OCR-D

KOORDINIERUNGSPROJEKT ZUR WEITERENTWICKLUNG VON OCR-VERFAHREN

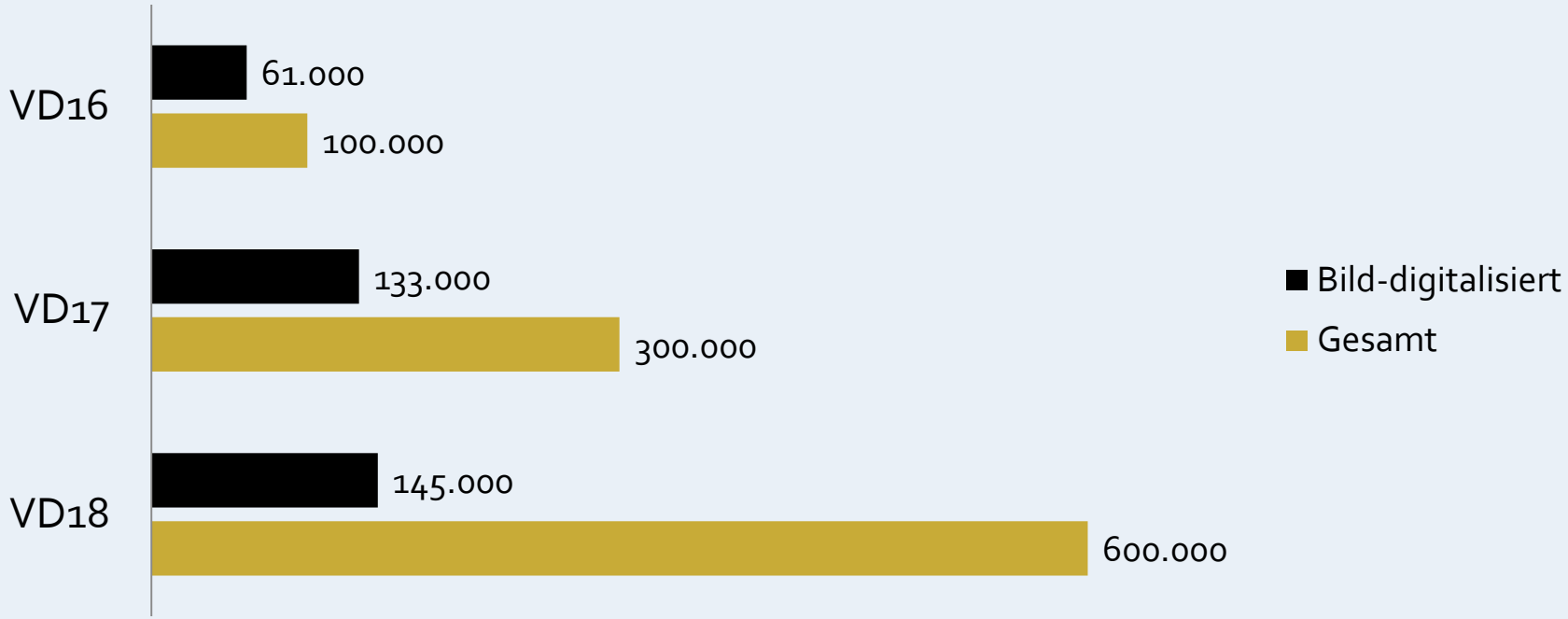
Gefördert von der Deutschen Forschungsgemeinschaft

25.02.2016

Elisa Herrmann

VD 16-18

Katalogisierung → Bild-Digitalisierung → OCR



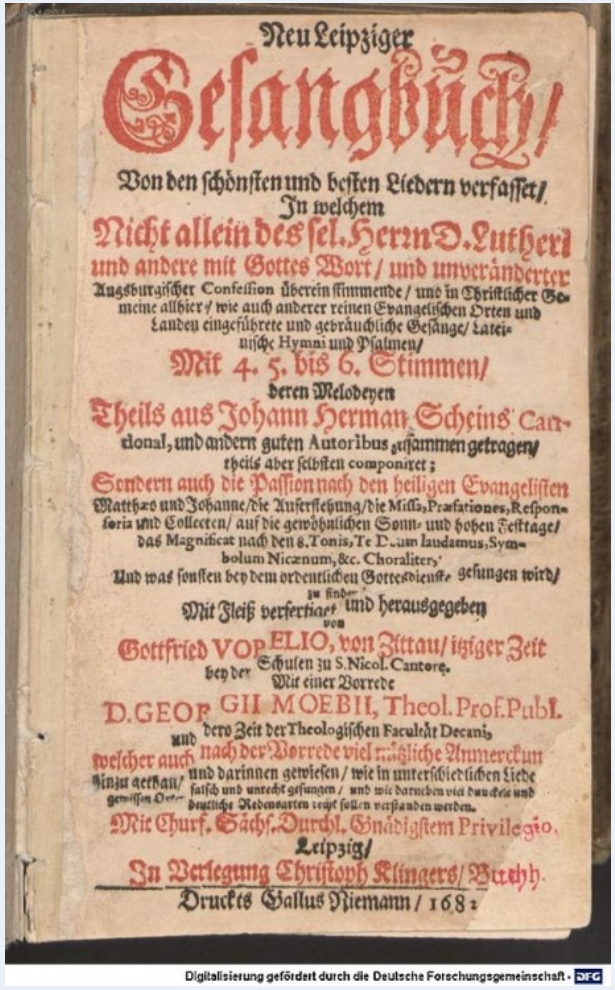
* Gerundete Zahlen

Ziel

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

Herausforderungen

- Material:
  - Sprachen: v.a. Latein, Deutsch
  - Schriftarten und -ausprägungen: u.a. Antiqua, Fraktur, Kursive
  - verschiedene Textsorten mit spezifischem Layout
- Uneinheitliche Standards
- Neuprozessierung, dynamische Datensicherung


Neu Leipziger
Gesangbuch/
Von den schönsten und besten Liedern verfasset/
In welchem
Nicht allein des sel. Herrn D. Lutheri
und andere mit Gottes Wort/ und unveränderter
Augsburgischer Confession überein stimmende/ und in Christlicher Gemeine allhier/ wie auch anderer reinen Evangelischen Orten und Landen eingeführete und gebräuchliche Gesänge/ Lateinische Hymni und Psalmen/
Mit 4. 5. bis 6. Stimmen/
deren Melodeyen
Theils aus Johann Herman Scheins Cantional, und andern guten Autoribus zusammen getragen/
theils aber selbsten componiret;
Sondern auch die Passion nach den heiligen Evangelisten
Matthæo und Johanne/ die Auferstehung/ die Missa, Præfationes, Responsoria und Collecten/ auf die gewöhnlichen Sonn- und hohen Festtage/
das Magnificat nach den 8. Tonis, Te Deum laudamus, Symbolum Nicænum, &c. Choraliter;
Und was sonsten bey dem ordentlichen Gottesdienste gesungen wird/
zu finden.
Mit Fleiß versertiget und herausgegeben
von
Gottfried VOPELIO, von Zittau/ itziger Zeit
bey der Schulen zu S. Nicol. Cantore.
Mit einer Vorrede
D. GEORGII MOEBII, Theol. Prof. Publ.
und dero Zeit der Theologischen Facultät Decani,
welcher auch nach der Vorrede viel nützliche Anmerckungen hinzu gethan/ und darinnen gewiesen/ wie in unterschiedlichen Liedern gewissen Oertern falsch und unrecht gesungen/ und wie darunter viel dunckele und undeutliche Redensarten recht sollen verstanden werden.
Mit Churf. Sächs. Durchl. Gnädigstem Privilegio.
Leipzig/
In Verlegung Christoph Klingers/ Buchh.
Druckts Gallus Niemann/ 1682.
Digitalisierung gefördert durch die Deutsche Forschungsgemeinschaft - DFG

http://daten.digitale-sammlungen.de/bsb00091604/image_5

Das Projekt

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

- 2 Phasen:
  1. Aufbau der Koordinierungsstruktur und Konzeption der Projektphase
  2. Ausschreibung und konzeptionelle Begleitung der Pilotprojekte

Arbeitspakete

Standards

Workflow &
Use Cases

Referenzkorpora

Qualitätskontrolle

Analyse
vorhandener
Tools

Langzeit-
archivierung

Funktionsmodell

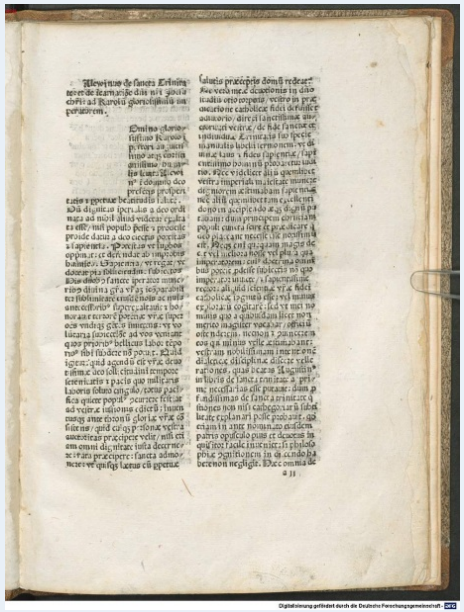
- Einbindung neuer Erkenntnisse aus den einzelnen Arbeitspaketen
- Einbindung bestehender Werkzeuge
- Modularer Aufbau
  - Adaptive Anpassung an verschiedene Bedürfnisse
  - Anpassung an zukünftige Entwicklungen

Funktionsmodell – I

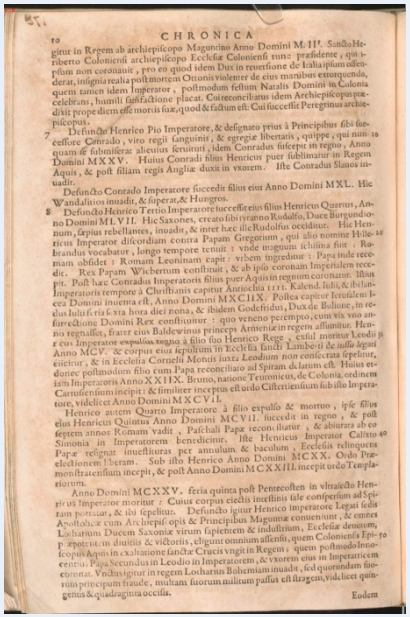
- Bild-Digitalisat bereitstellen
- Splitting (optional)
- Bildvorsortierung

Bildvorsortierung

- Bildvorsortierung
  - Z.B. Layoutanalyse



http://daten.digitale-sammlungen.de/bsb00005199/image_9



http://daten.digitale-sammlungen.de/bsb00010079/image_16

Funktionsmodell – II

- Preprocessing Page Level
- (z.B. Cropping, Deskewing, Binarisierung)
- Qualitätskontrolle
- OLR1: Page Segmentation
- Preprocessing Segment Level
- Qualitätskontrolle

Page Segmentation

- Segmentierung des Bilddigitalisats in
  - Textzonen
  - Nichttextzonen

http://daten.digitale-sammlungen.de/bsb00042833/image_3

Funktionsmodell - III

- OCR
- Qualitätskontrolle
- OLR 2: Region Classification
- Qualitätskontrolle

OCR & OLR

- OCR
  - Vgl. bzw. Kombination **klassischer Zeichenerkennungsverfahren** auf Glyphenebene mit **segmentierunsgfreien Ansätzen** (z.B. "Deep Learning"-Verfahren auf Basis neuronaler Netze)
- OLR
  - **Region Classification**: Bestimmung der layout-semantischen Funktion der einzelnen Regionen (Überschrift, Marginalie, Fußnote etc.)
  - **Document Analysis**: Extrapolierung der Dokumentstruktur aus den entsprechenden Strukturelementen (Überschrift)

Funktionsmodell - IV

- Export
- Nachkorrektur/ Crowdsourcing
- Qualitätskontrolle
- LZA

Ausblick

- Antworten auf technische, informationswissenschaftliche & organisatorische Herausforderungen

Konsolidiertes Verfahren zur OCR-Verarbeitung von Digitalisaten des schriftlichen deutschen Kulturerbes des 16.-19. Jh.

Kontaktdaten

- Webseite: www.ocr-d.de

- Elisa Herrmann, elisa.herrmann@hab.de, +49 5331 808-306