EVALUATION VON OCR-QUALITÄT UNTER BERÜCKSICHTIGUNG BIBLIOTHEKARISCHER METADATEN

Dienstag, 18. November, 15:45 – 16:15 Uhr

uwe.hartwig@bibliothek.uni-halle.de

UNIVERSITÄTS- UND LANDESBIBLIOTHEK SACHSEN - ANHALT

1. FRAGESTELLUNGEN

- **“Welche Qualität hat OCR für den Abschlussbericht von Projekt X?”**
- **“Welche Qualität hat OCR bei VD18 Frakturdrucken?”**
- **“Welche Qualität hat OCR bei Funeralschriften?”**
- **“Welche Qualität hat OCR bei RAHBAR-Periodika aus Teheran 1930er Jahre?”**

=>

Wie wird Qualität gemessen?
Wie kann man “Sinn” aus einer Menge von Messwerten gewinnen?

uwe.hartwig@bibliothek.uni-halle.de

2. DATA AND GROUNDTRUTH

Digitale Sammlungen ULB Sachsen-Anhalt

- **Share_it: ca. 75.000 / 75 % OCR**
  (https://opendata.uni-halle.de/handle/1981185920/31823)
- **Share_DIGit: 295.000 / 90 % OCR**
  (https://opendata2.uni-halle.de/handle/1516514412012/1)

GT-Daten

- **VD18 ODEM: 1.600 Seiten**
  (http://github.com/ulb-sachsen-anhalt)
- **Zeitungsdigitalisierung (intern): 100 Seiten**
- **FID MENA arab/pers (intern): 100 Seiten**
- **FID MENA Retrodig. (intern): 200 Seiten**

uwe.hartwig@bibliothek.uni-halle.de

3. MEASURING OCR QUALITY

Zeitrahmen	Kontext	Vorgehen
2013-18	Pilotierung Zeitungsdigitalisierung	methodisch Bernoulli-Experiment
2019-22	Zeitungsdigitalisierung	methodisch GT (Editor / Transkribus-SWT)
2020	FID MENA arabisch	subjektiv GT (Texteditor)
2022-23	VD18 OCR-D Phase 3	sampling GT (Transkribus-SWT + OCR-D)
2023-25	FID MENA persisch	subjektiv/sampling GT (Transkribus-SWT / eScriptorium)
2025-26	FID MENA Retrodigitalisate	sampling GT (OCR4all / eScriptorium)
2026-27	VD16/17/18 Alvensleben	sampling GT

uwe.hartwig@bibliothek.uni-halle.de

4. BIBLIOTHEKARISCHE METADATEN

OAI => METS/MODS

- **Deskriptive Merkmale (MODS)**
  - Sprache => `mods:language`
  - `Genre => mods:genre`
  - `Veröffentlichung`
    - `Ort => mods:originInfo/mods:place`
    - `Zeitraum => mods:originInfo/mods:dataIssued`
    - `Datum => mods:part`
- **Strukturmerkmale (METS)**
  - `Strukturtypen => mets:div@TYPE, mets:div@LABEL`

uwe.hartwig@bibliothek.uni-halle.de

5. SAMPLING A CORPUS

Vorraussetzung: eindeutige Identifizierbarkeit

- **Wie?**
  - Datenharvesting über Web (`OAI + METS/MODS`)
  - Pool aller Seiten aller Vorgänge eines OAI-Sets
  - Auswahl rand. Seiten aus diesem Pool
  - GT-Erstellung
  - Verknüpfung Seiten mit entsprechenden deskriptiven Metadaten (`METS/MODS`)

(https://github.com/ulb-sachsen-anhalt/ulb-groundtruth-eval-odem-ger)

uwe.hartwig@bibliothek.uni-halle.de

6. EVALUATION IST-ZUSTAND

`digital-eval`
(https://github.com/ulb-sachsen-anhalt/digital-eval)

- Abgleich Kandidatenmenge K mit Referenzdaten GT
  textbasierte OCR-/IR-Metriken
  (*keine* Visualisierung => dinglehopper)
- Deskriptive Statistik für Gesamt-u. Teilmengen => Sets

  ABER
  Sets implizit statisch über Dateistruktur vorgegeben

Set (Ordner) => Unterordner (Teilset) => …
Zeitung => Jahrgang => Seite Ausgabe
ODEM => Sprache

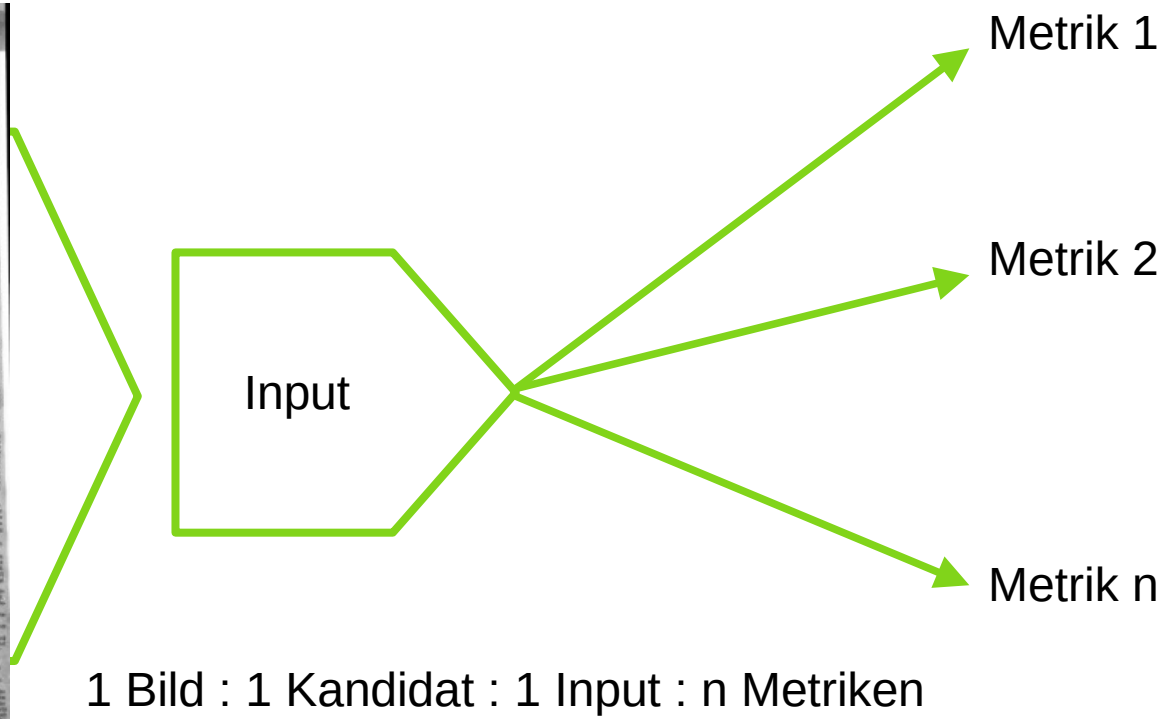
(https://github.com/ulb-sachsen-anhalt/ulb-groundtruth-eval-odem-other)

uwe.hartwig@bibliothek.uni-halle.de

7. EVALUATION IST-ZUSTAND

`digital-eval`

(https://github.com/ulb-sachsen-anhalt/digital-eval)

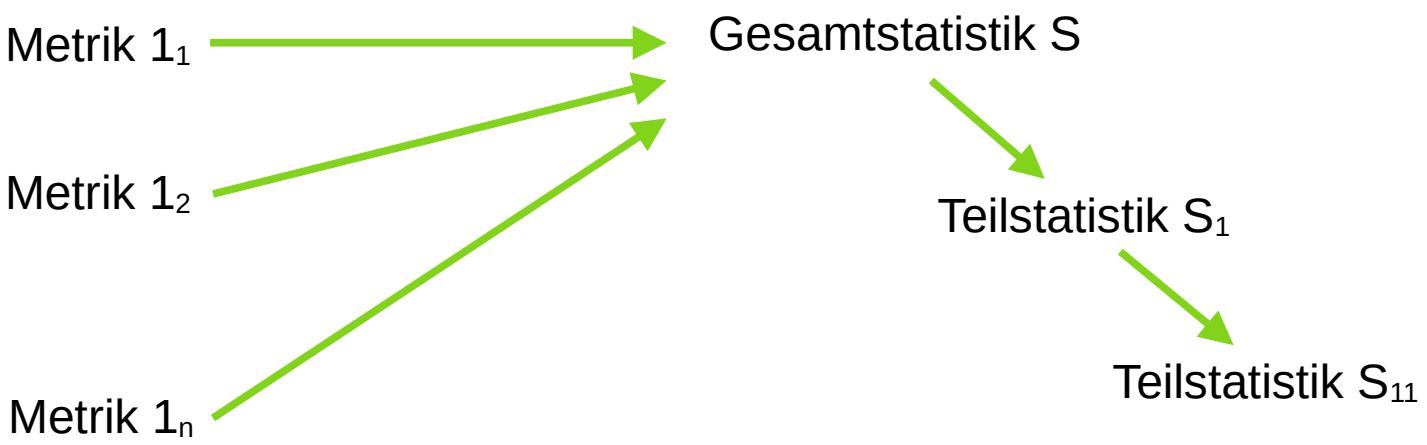


1 Bild : 1 Kandidat : 1 Input : n Metriken

uwe.hartwig@bibliothek.uni-halle.de

8. EVALUATION IST-ZUSTAND

digital-eval
(https://github.com/ulb-sachsen-anhalt/digital-eval)



uwe.hartwig@bibliothek.uni-halle.de

9. EVALUATION SOLL-ZUSTAND

- Statische Sets => dynamische Sets
- deskriptive Metadaten + Strukturinformationen => Facetten

- VD16/17/18 Alvensleben Evaluation dynamischer Sets
  Zeiträume, Sprachen
- Zeitungen
  Erscheinungsdatum, Publikationsort

- GT-Strukturinformationen nutzen?
  1 Eingangsbild => n Eingangsregionen

uwe.hartwig@bibliothek.uni-halle.de