



Ground Truth: Grundwahrheit oder
Ad-Hoc-Lösung?

Wo stehen die Digital Humanities?

*Matthias Boenig, Maria Federbusch, Elisa Herrmann,
Clemens Neudecker, Kay-Michael Würzner*

DHd Konferenz 2018, Köln, 26. Februar - 2. März 2018



Prämissen des Vortrags

Die Digital Humanities soll

- eine **empirische** Disziplin sein
 - transparent erhobene Datengrundlage
 - klar umrissene, wissenschaftlich anerkannte Auswertungsmethoden
 - Interpretation der Ergebnisse innerhalb eines theoretischen Modells
- **reproduzierbare** Forschungsergebnisse liefern
 - frei verfügbare Methoden
 - frei verfügbare Daten und Surrogate
 - frei verfügbare wissenschaftliche Veröffentlichungen



Digital Humanities im Vergleich

Digital Humanities	Kognitionswissenschaft/ Leseforschung	Machine Learning/ Dokumentanalyse	
empirisch			
???	+	+	transparente Datengrundlage
-	+	+	anerkannte Auswertungsmethode
-	+	+	theoretisches Modell
reproduzierbar			
+	+	-	frei verfügbare Methoden
+	-	+	frei verfügbare Daten
+	-	-	freie Veröffentlichungen



Was ist Ground Truth?

- »fundamentale, nicht weiter beweisbare wahrheit, auf deren evidenz sich andere wahrheiten, sätze, lehren usw. aufbauen«
[<http://www.woerterbuchnetz.de/DWB?lemma=grundwahrheit>]
- fehlerfreie, manuell erhobene Daten
- jeweils mit Bezug zur Aufgabenstellung, z.B. für OCR
 - Text
 - Bild
 - Text-Bild-Verknüpfung
 - ...



Aspekte der Verwendung von *Ground Truth*

- Trainingsdaten
 - automatische Induktion statistischer Modelle
- Evaluationsdaten
 - Überprüfung von Annotationsergebnissen
 - Qualitätsmessungen
- Referenzdaten
 - Repräsentation einer Grundgesamtheit
 - Grundlage von Analysen und Problembeschreibungen



Problembereich Ground Truth in den DH

- mangelhafte Dokumentation der Datenerhebung
 - unterschiedliche, nicht-standardisierte Formate
 - fehlende Metadaten
 - unzureichende Transkriptions-/Annotationsrichtlinien
- Umfang der verfügbaren Daten
 - aus unterschiedlichen Datensätzen kompilierte Daten
 - maschinell erzeugte Daten als GT



OCR-D *Ground-Truth-Guidelines*

- 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 der Dokumentation: **DITA**
 - themenfokussierte Gliederung (sog. Topics)
 - Export in verschiedene Formate
- Operationalisierung durch Validierung
 - Schema **und** Schematron

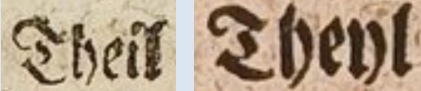



Ziel: Erweiterung des frei zugänglichen Ground-Truth-Bestandes für die Text- bzw. Strukturerkennung von historischen Drucken

1. Format-Dokumentation von vorhandenem Ground Truth
2. Handlungsanweisung für die Ground-Truth-Erstellung
3. automatische Überprüfbarkeit der Validität von Ground Truth
4. Schaffung eines Ground-Truth-Repositoryms
5. Integration bzw. Vereinheitlichung heterogener Datenbestände



Problembereich: Datenkonsistenz

Vorlage	vorlagengengetreu	normalisiert
	Theil, Theyl	Teil
	Bureau	Büro
	c.	etc.



Vergleich: DTA-Basisformat - OCR-D GT-Guidelines

DTA-Richtlinien	OCR-D GT-Guidelines	
DTA-Basisformat	PAGE-XML	Format
aufgespalten	verschiedene Level	Ligaturen (Zeichenebene)
Fußnoten werden direkt an die Stelle im laufenden Text gesetzt, von der aus sie referenziert werden.	Fußnoten erscheinen im unteren Seitenbereich des Satzspiegels.	Fußnoten (Strukturebene)



- **Transkription:** Interpretation von einzelnen typographischen und graphematischen Phänomenen
 - Definition und Dokumentation von Interpretationsspielräumen
 - unterschiedlichen (End-)Nutzeransprüchen gerecht werden
- **Struktur und Layout:** minimale - maximale Erfassung
 - Definition und Dokumentation von Entscheidungen
 - ressourcensparende, pragmatische Vorgaben



Beispiel einer Schematronregel für PAGE-Datei: Prüfung der ct-Ligatur

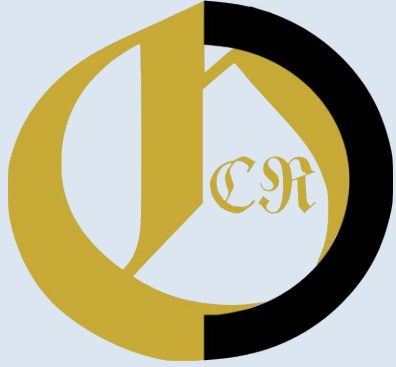
```
- <pattern id="ct_ligatur">
  <let name="x" value="//page:Unicode[text()][contains(., 'ct')]"/>
  - <rule context="//page:Unicode[text()][contains(., 'ct')]">
    - <report test="$x" role="WARNING">
      [W0001] The document contains splitted ligature ct. OCR-D Level 1
    </report>
  </rule>
</pattern>
```



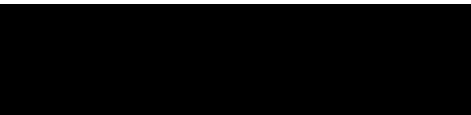
Zusammenfassung

- Nachholbedarf im Bereich Ground Truth in den Digital Humanities
 - Umfang
 - Dokumentation
- Schärfung des Bewusstseins im Umgang mit Ground-Truth-Daten nötig
- OCR-D Ground-Truth-Guidelines als Vorschlag zur standardisierten Erhebung von Ground-Truth-Daten für OCR-bezogene Verfahren
- http://ocr-d.de/gt_guidelines

Danke für Ihre Aufmerksamkeit!



RESTeRampe





Stand Ground Truth in OCR-D : Metadatensatz

15

Administrative Daten

- Dateiname
- ID

Digitalisierungsdaten

- Dateiformat
- Bildtyp

Struktur- und OCR-Problembereiche

Die Struktur- und OCR-Problembereiche sind in vier Kategorien eingeteilt:

Kategorien

Typographie/Layout

Fontmix (Absatzebene)

Fontmix (Wortebene)

Fontmix (Zeichenebene)

Schriftgrößenmix

Sprachmix

Langes s/Zeichen

komplexes Layout

Abbildung

Fußnoten/Marginalien

Sonderzeichen

Ligaturen

Zierbuchstaben

Fettschrift/Normschrift-Wechsel

Zahlen

Produktionsfehler

Tintendurchdruck

Benutzung

Flecken und Annotationen

Knicke/Falten

Digitalisierung

Verzerrung

Farbsprung

Variierende Kontrastverhältnisse

Gespiegelte Buchseite

vgl. die Kategorien von intrinsischen und extrinsischen Probleme nach Christoph Stollwerk:

Machbarkeitsstudie zu Einsatzmöglichkeiten von OCR-Software im Bereich "Alter Drucke" zur Vorbereitung einer vollständigen Digitalisierung deutscher Druckerzeugnisse zwischen 1500 und 1930. Göttingen 2016. S. 34-36

[<http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dw-p-2016-16.pdf>]



“**Uns stört**, dass diese Vorgehensweise impliziert, dass man die uneinheitliche Makrostruktur des Kochbuchs der Henriette Davidis in [der] digitale[n] Edition schlicht so übernommen hat, **wie man sie vorfindet.**” Norbert Luttenberger, Jesper Zedlitz Torsten Knauf 2017: Krümelmonsters Kochbuch – einige Gedanken zur domänenspezifischen Edition von Gebrauchstexten in: INFORMATIK 2017, Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn.

“Die [DTA] **Erfassung** erfolgt nach dem **Prinzip der Wahrung** des historischen Sprachstandes der Texte. [Es] wird darauf geachtet, bei der Texterfassung die Zahl der (unvermeidbaren) **Interpretationen** typographischer Gegebenheiten **gering zu halten.**” Dokumentation des DTA-Basisformates: <http://www.deutschestextarchiv.de/doku/basisformat/transkription.html>



Gegenüberstellung



Gegenüberstellung eines Modells eines Forschungsprozesses aus der empirischen Sozialforschung und dessen Anwendung auf den Prozess der automatischen Texterfassung.

Siehe dazu: **Schnell, Rainer / Hill, Paul B. / Esser, Elke** (2011): Methoden der empirischen Sozialforschung. 9., aktualisierte Aufl. München: Oldenbourg



Ziele dieses Vortrags

18

1. Die Datengrundlage des Ground Truth zur Induzierung von Modellen sollte sich in einem ähnlichen Verhältnis zu den auftretenden Phänomenen der zu untersuchenden Daten verhalten.
2. Eine Akklamation von Daten als Ground Truth reicht nicht aus. Es bedarf eines Ground-Truth-Konzepts. Dieses Konzept ist zu erarbeiten, zu unterstützen und mit diesem ist verpflichtend zu arbeiten.
3. Ein Referenzdaten-System ist zu schaffen.
 - a. Referenzdaten
 - b. Richtlinien zur Erfassung von Referenzdaten
 - c. persistente Referenzsysteme



“Literarische Texte mit Unterstützung von Informationstechnologie zu bearbeiten heißt, literaturwissenschaftliche Methoden teilweise **Computerprogrammen** zu übertragen.”

Christine Ivanovic: Die Vernetzung des Textes: Im Möglichkeitsraum digitaler Literaturanalyse. In: Zeitschrift für digitale Geisteswissenschaften. 2017. text/html Format. DOI: [10.17175/2016_010](https://doi.org/10.17175/2016_010)

Bildnachweis:

Das Dunkle in der Black Box : Die Maschine

http://www.deutschlandfunk.de/das-dunkle-in-der-black-box-die-maschine.740.de.html?dram:article_id=406190 [2017]



Administrative Daten

- Dateiname
- ID

Digitalisierungsdaten

- Dateiformat
- Bildtyp

Struktur- und OCR-Problembereiche

Die Struktur- und OCR-Problembereiche sind in vier Kategorien eingeteilt:

Kategorien

Typographie/Layout

Fontmix (Absatzebene)

Fontmix (Wortebene)

Fontmix (Zeichenebene)

Schriftgrößenmix

Sprachmix

Langes s/Zeichen

komplexes Layout

Abbildung

Fußnoten/Marginalien

Sonderzeichen

Ligaturen

Zierbuchstaben

Fettschrift/Normschrift-Wechsel

Zahlen

Produktionsfehler

Tintendurchdruck

Benutzung

Flecken und Annotationen

Knicke/Falten

Digitalisierung

Verzerrung

Farbsprung

Variierende Kontrastverhältnisse

Gespiegelte Buchseite

vgl. die Kategorien von intrinsischen und extrinsischen Probleme nach Christoph Stollwerk:

Machbarkeitsstudie zu Einsatzmöglichkeiten von OCR-Software im Bereich "Alter Drucke" zur Vorbereitung einer vollständigen Digitalisierung deutscher Druckerzeugnisse zwischen 1500 und 1930. Göttingen 2016. S. 34-36

[<http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dw-p-2016-16.pdf>]



Grenzen definieren

- “Eines der wesentlichen Probleme dabei ist die Wahl der geeigneten Granularität. Die Granularität muss gröber sein, als die in der Korpuslinguistik angewandte (typische Einheiten wären Wort oder Satz); sie sollte feiner sein als die etablierten literaturwissenschaftlichen Kategorien (typische Einheiten wären Gattungen oder Schreibweisen).“

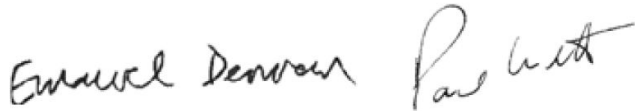
Christine Ivanovic: Die Vernetzung des Textes: Im Möglichkeitsraum digitaler Literaturanalyse. In: Zeitschrift für digitale Geisteswissenschaften. 2017. text/html Format. DOI: [10.17175/2016_010](https://doi.org/10.17175/2016_010)

MODELERS OF ALL MARKETS, UNITE!

You have nothing to lose but your illusions.

The Modelers' Hippocratic Oath

- ~ I will remember that I didn't make the world, and it doesn't satisfy my equations.
 - ~ Though I will use models boldly to estimate value, I will not be overly impressed by mathematics.
 - ~ I will never sacrifice reality for elegance without explaining why I have done so.
 - ~ Nor will I give the people who use my model false comfort about its accuracy.
- Instead, I will make explicit its assumptions and oversights.
- ~ I understand that my work may have enormous effects on society and the economy, many of them beyond my comprehension



Emanuel Derman
January 7 2009

Paul Wilmott
January 7 2009

Vorschlag eines Hippokratischen Eides
als Reaktion auf die Finanzkrise ab 2007 von dem
Wirtschaftswissenschaftler Emanuel Derman und
dem Mathematiker Paul Wilmot zum
verantwortungsvollen Umgang mit Modellen

siehe:

<https://www.soa.org/Library/newsletters/risk-management-newsletter/2009/september/jrm-2009-iss17-derman.pdf>



Definition Ground Truth

Mid 19th century; earliest use found in Henry Ellison (1811–1890). From ground + truth [adverb], in sense 1 probably after German Grundwahrheit.

Unter Ground Truth wird in diesem Kontext die Dokumentation ausgewählter Merkmale (Zeichen, Zeilen, Absätze, Spalten, Abbildungen, usw.) des Textes in Form einer digitalen Transkription verstanden. Dabei ist je nach Anwendung zwischen allgemeineren Referenz- und spezifischeren Trainingsdaten zu unterscheiden. Zum Ground Truth gehört neben dem digitalen Text das Bilddigitalisat

Forschungsprozesses	
Auswahl des Forschungsproblem	
Theoriebildung	
Konzeptspezifikation	Bestimmung der Untersuchungsform
Operationalisierung	
Auswahl der Untersuchungseinheiten	
Datenerfassung	
Datenanalyse	
Publikation	

Evaluation der Forschungs-
und Erkennungs-
ergebnisse

Texterfassung	
Frage nach den Textvorlagen	
Automatische zeichenbasiert, /Manuelle Erfassung z. B. double Keying	
Transkriptionsrichtlinien, Schematron-Regeln	Implementierung des Erfassungswflows
Korpusfestlegung, Erfassungstiefe	
Text- und Strukturerkennung	
Qualitätsprüfung	
Datenbereitstellung	