OPERANDI

OCR-D Performanzoptimierung und Integration

Kitodo-Anwendertreffen 25.11.21

GWDG Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

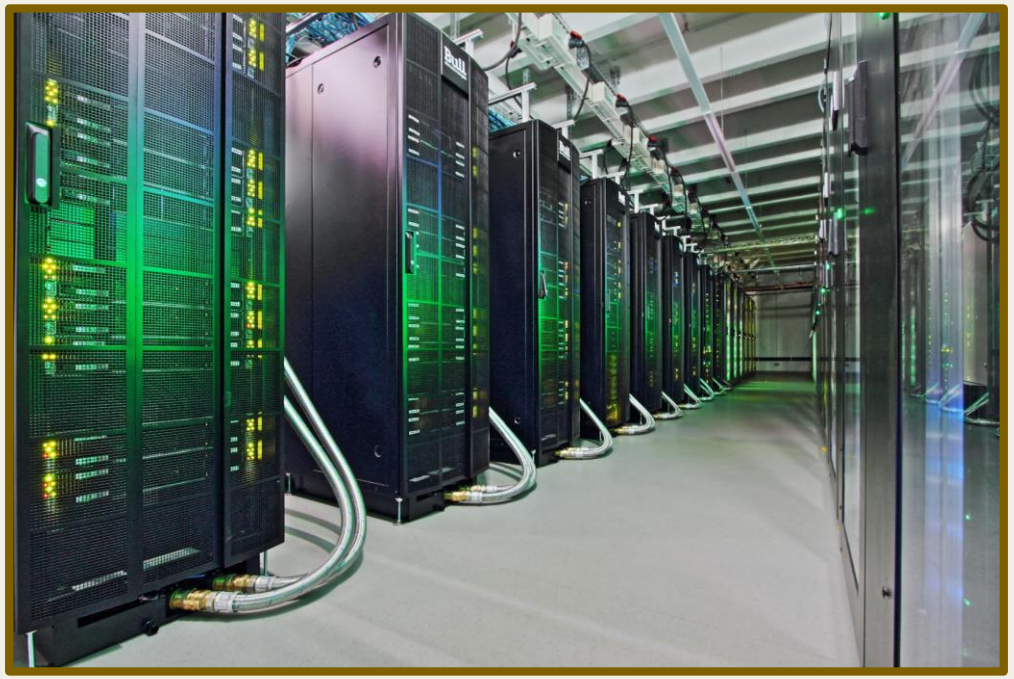
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN | SUB

OPERANDI – Projektziel

Entwicklung und Aufbau eines auf OCR-D basierenden Implementierungspaketes zur Massenvolltexterfassung mit verbessertem Durchsatz, bei besserer Qualität der Ergebnisse

Nachnutzbar von anderen Vorhaben und Einrichtungen mit vergleichbaren Anforderungen



OPERANDI – Projektinfo

- Teilnehmende Institutionen:
  - Niedersächsische Staats- und Universitätsbibliothek Göttingen
  - Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen

- Projektdauer:
  - 24 Monate

- Projektstart:
  - 1. Oktober 2021

- Gefördert im Rahmen der DFG-Initiative OCR-D

Hintergrund: DFG-Initiative OCR-D

- Ziel: konzeptionelle und technische Vorbereitung der Volltexttransformation der im deutschen Sprachraum erschienenen Drucke des 16.-19. Jahrhunderts
- Prozessschritte → OCR-D Software (Open Source) → optimale Workflows
- 3. Projektphase: 4 Implementierungsprojekte, 3 Modulprojekte

OPERANDI – Ziele

- Implementierung für die Massendigitalisierung
  - stabil, performant, fehlertolerant, skalierbar
- Adaptive, parallelisierte Workflows
- Taskmanagement und -priorisierung
- Asynchrone Interprozesskommunikation durch Schnittstellen
- Einfach bedienbares Implementierungspaket
- Einfache Datenspeicherung
- Leichte Übertragung in alternative Prozessierungsumgebungen (Cloud)

Zwei Szenarien

Szenario 1: OCR-Erzeugung für bereits digitalisierte Werke

Batch-Prozessierung

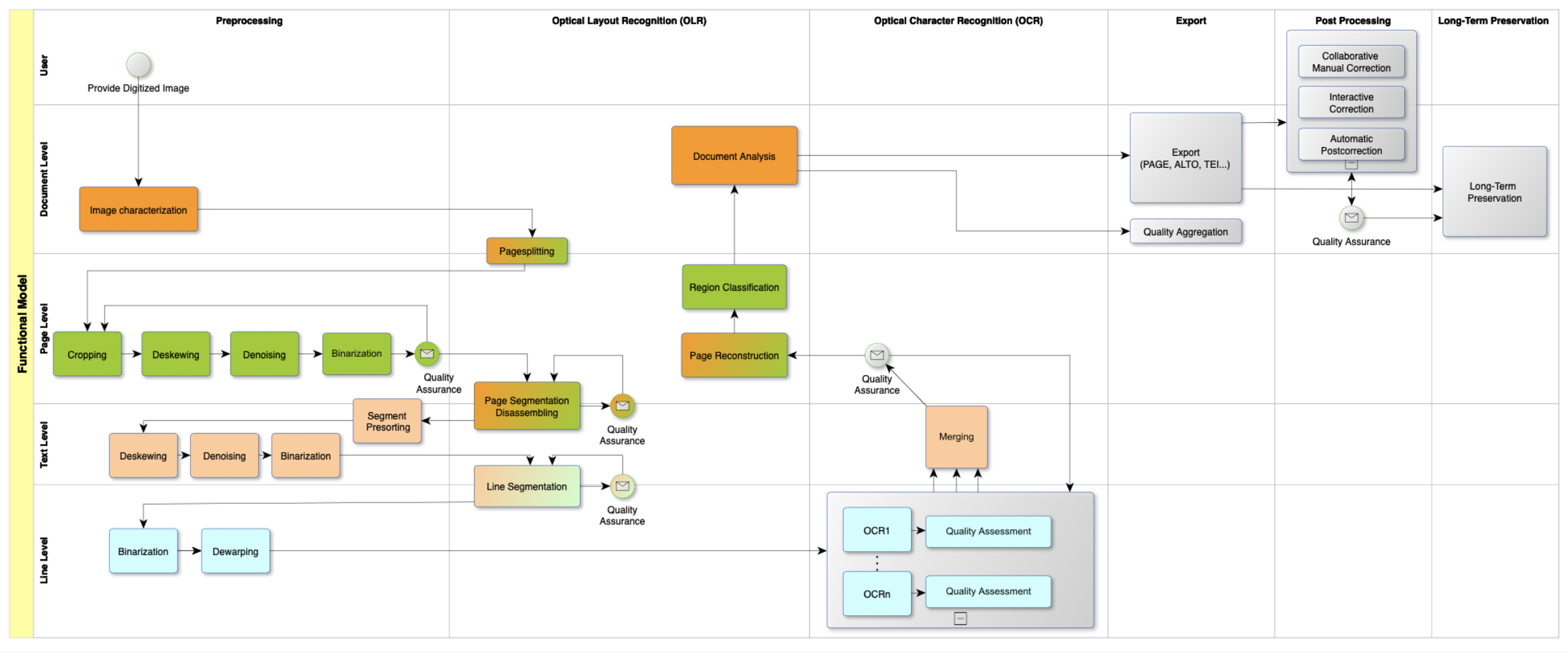
- Use Case: VD18-Bestand
- Pilotierung: 0,68 Seiten pro Minute
- Vision: 60 Seiten pro Minute

Szenario 2: OCR-Erzeugung für neu zu digitalisierende Werke

Integration

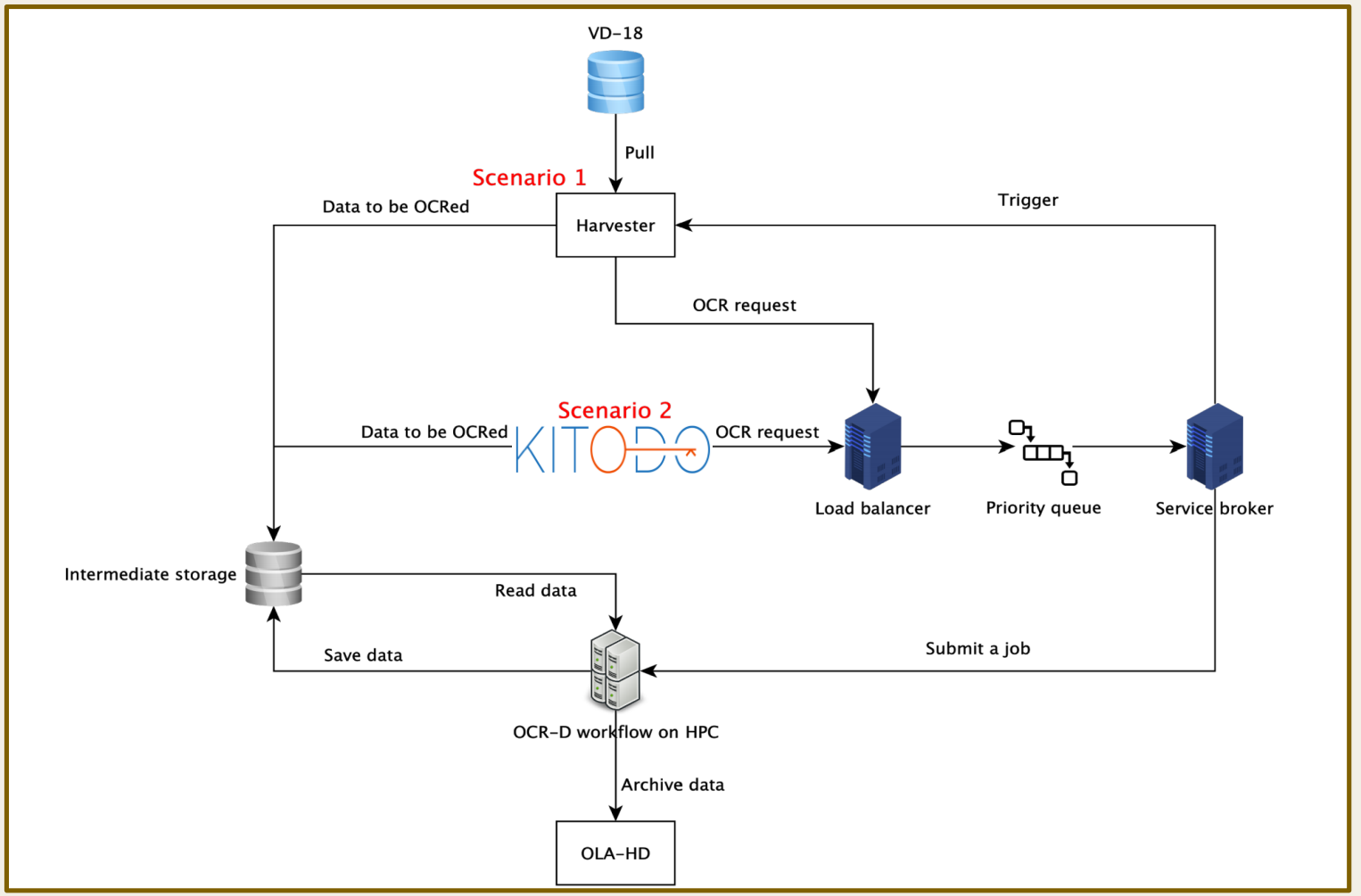
- Einbindung in Workflowschritte von Digitalisierungssoftware
- Goobi als Proof-of-Concept
- Projektziel: Portierung der Lösung auf Kitodo

OCR-D Workflow



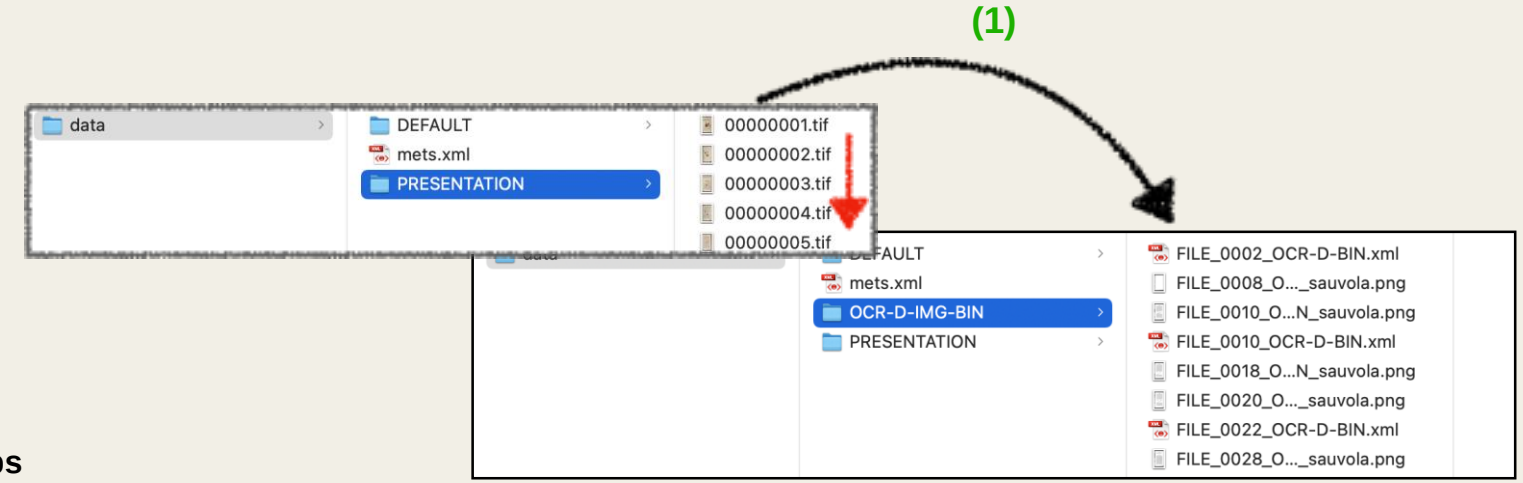
OCR-D Workflows. (https://ocr-d.de/en/workflows)

OPERANDI – Architektur



Test OCR-D Workflow

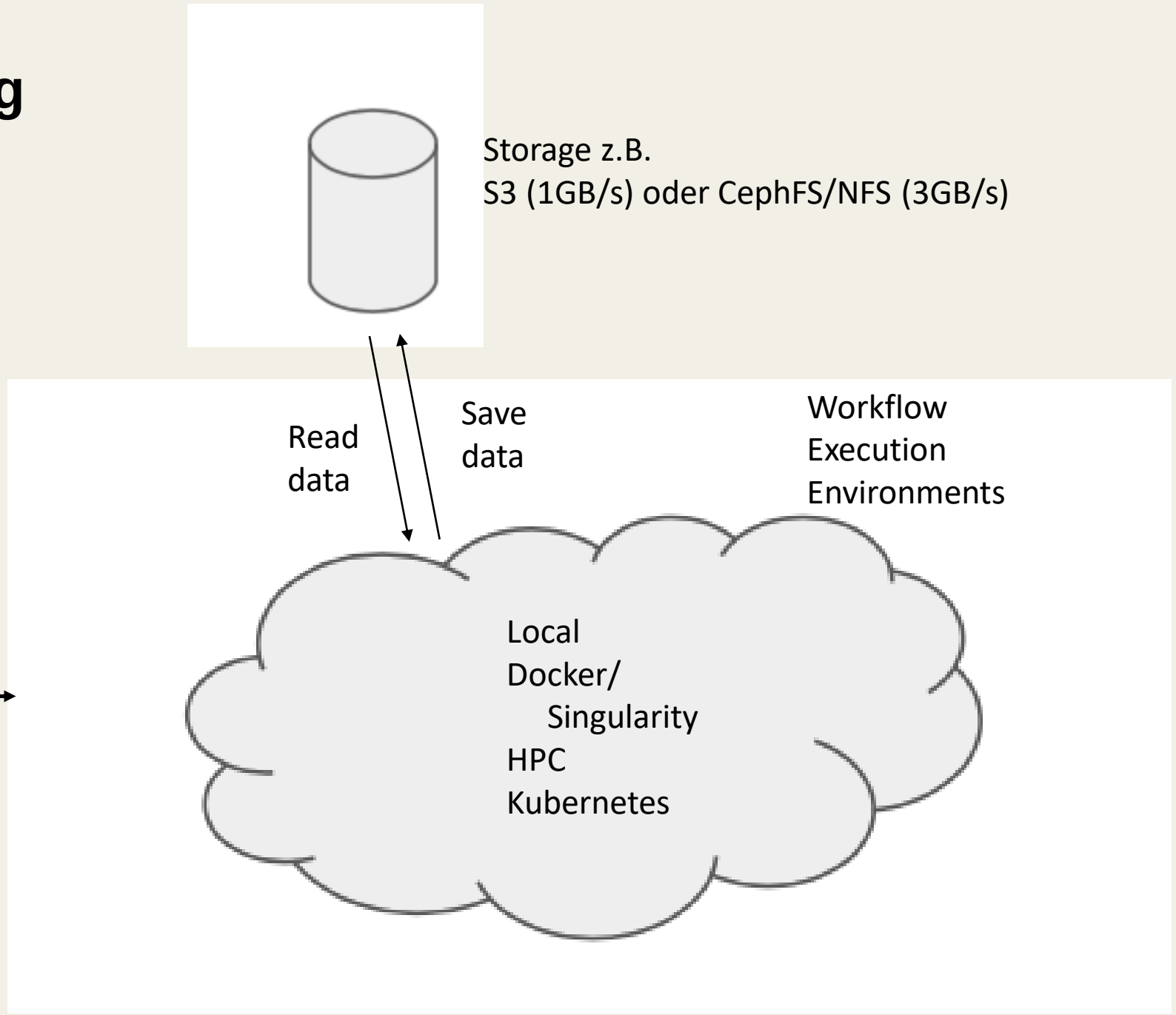
```
$> ocrd process    'tesserocr-binarize   -I PRESENTATION   -O OCR-D-IMG-BIN'   \                    (1)
                       'tesserocr-segment-region -I OCR-D-IMG-BIN -O OCR-D-SEG-REGION'   \            (2)
                       'tesserocr-segment-line -I OCR-D-SEG-REGION -O OCR-D-SEG-LINE'   \             (3)
                       'tesserocr-recognize -I OCR-D-SEG-LINE -O OCR-D-OCR-TESSEROCR'                 (4)
```



processor execution steps

sequential processing

Ausführungsumgebung



OPERANDI – Kontakt

Lilja Sautter (SUB Göttingen)
sautter@sub.uni-goettingen.de
Jörg-Holger Panzer (SUB Göttingen)
panzer@sub.uni-goettingen.de
Triet Doan (GWDG)
triet.doan@gwdg.de

GWDG Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN
NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN | SUB