UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

Tesseract OCR – News

Stefan Weil, University Library Mannheim

ELAG 2019 Berlin

Text recognition (OCR) Software

commercial software

fat = used in libraries

ABBYY Finereader
BIT-Alpha
Readiris
OmniPage

Adobe Acrobat
CorelDraw
Microsoft OneNote
…

Tesseract
OCRopus / Kraken / Calamari
CuneiForm
…

free software

ABBYY Cloud OCR
Google Cloud Vision
Microsoft Azure Computer Vision
OCR.space Online OCR …

Cloud OCR

Tesseract OCR Features

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Open Source, GitHub https://github.com/tesseract-ocr/
- All-in-1 solution (preprocessing, layout and text recognition, postprocessing)
- More than 100 languages / more than 30 scripts
- Reads all common image formats (not PDF!)
- Creates text, PDF, hOCR, ALTO, TSV
- Large, wordwide user community
- Supports text recognition using neural networks
- Active development also in DFG project OCR-D

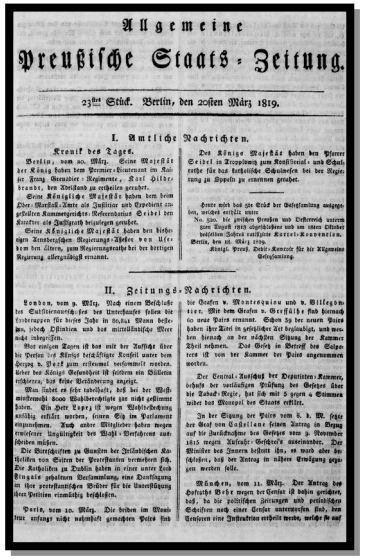
Tesseract OCR News

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- 1984–1994 Ray Smith, Hewlett-Packard
- 2006– Google (currently inactive)
- 2016–11 New LSTM-based neural network line recognizer
- 2017–02 Release 3.05 (only with old = legacy text recognition engine)
- 2017–09 Current fast and best LSTM models
- 2018–10 Release 4.0 (included in Linux distributions), 3.05 compatible
- 2018–12 ALTO support
- 2019–06 (?) Release 4.1

Tesseract at UB Mannheim

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM


Allgemeine
Preußische Staats-Zeitung.
23ßtes Stück. Berlin, den 20sten März 1819.
I. Amtliche Nachrichten.
II. Zeitungs-Nachrichten.

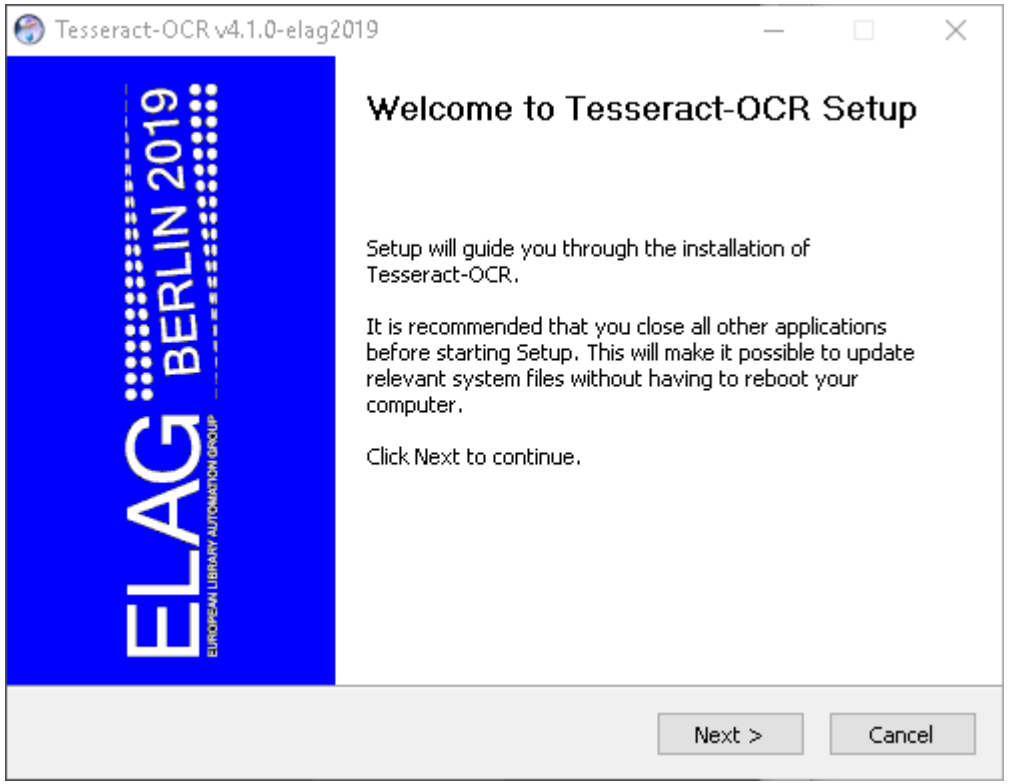
- Used in DFG project *Aktienführer* https://digi.bib.uni-mannheim.de/aktienfuehrer/
- Used for historical newspaper *Deutscher Reichsanzeiger* https://digi.bib.uni-mannheim.de/periodika/reichsanzeiger

- DFG project *OCR-D* http://www.ocr-d.de/,
  Module project *Optimized use of OCR methods – Tesseract as a component of the OCR-D workflow*

OCR-D
Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)

- Project *OCR-BW*,
  text recognition for archives and libraries in Baden-Württemberg

Tesseract 4.1 ELAG 2019 Edition



Latest installer for Windows
(still untested)

https://digi.bib.uni-mannheim.de/tesseract