle für die von OCR-D genutzten Engines trainieren können. Darüber hinaus sollen die implementierenden Einrichtungen beim Training ihrer Modelle mit Low-Level-Support und Beratung unterstützt werden.

Prüfung und Weiterentwicklung der Software *(Schwerpunkt SBB, GWDG und SUB)*

- (automatisches) Benchmarking
- Registry für Prozessoren inkl. Benchmarkergebnissen
- Ermittlung der Robustheit eines Prozessors/Workflows
- Empfehlung eines Workflows anhand von wenigen Beispielseiten
- Dokumentation der Prozessoren und ihrer Interaktionsmöglichkeiten

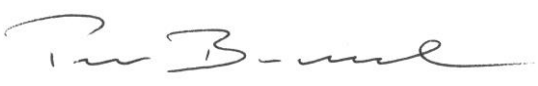
- Weiterentwicklung der Software-Lösung (aktuell: Taverna bzw. Makefile-basiert) zur Verkettung der einzelnen OCR-D-Prozessoren zu Workflows. Mit Blick auf den geplanten Einsatz der OCR-D-Software in der Massenverarbeitung kommt der robusten, teils automatisierten Workflow-Erstellung eine besondere Bedeutung zu.
- Vorbereitung der technischen Integration der Qualitätsanalyse in die OCR-D-Software
- In enger Zusammenarbeit mit den Entwicklern in den OCR-D-Projekten wird an der Realisierung einer bedarfsgerechten Erstellung von Ground Truth vor allem für den Bereich der Layout-Erkennung auf Basis von vorhandenen Textressourcen (u.a. Deutsches Textarchiv) sowie synthetisch erstellten Dokumenten gearbeitet.
- Weiterentwicklung, Pflege und Hosting der OCR-D-Repositorien für GT und Forschungsdaten. Die im Verlauf der Pilotierungs- und Implementierungsphase entstehenden Daten müssen in die öffentlich zugänglichen Repositorien eingepflegt und deren Funktionalität an die Anforderungen der zu erwartenden größeren Datenmengen angepasst werden.
- Erarbeitung und Umsetzung eines (dauerhaften) Maintenance- und Sunsetting-Konzeptes für die OCR-D-Software. Dafür ist auch ein Konzept für eine dauerhafte Support-Struktur zur Wartung und Pflege der Ergebnisse der Modulprojekte zu erarbeiten und umzusetzen. Um die OCR-D-Software zu professionalisieren, ist im Rahmen der Implementierungsphase die Optimierung der OCR-D-Software-Komponenten zu organisieren und zu gewährleisten.
- Erarbeitung eines Konzeptes für eine Betriebs- und Hostinginfrastruktur, in das die Ergebnisse vorangegangener DFG-Projekte mit einbezogen werden.

Standardisierung *(Schwerpunkt BBAW)*

- Entwicklung und Realisierung von Standardisierungen im Bereich der OCR, einschließlich Zeichencodierung, Training, Ground Truth, Formate (u.a. TEI) und Evaluation, unter Einbeziehung von relevanten Gremien und Organisationen.
- Standards und Richtlinien wie die DFG-Praxisregeln Digitalisierung müssen gepflegt, vermittelt und weiterentwickelt werden. Durch verschiedene Formate wie Workshops sowie Gremienarbeit ist der Wissenstransfer für die OCR von historischen Dokumenten zu gewährleisten.

Die angeführten Projektziele werden im Verlauf der Pilotierungsphase der Implementierungsvorhaben anhand der Auseinandersetzung mit der OCR-D-Software, durch die Identifikation von Entwicklungsbedarfen sowie ggf. mittels Konsolidierung der Implementierungsprojekte (eventuell Bildung von Konsortien oder Verbünden) präzisiert und ggf. stärker an die Bedarfe angepasst. Die organisatorische und institutionelle Absicherung der Ergebnisse von OCR-D kann von den beteiligten Antragstellern sowohl in der Pilotierungsphase der Implementierungsprojekte als auch in der Projektphase gewährleistet werden. Das geplante Vorhaben begleitet die Implementierungsprojekte vom Beginn der Pilotierungsphase an, um die OCR-D-Software koordiniert in den praktischen Einsatz überführen zu können.

Wolfenbüttel, den 20. Mai 2020



Prof. Dr. Peter Burschel
Direktor Herzog August Bibliothek