Tesseract: Turbo und Training

Daniel Brenn
Universitäts- und Landesbibliothek Sachsen-Anhalt
daniel.brenn@bibliothek.uni-halle.de

UNIVERSITÄTS- UND
LANDESBIBLIOTHEK
SACHSEN - ANHALT

Gliederung

- Warum OpenSource?
- Tesseract mit neuronalen Netzen
- Turbo in der Digitalisierung
- Training für besondere Schriften

Ausgangslage

- Digitalisierung läuft aktuell noch über Visual Library der Firma semantics
- OCR mit AbbyyFine
- Volumenlizenzen zur Digitalisierung
- festgelegt auf bestimmte Sprachen und Schriften
- OCR nicht für alle Digitalisate möglich, kostenintensiv und eher unflexibel

- Entscheidung für Nutzung von Kitodo und DSpace
- Aufbau eines neuen, modularen Digitalisierungsworkflows mit OpenSource-Komponenten
- Plan zur Nutzung von AbbyyCloud in Kombination mit Tesseract OCR

Gliederung

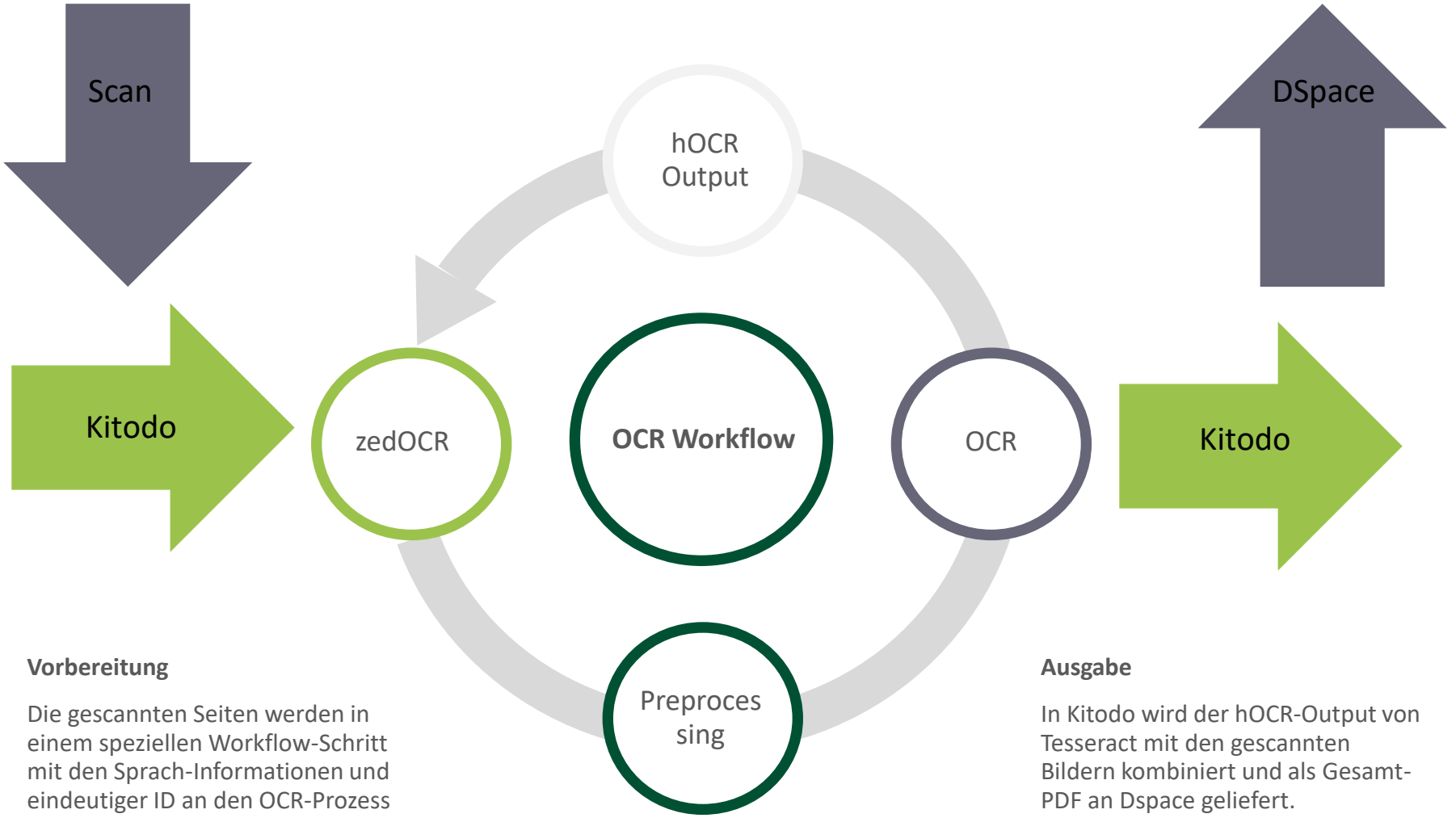
- Warum OpenSource?
- Tesseract mit neuronalen Netzen
- Turbo in der Digitalisierung
- Training für besondere Schriften

Warum OpenScource?

- Entscheidung zum Einsatz von OpenSource-Software für größere Flexibilität
- Insbesondere Tesseract bietet durch lange Entwicklung Zugriff auf sehr breites Feld von Trainingsdaten
- Nutzung für besondere Projekte durch eigene Weiterentwicklung
- Gute Möglichkeiten zur Kollaboration mit anderen Bibliotheken/ Projekten

- aber: OpenSource benötigt intensive Auseinandersetzung mit der Software und häufig mehr Arbeitszeit!
- zentral wichtig: transparente Dokumentation, um Wissen nicht nur bei einzelnen Kollegen zu bündeln

Warum Open Source? - Digitalisierungsworkflow



Vorbereitung

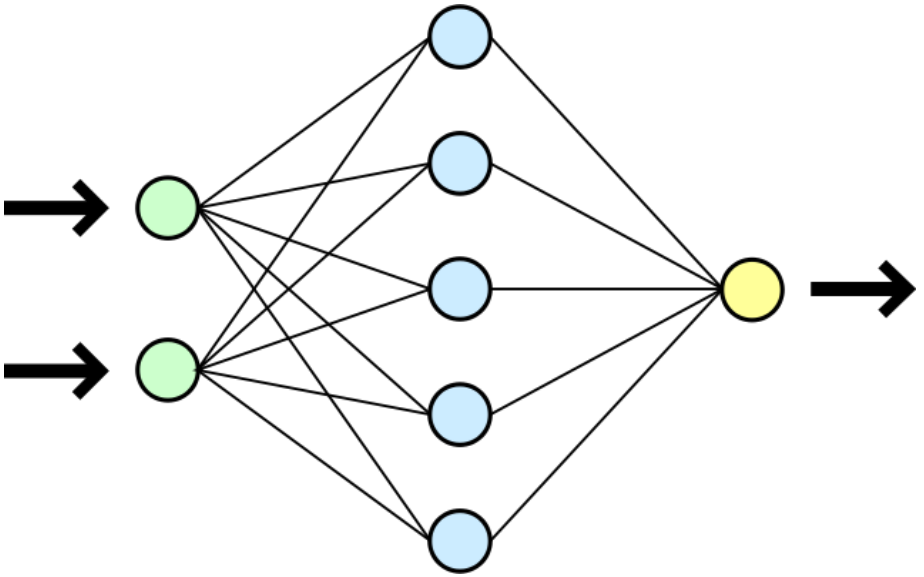
Die gescannten Seiten werden in einem speziellen Workflow-Schritt mit den Sprach-Informationen und eindeutiger ID an den OCR-Prozess weitergeleitet

Ausgabe

In Kitodo wird der hOCR-Output von Tesseract mit den gescannten Bildern kombiniert und als Gesamt-PDF an Dspace geliefert.

Gliederung

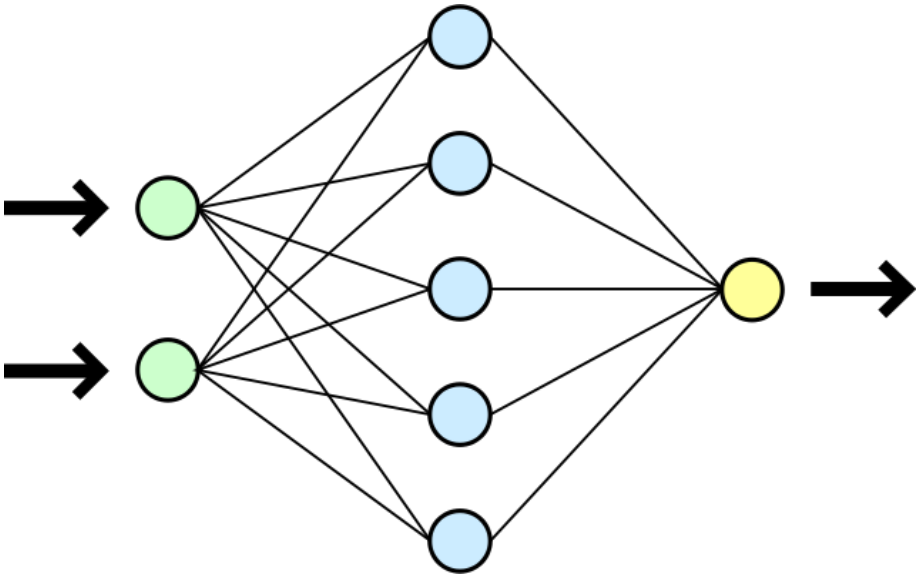
- Warum OpenSource?
- Tesseract mit neuronalen Netzen
- Turbo in der Digitalisierung
- Training für besondere Schriften



Quelle: Dake, Mysid –
CC BY 1.0, https://commons.wikimedia.org/w/index.php?curid=1412126

Tesseract mit neuronalen Netzen

- Tesseract nutzt seit Version 4 neuronale Netze
- konkret: LSTM (Long Short-Term Memory)
- dadurch erhebliche Verbesserung der Genauigkeit bei schwierigerem Layout und schlechter lesbaren Texten erkennbar
- ermöglicht OCR auch bei Schriften, die vorher nur schwer oder gar nicht verwendbar waren
- erhöhter Leistungsbedarf im Vergleich zu früheren Versionen



Gliederung

- Warum OpenSource?
- Tesseract mit neuronalen Netzen
- Turbo in der Digitalisierung
- Training für besondere Schriften

Nr. 6. Mittwoch, den 27. März 1889. Auflage 40 000.

General Anzeiger

für Halle und den Saalkreis.

Die heutige Nummer umfaßt 8 Seiten.

Zur gefl. Beachtung!

Der „General-Anzeiger für Halle und den Saalkreis" wird im gesammten Verbreitungsbezirke in jedem Hause bis zur 3. Etage

gänzlich kostenlos

ausgegeben.

Sollten unsere Träger die Ablieferung der Blätter nicht pünktlich besorgen oder unterlassen, so bitten wir die verehrl. Leser, uns dergl. unsere Filialen hiervon gefälligst zu benachrichtigen.

Anzeigen, betreffend Kauf, Miets- und Stellengesuche, sowie Angebote und Vermiethungen jeder Art werden bis zu 5 Zeilen einspaltig

kostenlos

aufgenommen.

Verlag des General-Anzeigers für Halle u. d. Saalkreis.

Haupt-Expedition:

Grosse Ulrichstr. 36.

Pönulde und Consorten.

Politische Uebersicht.

Deutsches Reich.

Die Springwurzel.

Turbo in der Digitalisierung

- für regulären Digitalisierungsworkflow ist relativ hohe OCR-Geschwindigkeit erforderlich
- nur Teilschritt der Digitalisierung
- Tesseract kann Prozesse auf einzelne Prozessorkerne auslagern und so parallelisieren
  - „ab Werk“ bis zu 4 parallele Prozesse
- Zur Verbesserung der Geschwindigkeit ist weitere Parallelisierung mit pytesseract möglich
- durch Installation auf VM ist ein nachträgliches Erhöhen der Kapazität möglich, die Vorplanung sollte die Notwendigkeit dafür aber minimieren
- Qualitätssicherung erfolgt vor Freischaltung im Repositorium

Nr. 6. Mittwoch, den 27. März 1889. Auflage 40 000.

General Anzeiger

für Halle und den Saalkreis.

Die heutige Nummer umfaßt 8 Seiten.

Zur gefl. Beachtung!

Der „General-Anzeiger für Halle und den Saalkreis“ wird im gesammten Verbreitungsbezirke in jedem Hause bis zur 3. Etage

gänzlich kostenlos

ausgegeben.

Stellen unsere Leser die Ablieferung der Blätter nicht pünktlich besorgen oder unterlassen, so bitten wir die verehrl. Leser, uns dergl. unsere Filialen hiervon gefälligst zu benachrichtigen.

Anzeigen, betreffend Kauf, Miethe und Stellengesuche, sowie Angebote und Vermiethungen jeder Art werden bis zu 5 Zeilen einspaltig

kostenlos

aufgenommen.

Verlag des General-Anzeigers für Halle u. d. Saalkreis.

Haupt-Expedition:

Grosse Ulrichstr. 36.

Pétroulede und Consorten.

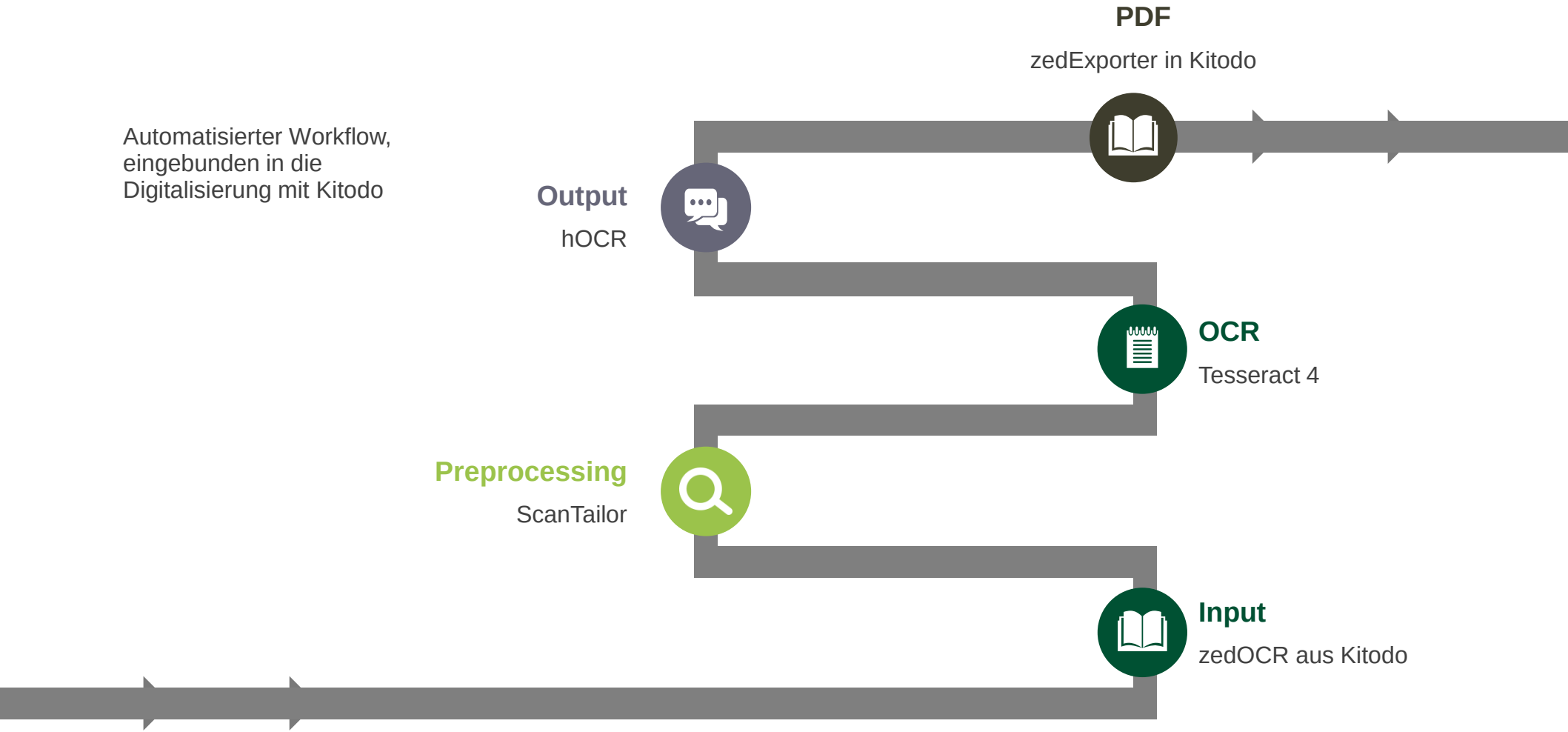
Politische Uebersicht.

Deutsches Reich.

Die Springwurzel.

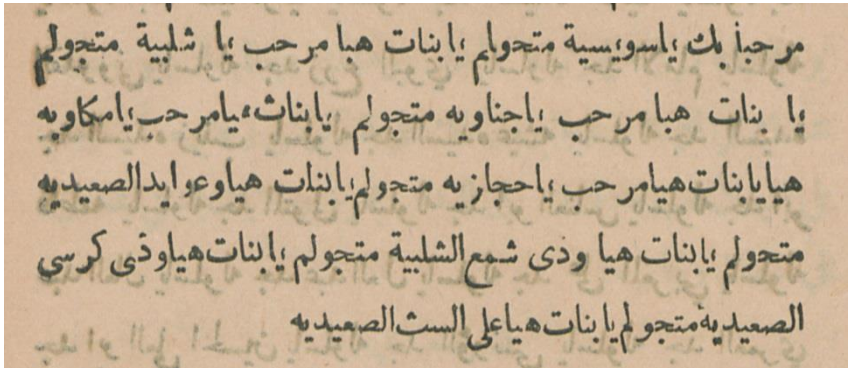
Automatisierter Digitalisierungsworkflow

Automatisierter Workflow, eingebunden in die Digitalisierung mit Kitodo



Gliederung

- Warum OpenSource?
- Tesseract mit neuronalen Netzen
- Turbo in der Digitalisierung
- Training für besondere Schriften

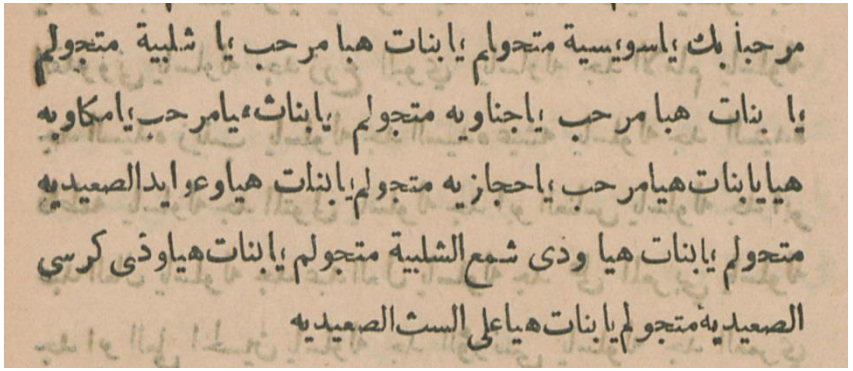


مرحبأ بك :أسورسية متحوام يأبنات هبا مرحب وا شلبية متعجولم
يا نات هيا مرحب يأجناوءه متجولم يأبناث#يأمر<بوامكاوبه
هيأيابنات هيام رحب ياحجاز به متجواليرالنات هياوءو ابدالصميديه
متدو م يالنات هيا ودى تشممالشلبسة متجولم يأناتهياود
الصعيديةمتجو لجرا بنات هياعل الس ث الصعيديه

Quelle: Darwiš, Kāmal Afandī: Kitāb bidaʿ al-faǧār fī ḥaflat az-zār, Kairo 1911, urn:nbn:de:gbv:3:5-86614, CC-BY-SA 4.0.

Training für besondere Schriften

- in ersten Tests erreichte Tesseract auch beeindruckende Ergebnisse für als schwierig einzustufende Schriften, insb. Arabisch und Persisch
  - Titel ab ca. 2. Hälfte des 20. Jhd. erreichten eine Genauigkeit von ca. 95%
  - Titel aus der ersten Hälfte des 20. Jhd. erreichten mit den Trainingsdaten von Tesseract 85-90%
- der neue Digitalisierungsworkflow kommt damit direkt unserem FID Nahost zugute
- im Rahmen der aktuellen FID-Förderphase ist zudem Training für schlechter zu lesende Schriftarten geplant
- Die hieraus gewonnen Erkenntnisse werden natürlich auch für weniger exotische Schriften nutzbar sein



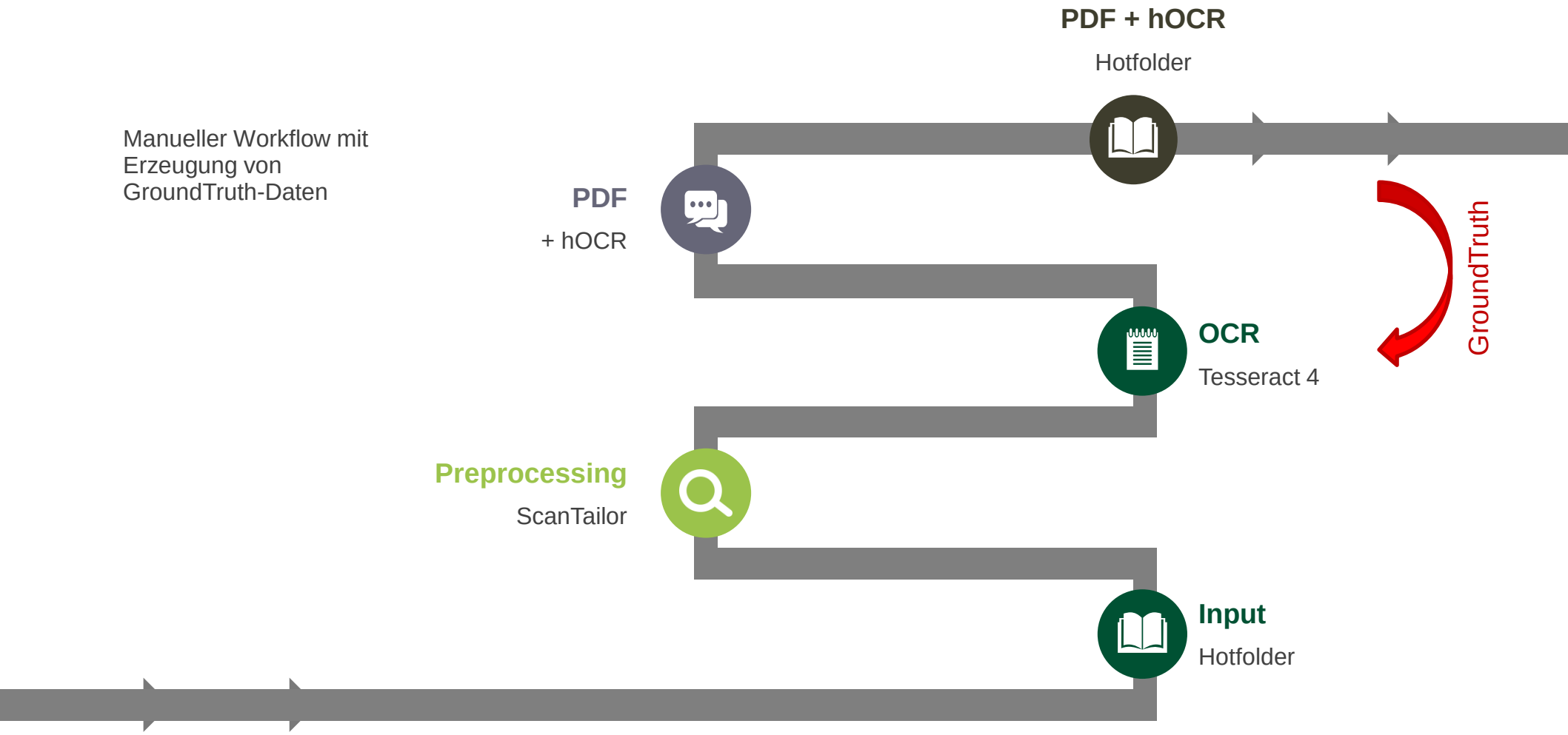
مرحبأ بك :أسورسية متحوام يأبنات هبا مرحب وا شلبية متعجولم
يا نات هيا مرحب يأجناوءه متجولم يأبناث#يأمر<بوامكاوبه
هيأيابنات هيام رحب ياحجاز به متجواليرالنات هياوءو ابدالصميديه
متدو م يالنات هيا ودى ششممالشلبسة متجولم يأناتهياود
الصعيديةمتجو لجرا بنات هياعل الس ث الصعيديه

Quelle: Darwiš, Kāmal Afandī: Kitāb bidaʿ al-faǧār fī ḥaflat az-zār, Kairo 1911, urn:nbn:de:gbv:3:5-86614, CC-BY-SA 4.0.

ج ح خ چ خ – ب ت ث ن پ ي ث ي – ي ب
س ش ش ز ر ذ د ز ژ - ص ض ط ظ

Manuelle OCR mit Training

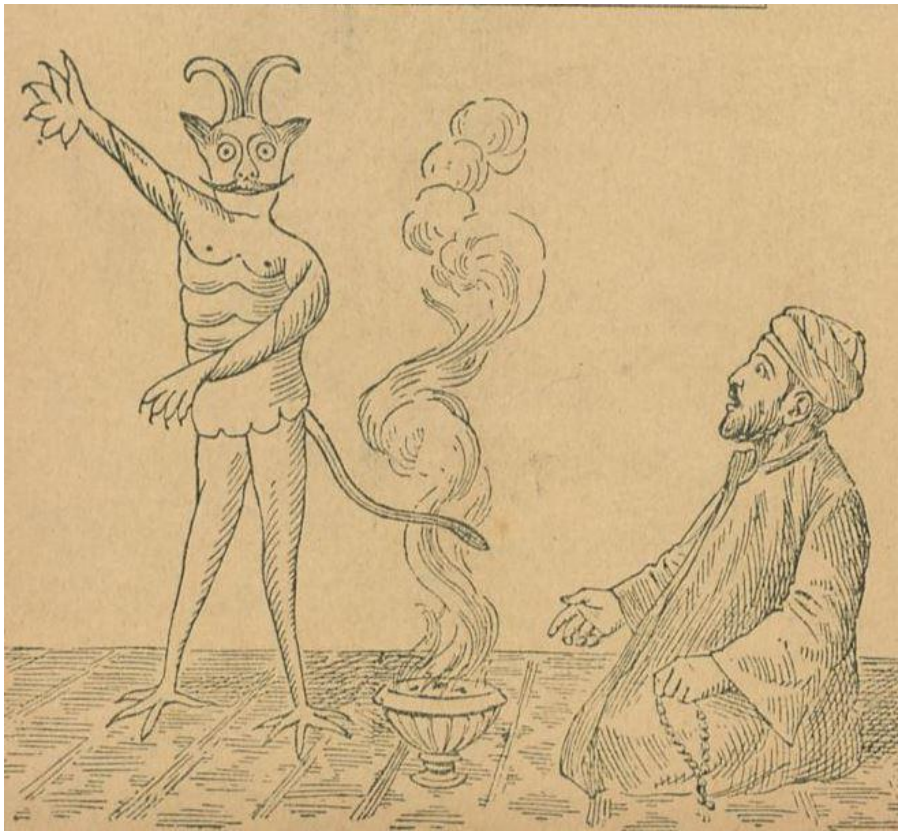
Manueller Workflow mit Erzeugung von GroundTruth-Daten



… Wir haben noch eine Menge Arbeit vor uns

Quelle: diverse arab. Titel von https://menadoc.bibliothek.uni-halle.de, CC-BY-SA 4.0.

Vielen Dank, haben Sie noch Fragen?



Quelle: Darwiš, Kāmal Afandī: Kitāb bidaʿ al-faǧār fī ḥaflat az-zār, Kairo 1911, urn:nbn:de:gbv:3:5-86614, CC-BY-SA 4.0.