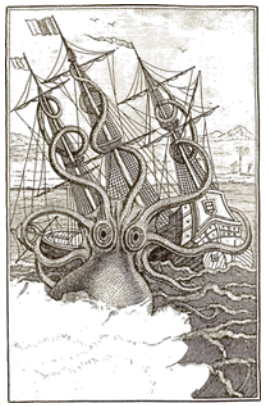
Open Source OCR-Systeme

Tesseract OCR



Calamari OCR

EasyOCR

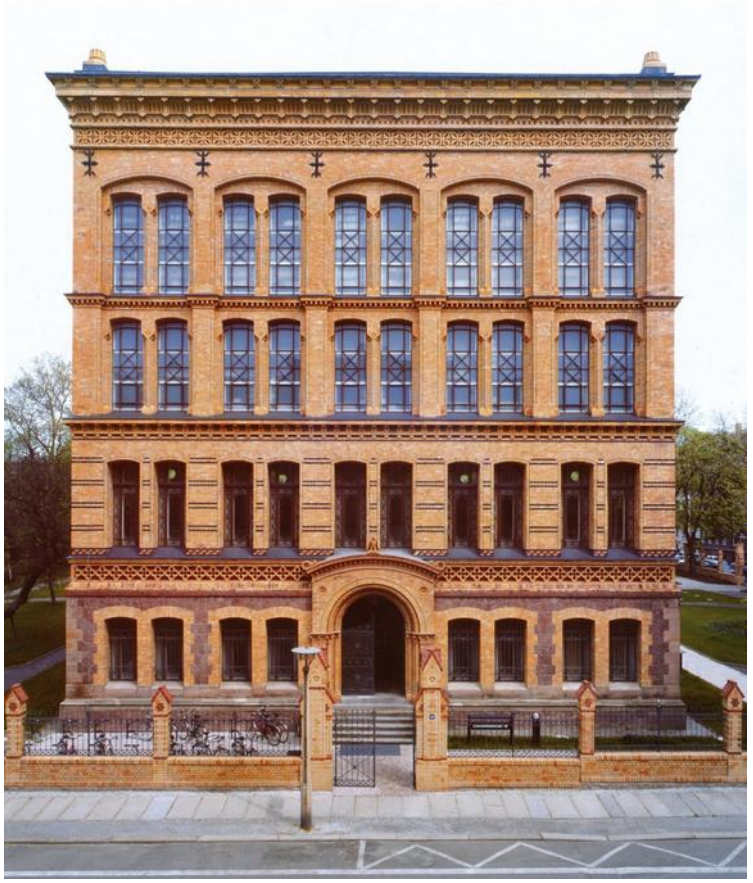
Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Uwe Hartwig
Universitäts- und Landesbibliothek Sachsen-Anhalt
uwe.hartwig@bibliothek.uni-halle.de

UNIVERSITÄTS- UND LANDESBIBLIOTHEK SACHSEN - ANHALT

ULB Sachsen-Anhalt

- eine der größten Zeitungssammlungen Deutschlands
- 2012-2014 Pilotierungsprojekt
  Pilotphase zur Zeitungsdigitalisierung
  externer Dienstleister
  ca. 130.000 Seiten (Original)
- 2019-2021 Projekt
  Digitalisierung Historischer Deutscher Zeitungen I
  Volltexterzeugung mit Open Source Komponenten
  ca. 530.000 Seiten (Mikrofilm)
- Zeitungsportal
  (https://digitale.bibliothek.uni-halle.de/zd/date)
  aktuell zugänglich:
  - „Hallesches Tageblatt“ (1799-1892)
  - „Hallische Nachrichten“ (1918-1944)



Übersicht I

	Schwerpunkt	Modelle	Training
Tesseract https://github.com/tesseract-ocr/tesseract	--	300+	möglich
EasyOCR https://github.com/JaidedAI/EasyOCR	Südostasien	80+	möglich
Calamari https://github.com/Calamari-OCR/calamari	Historische Drucke	20+ (GitHub)	möglich
Kraken https://github.com/mittagessen/kraken	--	2 (zenodo)	möglich
OCR-D https://github.com/ocr-d	Historische Drucke / VD	--	möglich

Übersicht II

	Plattform	Integration	Input	Output
Tesseract	Linux / MacOS / Windows	CLI/Library	tif, jpg, png	ALTO, ...
easyOCR	Linux (Torch)	CLI/Python	tif, jpg, png	nativ
Calamari	Linux (TensorFlow)	CLI/Python	png, tif	PAGE
Kraken	Linux (Torch)	CLI/Python	tif, png, ALTO, PAGE, hOCR	PAGE, ... nativ
OCR-D	Linux	CLI/Python	jpg, ...	PAGE, ...

Übersicht III

	Preprocessing	Konvertierung	QS	Sonstiges
Tesseract	nein	nein	Konfidenz	PDF-Textlayer
easyOCR	nein	nein	Konfidenz	Enterprise Support
Calamari	ja	nein	Konfidenz	Bestandteil von OCR4all
Kraken	ja	nein	Konfidenz	modular; PDF-Textlayer
OCR-D	möglich	möglich	Konfidenz + Tools	Workflowsystem