OCR-D

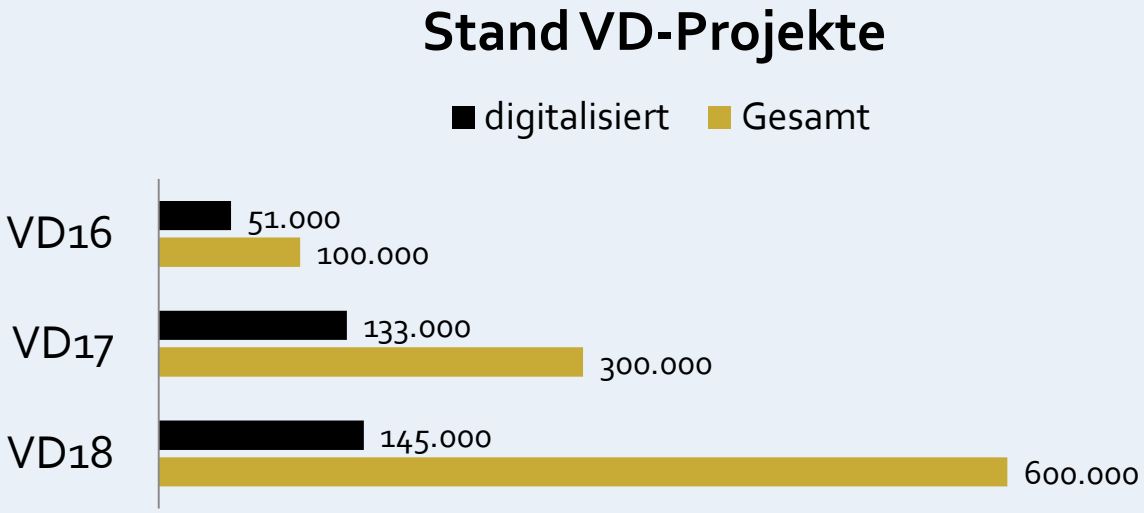
KOORDINIERUNGSPROJEKT ZUR WEITERENTWICKLUNG VON OCR-VERFAHREN

Gefördert von der Deutschen Forschungsgemeinschaft

09.02.2016

Dr. Thomas Stäcker, Elisa Herrmann

Einleitung



- Herausforderungen:
  - Unterschiedliche Sprachen: Latein, Frühneuhochdeutsch, Griechisch, Hebräisch u.a.
  - Unterschiedliche Schrifttypen: Antiqua, Fraktur, Kursive
  - Unterschiedliches Layout

Ziel

Konzeptionelle Vorbereitung der Transformation der VD-Drucke (16.-18. Jh.) und der Drucke des 19. Jh. in maschinenlesbare Form.

Das Projekt

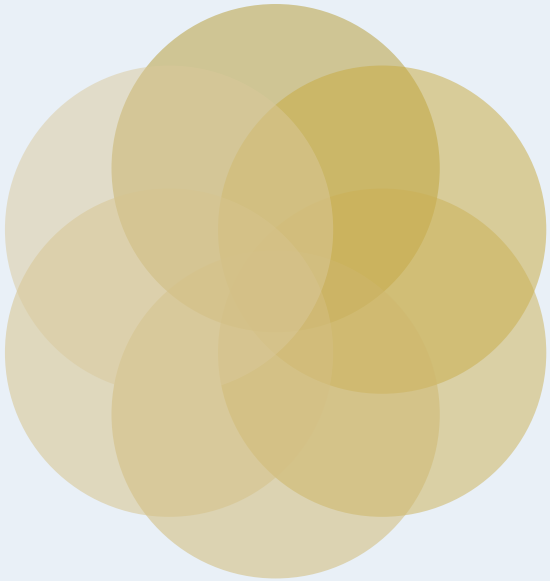
- Projektpartner
  - Herzog August Bibliothek Wolfenbüttel
  - Berlin-Brandenburgische Akademie der Wissenschaften,
    insb. Deutsche Textarchiv in Berlin (DTA)
  - Bayerische Staatsbibliothek in München

- 2 Phasen:
  - 1. Aufbau der Koordinierungsstruktur und Konzeption der Praxisphase (12 Monate)
  - 2. Ausschreibung und konzeptionelle Begleitung der Pilotprojekte (24 Monate)

Arbeitspakete

Standards

Workflow & Use Cases



Referenzkorpora

Qualitätskontrolle

Analyse vorhandener Tools

Langzeit-archivierung

Standards

- Richtlinien zur Bereitstellung von OCR-Rohdaten
- Schwerpunkte
  - Definition von Formaten
  - Notwendige strukturelle Minimalkodierung
  - Fehlerdefinition
  - Messung von Strukturgenauigkeit
  - Standards für Referenzkorpora und Trainingsdaten

- Verantwortlich: BBAW

Referenzkorpora

- Identifizierung vorhandener Referenzkorpora
- Neue forschungsorient. Trainingskorpora
- Ground Truth-Daten

- Verantwortlich: Bayerische Staatsbibliothek München

Analyse vorhandener Tools

- Interoperabilität , Robustheit, Skalierbarkeit
- Aktueller Forschungsstand in etablierten OCR-Workflows
- Editoren zur Nachkorrektur
  - Crowdsourcing
- Text/Bild-Strukturerkennung
- Annotation der Korpora
- Werkzeuge für die Qualitätskontrolle
- Austausch zwischen Entwicklern und wiss. Nutzern

- Verantwortlich: BBAW

Langzeitarchierung & Persistenz

- Spezifikation v. Datenformaten und Metadaten
- Dokumentation dynamischer Prozesse
- Spezifikation von Archivobjekten & LZA-Workflows

- Verantwortlich: Bayerische Staatsbibliothek

Qualitätsprüfung

Quelle: Martin Fisch, via Flickr (CC-BY)

- Derzeit: t-Studentische Verteilung
  - einfach anzuwenden, zeitaufwändig
  - Qualitätsprüfung erfolgt am Ende

Qualitätsprüfung

- Neuer Ansatz, möglichst ohne Ground Truth:
  - Vorhersage von Binarisierungserfolg
  - Messung der OLR
  - Messung von OCR-Erfolg/ Vorhersage von OCR-Erfolg

- Abstufung der OCR-Ergebnisse nach Use Cases
  - Kommunikation zum Nutzer: BnF, Anzeige der OCR-Ergebnisse

Verantwortlich: HAB

Qualitätsprüfung

EPIS J-OLJE

JOANNEM AKEPPLERUM

MATHEMATICUM C^SAREUM

S C R I P T i;
INSERTIS AD EASDEM

RESPONSIONIBUS KEPPLERIANIS,

QUOTQUOT HACTENUS REPERIRI

OPUS NOVUM, Qy6°RECONDITA KEPPLERIAN^E

DOCTRIN^E CAPITA DILUCIDE EXPLICANTUR,ET HISTORIA
LITERARIA IN UNIVERSUM MIRIHCE ILLUSTRATUR,

NUNC PRIMUM

CUM PR^FATIONE DE MERITIS GERMANORUM IN MATHESIN,

INTRODUCTIONE IN HISTORIAM LITERARIAM S./ECULORUM
XVI. ET XVII. ET JO. KEPPLERI VITA

JUSSU ET AUSPICIIS

ROMANOR. IMPERAT.

S. A. P. F. V. P. P.

EX MANUSCRIPTIS EDITUM.

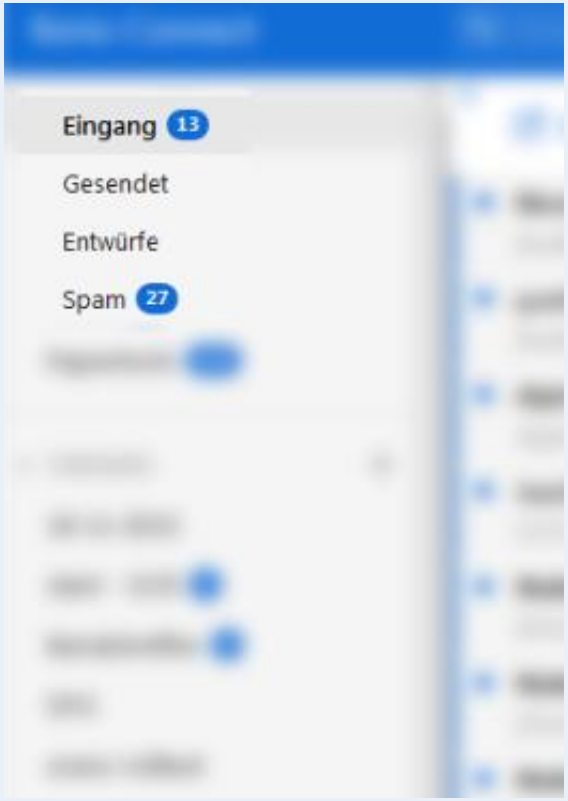
Le texte affiché peut comporter un certain nombre d'erreurs. En effet, le mode texte de ce document a été généré de façon automatique par un programme de reconnaissance optique de caractères (OCR). Le taux de reconnaissance estimé pour ce document est de **86.43** %.
En savoir plus sur l'OCR

Source gallica.bnf.fr / Bibliothèque nationale de France

Le texte affiché peut comporter un certain nombre d'erreurs. En effet, le mode texte de ce document a été généré de façon automatique par un programme de reconnaissance optique de caractères (OCR). Le taux de reconnaissance estimé pour ce document est de **86.43** %.
En savoir plus sur l'OCR

http://gallica.bnf.fr/ark:/12148/bpt6k57815621/f8.item.r=Kepler,%20Johannes.texteImage

Qualitätsprüfung



Hidden Markov Modelle

→ BBN Byblos

→ On-line Handschriftenerkennungs-system (Uni Duisburg)

Qualitätsprüfung

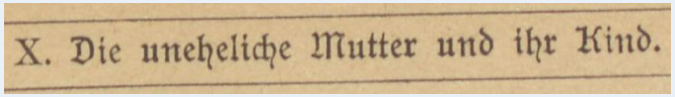
Guter Rat ist ~~Feuer~~ teuer

A Feuer

B Ungeheuer

C teuer

D blauer



X. Die uneheliche Mutter und ihr Rind.

Markuse, Max (1906): Uneheliche Mütter. urn:nbn:de:kobv:109-1-6385137

Qualitätssicherung

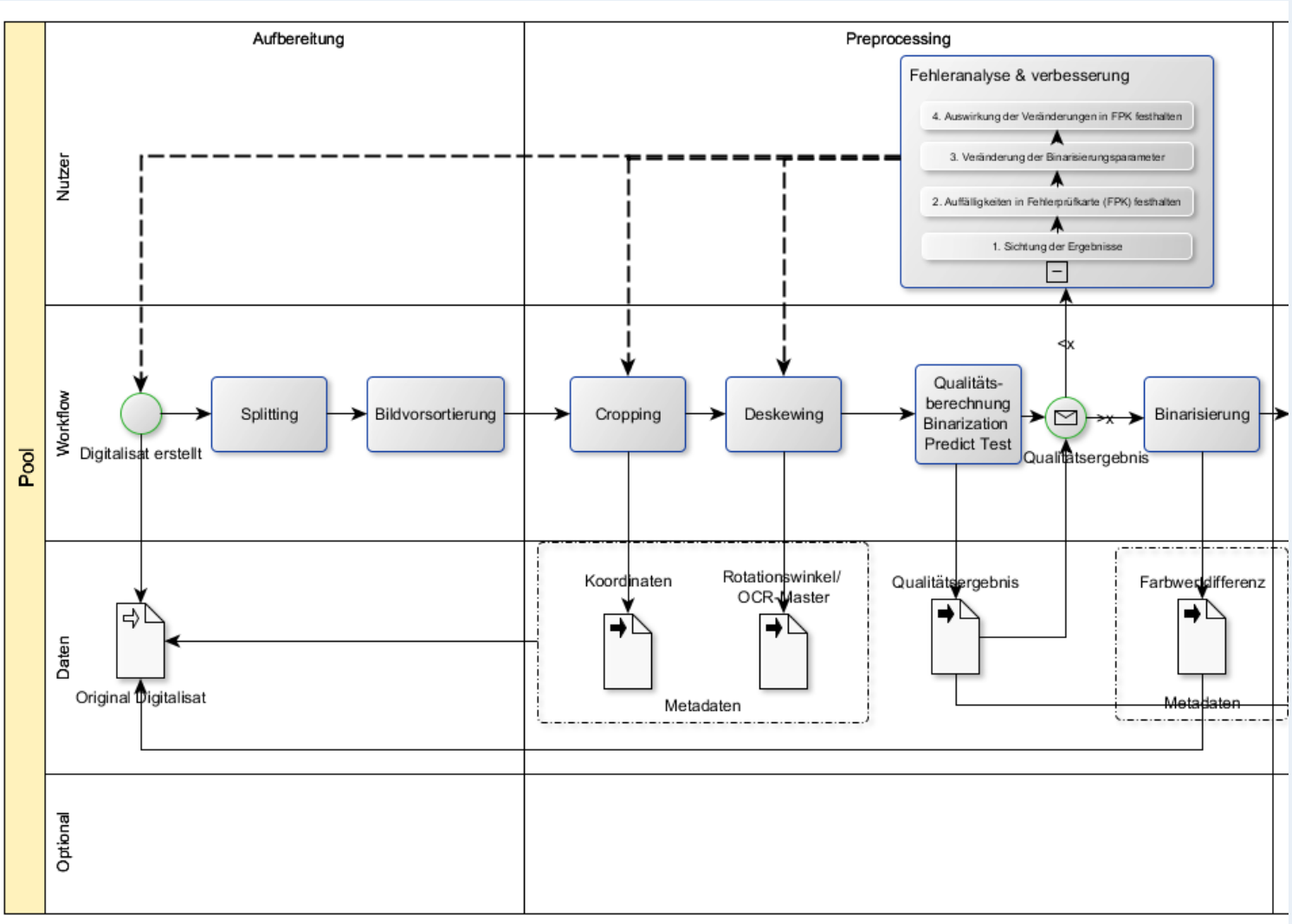
- Grundmethode: Six Sigma

Workflow

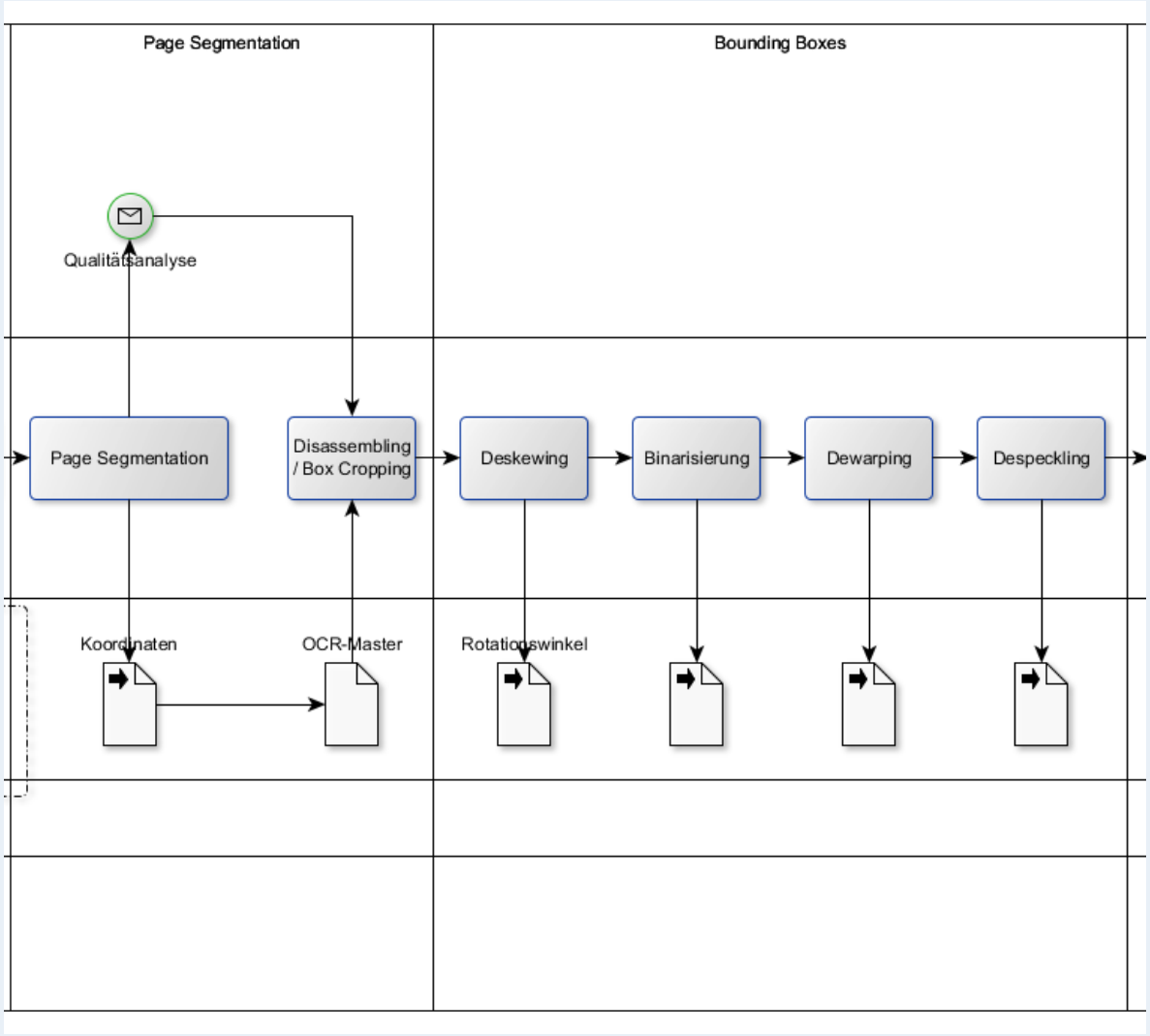
- Einbindung neuer Erkenntnisse aus den anderen Arbeitspaketen (v.a. OLR und Qualitätsprüfung)
- Modularer Aufbau
  - Adaptive Anpassung an verschiedene Bedürfnisse

- Verantwortlich: Bayerische Staatsbibliothek München

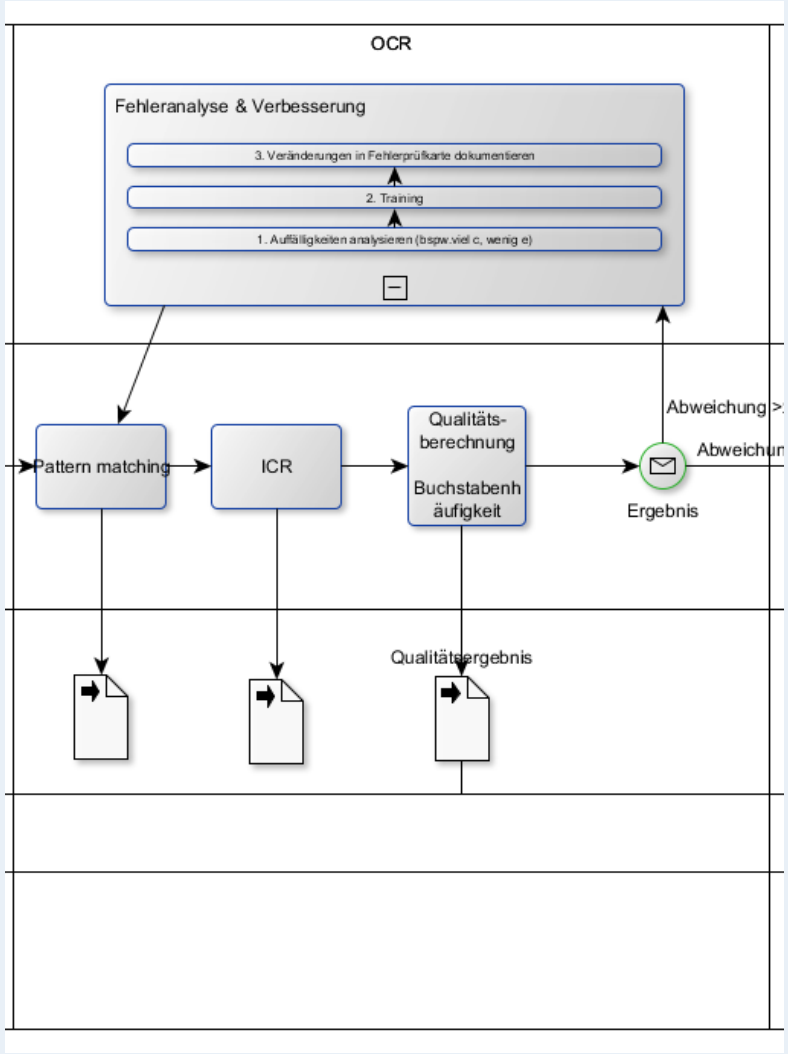
Workflow - I


Aufbereitung
Preprocessing
Pool
Nutzer
Workflow
Daten
Optional
Fehleranalyse & verbesserung
4. Auswirkung der Veränderungen in FPK festhalten
3. Veränderung der Binarisierungsparameter
2. Auffälligkeiten in Fehlerprüfkarte (FPK) festhalten
1. Sichtung der Ergebnisse
Digitalisat erstellt
Splitting
Bildvorsortierung
Cropping
Deskewing
Qualitäts-
berechnung
Binarization
Predict Test
Qualitätsergebnis
Binarisierung
Koordinaten
Rotationswinkel/
OCR-Master
Metadaten
Qualitätsergebnis
Farbwertdifferenz
Metadaten
Original Digitalisat

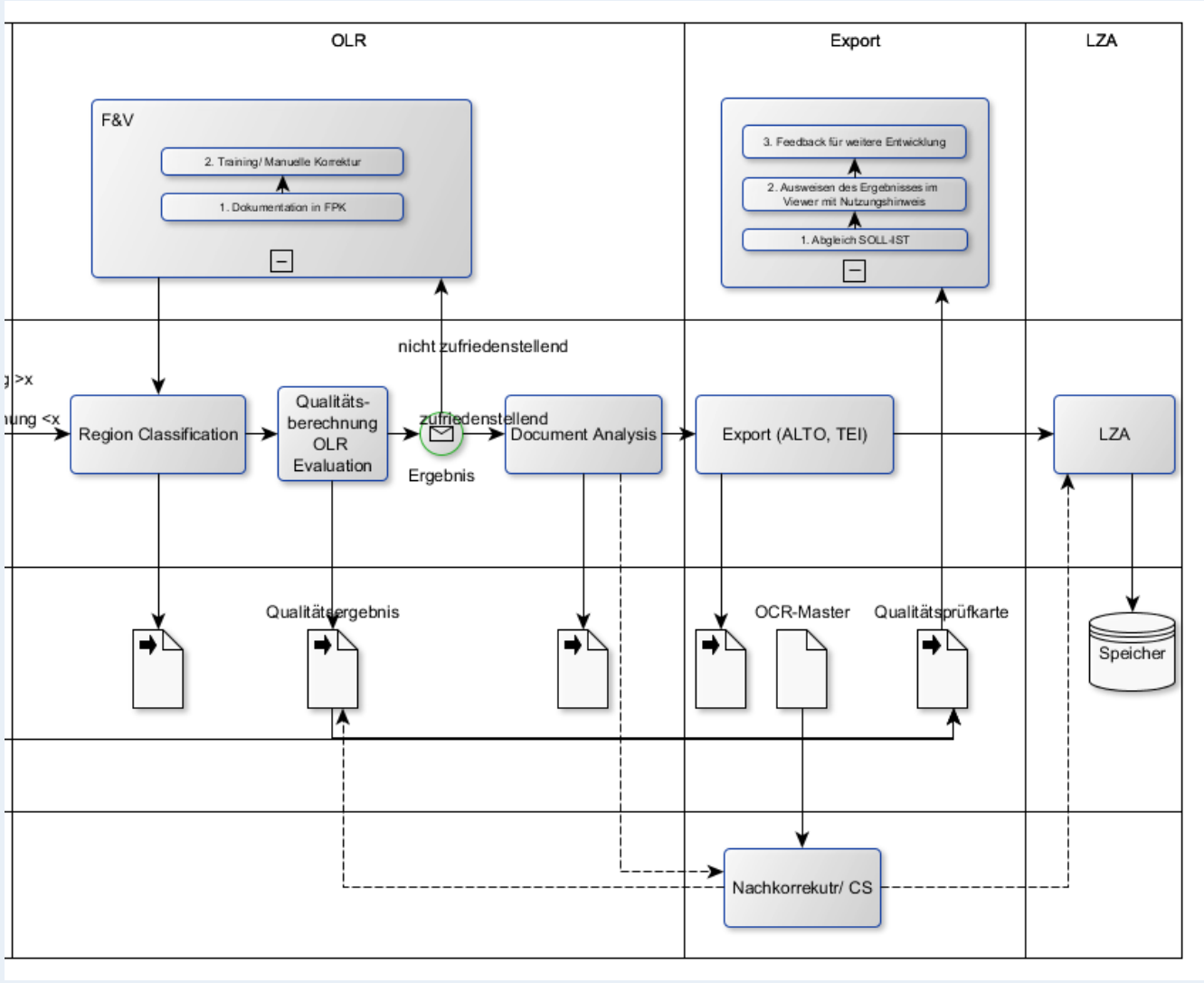
Workflow - II


Page Segmentation
Bounding Boxes
Qualitätsanalyse
Page Segmentation
Disassembling / Box Cropping
Deskewing
Binarisierung
Dewarping
Despeckling
Koordinaten
OCR-Master
Rotationswinkel

Workflow - III



Workflow - IV


OLR
Export
LZA
F&V
2. Training/ Manuelle Korrektur
1. Dokumentation in FPK
3. Feedback für weitere Entwicklung
2. Ausweisen des Ergebnisses im Viewer mit Nutzungshinweis
1. Abgleich SOLL-IST
nicht zufriedenstellend
Region Classification
Qualitäts-berechnung OLR Evaluation
zufriedenstellend
Ergebnis
Document Analysis
Export (ALTO, TEI)
LZA
Qualitätsergebnis
OCR-Master
Qualitätsprüfkarte
Speicher
Nachkorrekutr/ CS

Ausblick

- Begutachtung durch den wissenschaftlichen Beirat
- Konzeption: Mitte April an DFG
- Ausschreibung der Pilotprojekte

Vielen Dank

- Kontakte
  - Webseite: www.ocr-d.de
  - Elisa Herrmann
elisa.herrmann@hab.de
05331 808 306