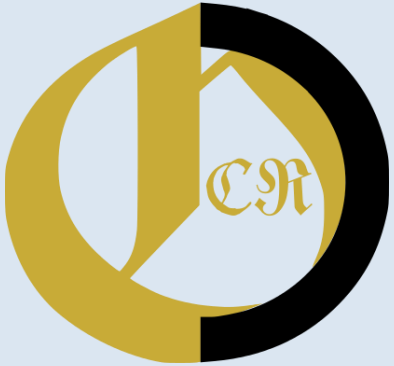


# Bibliothekarische Digitalisierungspraxis und die OCR-D-Software

12.02.2020

Elisabeth Engl

# Gliederung



1 Derzeitige (Bild-)Digitalisierungspraxis in Bibliotheken

2 Bisherige Erfahrungen mit OCR-Projekten und weitere Pläne

3 Anforderungen an OCR-Software

4 Ergebnisse der OCR-D-Teststellung

5 Bibliothekarische Anforderungen und die OCR-D-Software

# 1 Derzeitige (Bild-)Digitalisierungspraxis in Bibliotheken

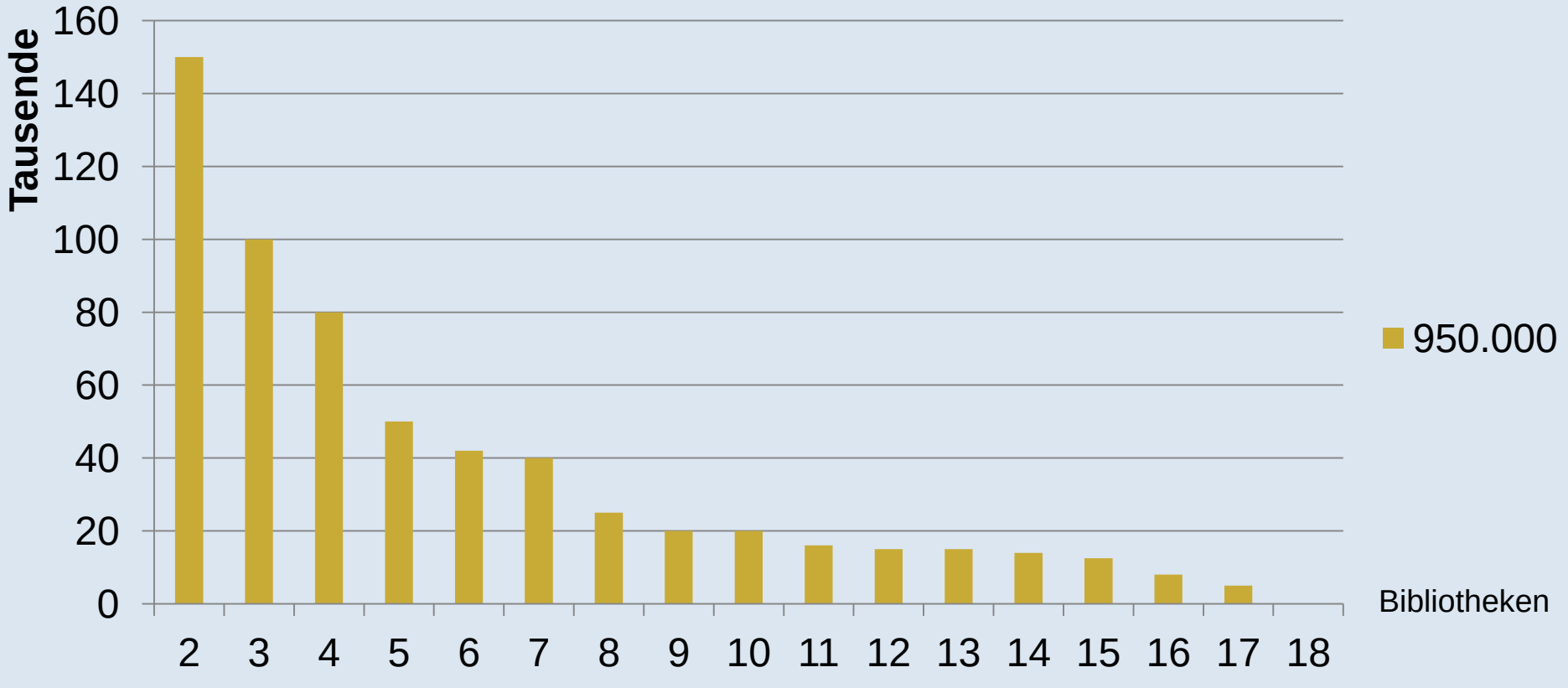


Hintergrund zur Erhebung

- Umfrage mit VD-Bibliotheken zur derzeitigen Digitalisierungspraxis und Erfahrungen mit OCR-Projekten durchgeführt
- Fragebogen wurde von allen vier VD-Trägerbibliotheken und 14 weiteren, v.a. im Bereich der Digitalisierung größeren Bibliotheken ausgefüllt

# Anzahl der durchschnittlich pro Monat digitalisierten Seiten (in Tausende)





# 2 Bisherige Erfahrungen mit OCR-Projekten und weitere Pläne



- 61 % haben bereits mindestens 2 OCR-Projekte durchgeführt
- Davon führt die Hälfte OCR in-house durch (v.a. mit ABBYY Finereader)
- OCR wird meist bei Neudigitalisierungen von Drucken des 18.–20. Jhd. eingesetzt
- 82 % aller Umfrageteilnehmer haben grundsätzlich Interesse an OCR-Projekten zu VD-Titeln

# 3 Anforderungen an OCR-Software



- Sehr hohe Erkennungsrate
- Integrierbarkeit in bestehende Digitalisierungsworkflows
- Einfache Bedienung
- Kosten- und zeiteffiziente Prozessierung

- Vortrainierte Modelle

- Inbetriebnahme auf verschiedenen Plattformen
- Layout- und Strukturerkennung
- Gesicherte Weiterentwicklung
- Aktive Nutzer-Community
- Offener Quellcode

# 4 Ergebnisse der OCR-D-Teststellung



- Software wurden zwischen November 2019 und Januar 2020 in neun Pilotbibliotheken getestet

MARTIN-LUTHER-UNIVERSITÄT HALLE-WITTENBERG
ULB Sachsen-Anhalt

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

ULB Universitäts- und Landesbibliothek Darmstadt

SLUB Wir führen Wissen.

Universitätsbibliothek Heidelberg

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

Staatsbibliothek zu Berlin Preußischer Kulturbesitz

HERZOG AUGUST BIBLIOTHEK

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

# 4 Ergebnisse der OCR-D-Teststellung



- OCR-D-Software ist in allen Pilotbibliotheken installierbar
  - Zwischenzeitlich durch ocrd_all deutlich vereinfacht
- Problematisch war die zum Testzeitpunkt noch fragmentierte Dokumentation (v.a. zu möglichen Workflows)
- Software läuft sehr stabil
- Laufzeit einzelner Prozessoren ist für Massenprozessierung noch zu lang
- Bereits gute Erkennungsergebnisse

# 5 Bibliothekarische Anforderungen und die OCR-D-Software



- Sehr hohe Erkennungsrate  bereits sehr gute Testergebnisse
- Integrierbarkeit in bestehende Digitalisierungsworkflows Technisch möglich
- Einfache Bedienung  robuste, gut dokumentierte Software mit nötigen Schnittstellen für Massendigitalisierung und vorkonfigurierten Workflows
- Kosten- und zeiteffiziente Prozessierung  lizenzfreie Software, noch zu bestimmen

- Vortrainierte Modelle  ja

- Inbetriebnahme auf verschiedenen Plattformen  ja
- Layout- und Strukturerkennung  ja, wird noch weiter verbessert
- Gesicherte Weiterentwicklung  wird angestrebt
- Aktive Nutzer-Community  wird angestrebt
- Offener Quellcode  ja

Vielen Dank für Ihre Aufmerksamkeit!

# Diskussionsbereiche



- OCR-D in der Bibliothek
- OCR-Volltexte: ein Angebot der Bibliotheken für die Forschung
- Herausforderung OCR-D Software