



Spezifikationen und Lessons Learned

12.02.2020

Matthias Boenig



Übersicht

- Prämissen
 - Standards, Modul, Definition, Open Source
- Spezifikationen
 - Formatdefinitionen
 - Ground Truth, METS, PAGE, OCRD-ZIP (Bagit)
 - Schnittstellen und Dokumentation
 - CLI, ocrd-tool.json



Prämissen

- Standards
 - XML, METS/MODS, ALTO, PAGE
- Modul
 - OCR: Summe aus verschiedenen einzelnen Prozessen
- Definition
 - Sicherung und Sicherheit der Kompatibilität der Module (Prozessoren) untereinander
 - Möglichkeit der Weiterentwicklung der OCR-D Software
- Open Source
 - uneingeschränkte Nutzung sowie Lizenz
 - öffentlicher Zugang



Spezifikation : Ground-Truth-Daten

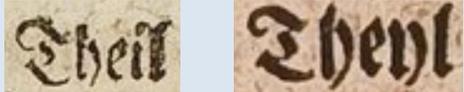
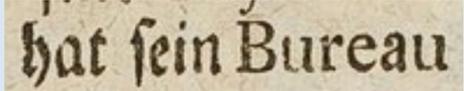
4

- GT Bereich: **Texterkennung**
 - zeilen- und wortweise Transkription
 - Texte mit Fraktur und Antiqua (Zeitraum 16.–19. Jahrhundert)
 - Sprachen: Deutsch, Latein, Griechisch, Hebräisch
- GT Bereich: **Layouterkennung**
 - semantische Auszeichnung der Dokument- und Seitenstruktur
 - u.a. Kapitelüberschriften, Marginalien, Fußnoten
- GT Bereich: **Störungen**
 - Verunreinigungen, Beschädigungen
 - Provenienzspuren



Problembereich: Datenkonsistenz

5

Vorlage	vorlagengengetreu	normalisiert
	Theil, Theyl	Teil
	hat fein Bureau	hat sein Büro
	□c.	etc.



Spezifikation : Ground-Truth-Daten

6

Korpus	Umfang Zeilen	Umfang Strukturen (Region)
Gt_daten_01	6014	1379
Gt_daten_02	9840	1435
Korpus	Umfang Seiten	Umfang Strukturen (Region)
Struktur	1.100	6.846
Dokument-Struktur	180.000	ca. 1.1 Mio



Spezifikation : Ground-Truth-Daten

- **Nach-Nutzung** von Referenzdaten, Korpora
 - Daten aus: Deutsches Textarchiv, InternetArchiv, Digitalisierungsprojekten
- Schwierigkeit:
 - **Transformation** von spezifischen Projektkontexten
 - Dienstleister
 - Format-**Differenzen** (Transkribus PAGE \neq Aletheia PAGE)
 - Heterogenität der GT-Ansprüche $GT = \{\text{Zeichen, Wort, Zeile, Region, Dokument}\}$
 - Definition der **Repräsentativität** und **Quantität**



Formatdefinition: OCR-D *Ground-Truth-Guidelines*

8

- Inhalte:
 - **Text:** Richtlinien zur Transkription der Volltexte für die Nutzung als Ground Truth
 - **Layout und Struktur:** Richtlinien zur Erfassung des Layouts und der Struktur von gedruckten Texten für die Nutzung als Ground Truth
 - **Repräsentation:** Dokumentation zum *PAGE XML Format for Page Content*
- Format:
 - englisch/deutsch
 - DITA : themenfokussierte Gliederung (sog. Topics)
- <https://ocr-d.github.io/en/gt-guidelines/trans/>
- <https://ocr-d.github.io/de/gt-guidelines/trans/>



Formatdefinition: METS

- seit **2004** Etablierung des **Metadata Encoding and Transmission Standard** (METS) bei der Erfassung von Metadaten im Zug der Digitalisierung
- **Zugang** und **Ergebnis** der Volltextdigitalisierung muss sich im METS-Datensatz widerspiegeln
- **OCR-D Empfehlungen** für:
 - Dokumentation der einzelnen Workflow-Schritte
 - Namensschema von **mets:fileGrp** (`<mets:fileGrp USE="OCR-D-IMG">`)
- <https://ocr-d.github.io/en/spec/mets>



Formatdefinition: PAGE

- seit **2009** definiert
- **XML** basiertes Schema
- Dokumentation und Beschreibung eines digitalen Volltextes auf **verschiedenen Erfassungsebenen** (Zeichen, Wort, Zeile, Region) sowie Möglichkeit einer allgemein **typographischen Klassifizierung** der Regionen
- Transformation von PAGE nach **ALTO**
- <http://ocr-d.de/de/spec/page>



Formatdefinition: OCRD-ZIP (Bagit)

- definierter Transport-Container
 - data (Verzeichnis Daten + METS-Datei)
 - Metdatendaten zum Bagit
 - tagmanifest-sha512.txt
 - manifest-sha512.txt
 - bagit.txt
 - bag-info.txt
- Verwendung: GT-Transport-Container, LZA-Container
- <https://github.com/LibraryOfCongress/bagger>
- https://ocr-d.github.io/en/spec/ocrd_zip



Schnittstellen: CLI

- Definition von **einheitlichen Parametern** für die Steuerung der modulbasierten Workflow-Schritte
- Input ↔ Output
 - I --input-file-grp
 - O --output-file-grp
 - m **--mets**
 - w --working-dir
 - g --page-id
 - p --parameter
 - l --log-level
 - J --dump-json



Schnittstellen: CLI

```
$> ocrd-olena-binarize \  
  --mets "path/to/file/mets.xml" \  
  --working-dir "path/to/workingDir/" \  
  --parameters '{"impl": "sauvola"}' \  
  --page-id PHYS_0001,PHYS_0002,PHYS_0003 \  
  --input-file-grp OCR-D-IMG \  
  --output-file-grp OCR-D-IMG-BIN-KRAKEN
```



Technische Dokumentation: ocrd-tool.json

- **normierte** Anwendungsbeschreibung auf Basis von JSON (JavaScript Object Notation)
- das **Modul/Tool** muss mit einer ocrd-tool.json **beschrieben** werden
 - Allgemeine Informationen (type, description...)
 - Beschreibung des Tools (input_file_grp, output_file_grp)
- ocrd-tool.json **steuert** die CLI.
- http://kba.cloud/de/spec/ocrd_tool



Fazit

- mit der OCR-D-Software: **belastbare** Spezifikationen auf Basis
 - Standards, Modul, Definition, Open Source
- **Konsolidierung** der Spezifikationen
 - Empfehlungen für die **DFG: Praxisregeln Digitalisierung**