Massendigitalisierung alter Drucke

OCR-D in Bibliotheken

01.03.2021

Dr. Elisabeth Engl

Gliederung

1 Aufbau und Ziele des OCR-D-Projekts

2 Funktionsweise der OCR-D-Software

3 Ergebnisse der OCR-D-Teststellung

4 Bibliothekarische Anforderungen und die OCR-D-Software

5 Aktueller Stand und Ausblick auf die 3. Projektphase

1 Aufbau und Ziele des OCR-D-Projekts

Hauptziel von OCR-D ist die konzeptionelle Vorbereitung der Transformation der VD-Drucke (16.–19. Jh.) in maschinenlesbare Form.

- die Erstellung von Ground Truth
- die Erarbeitung von Standards hinsichtlich Metadaten
- die Weiterentwicklung der Optical Layout Recognition (OLR)
- die Analyse vorhandener Tools, auch zur Nachkorrektur
- die Entwicklung von technischer Dokumentation, Schnittstellen, Spezifikationen
- die Erstellung eines Workflows zur Massenvolltextdigitalisierung
- die Erstellung von Verfahren der Qualitätssicherung

1 Aufbau und Ziele des OCR-D-Projekts

SUB | NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

GWDG Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

HERZOG AUGUST BIBLIOTHEK

KIT Karlsruher Institut für Technologie

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

DFKI Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

BSB Bayerische StaatsBibliothek Information in erster Linie

Julius-Maximilians-UNIVERSITÄT WÜRZBURG

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

IN OMNIBUS VERITAS UNIVERSITÄT MANNHEIM

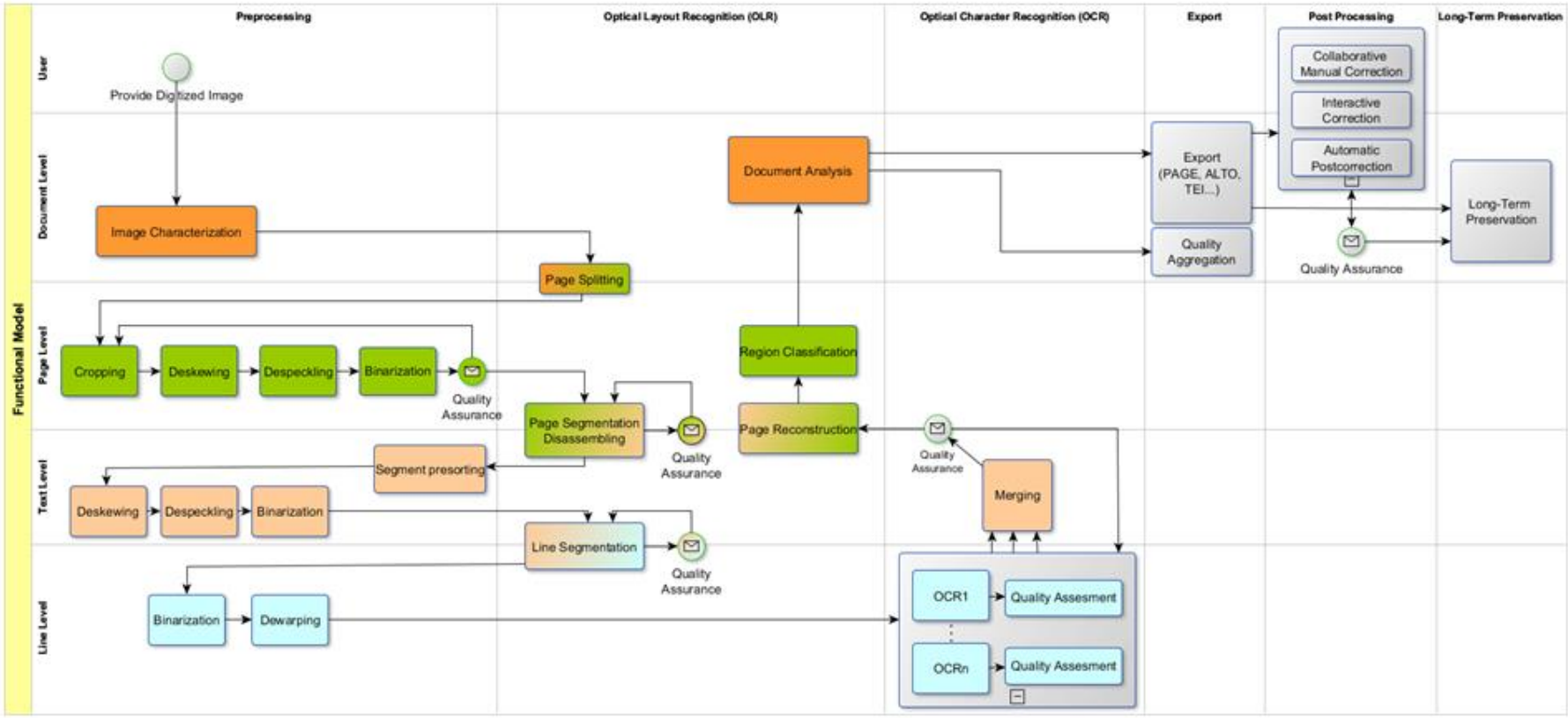
JG|U JOHANNES GUTENBERG UNIVERSITÄT MAINZ

UNIVERSITÄT LEIPZIG

Staatsbibliothek zu Berlin Preußischer Kulturbesitz

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

2 Funktionsweise der OCR-D-Software



2 Funktionsweise der OCR-D-Software

- Modular aufgebaute Software mit offenem Quellcode
- METS als Steuerungsdatei
- PAGE-XML
- JSON als Schemadatei
- Python API und CLI

3 Ergebnisse der OCR-D-Teststellungen

3 Ergebnisse der OCR-D-Teststellungen

- OCR-D-Software ist in allen Pilotbibliotheken installierbar
- Dokumentation zu Installation, Verwendung und Workflows vorhanden
- Software läuft sehr stabil
- Laufzeit einzelner Prozessoren ist für Massenprozessierung noch zu lang
- Bereits gute Erkennungsergebnisse

4 Bibliothekarische Anforderungen und die OCR-D-Software

- Sehr hohe Erkennungsrate bereits gute Testergebnisse
- Integrierbarkeit in bestehende Digitalisierungsworkflows Ziel von Phase III
- Einfache Bedienung robuste, gut dokumentierte Software mit nötigen Schnittstellen für Massendigitalisierung und vorkonfigurierten Workflows
- Kosten- und zeiteffiziente Prozessierung liberal lizenzierte Software, noch zu bestimmen

- Vortrainierte Modelle ja

- Inbetriebnahme auf verschiedenen Plattformen ja
- Layout- und Strukturerkennung ja, wird in Phase III weiter verbessert
- Gesicherte Weiterentwicklung Ziel von Phase III
- Aktive Nutzer-Community wird angestrebt
- Offener Quellcode ja

5 Aktueller Stand und Ausblick auf die dritte Projektphase

- Modulprojekte sind alle abgeschlossen
- OCR-D-Software ist als lauffähiger Prototyp auf GitHub verfügbar
- Ab 2021 Implementierung der OCR-D-Software in bestandshaltenden und -verarbeitenden Einrichtungen

Vielen Dank für Ihre Aufmerksamkeit!

OCR-D mitverfolgen:

Ausprobieren: https://github.com/OCR-D/

Mitreden: https://gitter.im/OCR-D/Lobby

Nachlesen & Lernen: https://www.ocr-d.de ; https://github.com/OCR-D/ocrd-website/wiki

Ausschreibung zur Implementierung der OCR-D-Software

- Implementierung des OCR-D-Prototyp in bestandshaltenden und -verarbeitenden Einrichtungen
- Dafür sind möglichst generische Implementierungspakete zu entwickeln, die den gesamten OCR-Workflow umfassen
- Ziel der Implementierungspakete sind verschiedene Anwendungsszenarien (bspw. Implementierung auf Arbeitsplatzrechnern, als Webdienst, etc.)

⇒ Am Ende der 3. Projektphase soll die OCR-D-Software in einer breiten Auswahl an Einrichtungen mit möglichst geringem Einrichtungsaufwand zur Volltextdigitalisierung von VD-Materialien eingesetzt werden können