



Die OCR-D-Workflowengine

18.02.2020

Elisabeth Engl



Gliederung

2

1. Aufbau und Ziele des OCR-D-Projekts
2. Funktionsweise der OCR-D-Software
3. Vom Bild zum Text



1. Aufbau und Ziele des OCR-D-Projekts

3

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

4



H E R Z O G
A U G U S T
B I B L I O
T H E K



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ



UNIVERSITÄT
LEIPZIG



berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

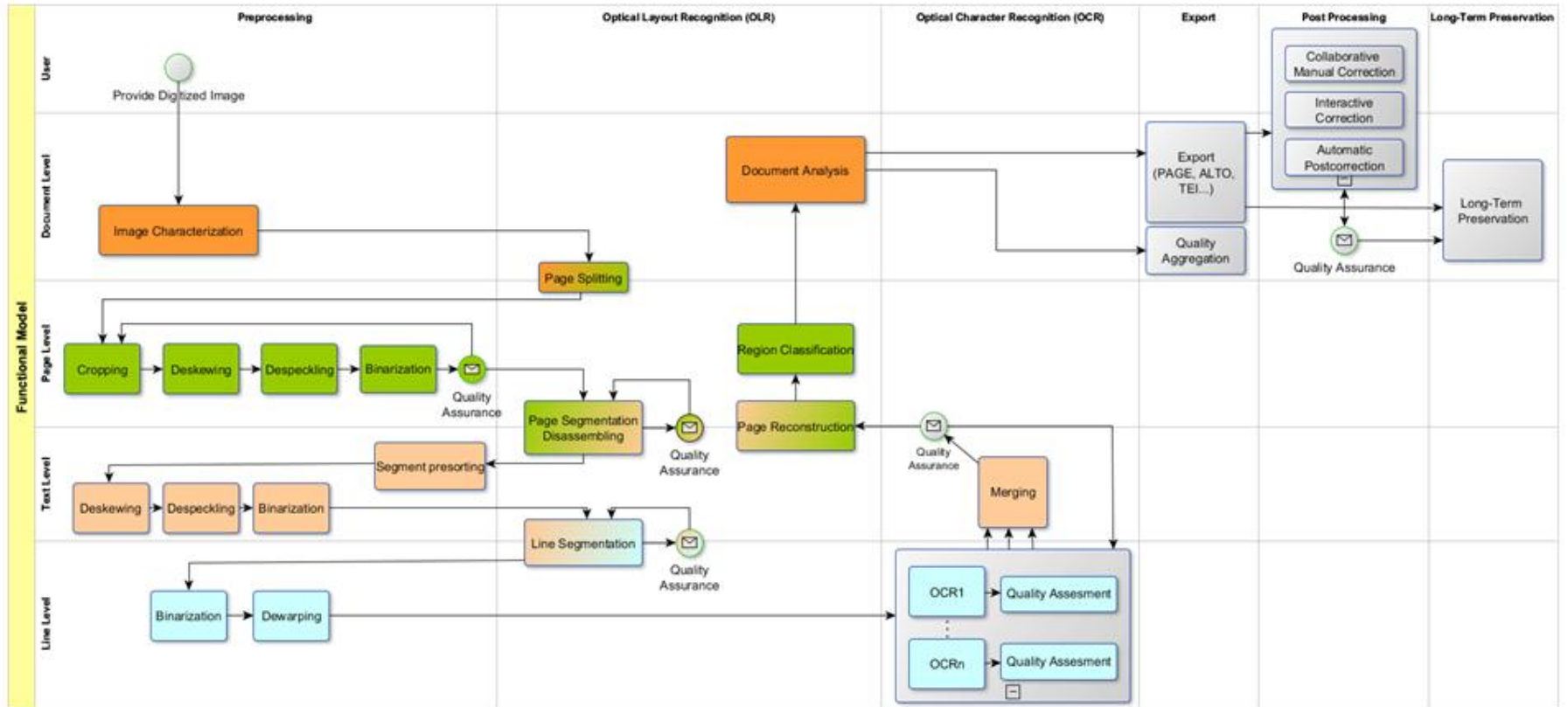
Staatsbibliothek
zu Berlin

Preußischer Kulturbesitz



2. Funktionsweise der OCR-D-Software

5





2. Funktionsweise der OCR-D-Software

6

- Modular aufgebaute Software mit offenem Quellcode
- METS als Steuerungsdatei
- PAGE als Ausgabeformat
- JSON als Schemadatei
- API für Massenprozessierung



3. Vom Bild zum Text

7

1. METS-Datei clonen

```
$ ocrd workspace clone LINK_TO_METS .
```

2. Dateigruppe mit benötigten Bildern auswählen

```
<mets:fileGrp USE="ORIGINAL">  
...  
</mets:fileGrp>
```

3. Gewünschte Werkzeuge aufrufen

```
$ ocrd-cis-ocropy-binarize -I ORIGINAL -O BIN -m mets.xml  
$ ocrd-tesseractocr-segment-region -I BIN -O REGS -m mets.xml
```



Vielen Dank für Ihre Aufmerksamkeit!

8

OCR-D mitverfolgen:

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

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

Nachlesen & Lernen: <https://www.ocr-d.github.io>