

# Ground Truth: Grundwahrheit oder Ad- Hoc-Lösung? Wo stehen die Digital Humanities?

## Boenig, Matthias

boenig@bbaw.de  
Berlin-Brandenburgische Akademie der Wissenschaften,  
Deutschland

## Federbusch, Maria

maria.federbusch@sbb.spk-berlin.de  
Staatsbibliothek zu Berlin Preußischer Kulturbesitz,  
Deutschland

## Herrmann, Elisa

herrmann@hab.de  
Herzog August Bibliothek Wolfenbüttel, Deutschland

## Neudecker, Clemens

clemens.neudecker@sbb.spk-berlin.de  
Staatsbibliothek zu Berlin Preußischer Kulturbesitz,  
Deutschland

## Würzner, Kay-Michael

wuerzner@bbaw.de  
Berlin-Brandenburgische Akademie der Wissenschaften,  
Deutschland

## Einleitung

Die Verwendung von Referenzdaten für das Training und die Auswertung statistischer Annotations- und Analyseverfahren ist ein Kernmerkmal empirischer Forschung, zu der auch die Digital Humanities zählen möchten.<sup>1</sup> Die wichtigste Grundlage für den erfolgreichen Einsatz statistischer Verfahren liegt in der Verwendung geeigneter, den Algorithmen zugrunde liegender Modelle. Für deren Erstellung ist neben einem passenden Lernverfahren das Vorhandensein von Trainingsdaten eine wesentliche Voraussetzung. Werden Forschungsdaten, die mit quantitativen Methoden entstanden sind, mit Referenzdaten verifiziert und interpretiert, ist ein kritischer Blick auf Auswahl, Erstellung und Umgang mit selbigen ein häufig vernachlässigter Bereich.

Vor diesem Hintergrund richtet der vorliegende Beitrag einen kritischen Blick auf die Rolle von und den Umgang mit Ground-Truth-Daten im Bereich der Digital Humanities. Drei Beobachtungen sollen zur Diskussion gestellt werden:

1. Es fehlt den Digital Humanities an einheitlichen und etablierten Ground-Truth-Datensets für die Evaluierung von Forschungsergebnissen auf Basis quantitativer Methoden.
2. Es fehlt den Digital Humanities an Richtlinien zur Erstellung und Verfahren zur Verifizierung von Ground Truth.
3. Es fehlt den Digital Humanities an akzeptierten und operationalisierbaren Metriken zur Qualitätsbestimmung von Ground Truth und abgeleiteten Datenanalysen.

Anhand der automatischen Texterfassung, die als Analogie für ein allgemeines Modell eines empirischen Forschungsprozesses<sup>2</sup> gesetzt wird, sollen im Folgenden die Probleme diskutiert und Möglichkeiten gezeigt werden, wie diesen Defiziten begegnet werden kann.



Abbildung : Gegenüberstellung eines Modells eines Forschungsprozesses aus der empirischen Sozialforschung<sup>3</sup> und dessen Anwendung auf den Prozess der automatischen Texterfassung

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.

Die Volltext-Digitalisierung von Archiv- und Bibliotheksbeständen, größeren Dokumentsammlungen oder Korpora wird heute von unterschiedlichen Seiten verfolgt: So werden beispielsweise seit 2005 massenhaft Bibliotheksbestände von Google im Rahmen von öffentlich-privaten Partnerschaften sowohl als Bild als auch als Text digitalisiert. Daneben unterstützen Stiftungen, Fördereinrichtungen wie die DFG sowie die Haushaltsmittel der Institutionen die Digitalisierungen im Rahmen spezifischer Projekte. Im Ergebnis dieser Digitalisierungsbemühungen stehen Volltextsammlungen höchst unterschiedlicher Qualität, Vollständigkeit, Interoperabilität und Nachnutzbarkeit.

Mit der OCR-D-Initiative wird erstmals versucht, die technischen und organisatorischen Grundlagen dafür zu schaffen, einen breiten heterogenen Bestand von Drucken aus dem 16. – 18. Jahrhundert vollständig und einheitlich in elektronischen Volltext umzuwandeln und frei zur Verfügung zu stellen. Unter Einbeziehung einzelner Modulprojekte wird vom DFG-geförderten Koordinierungsprojekt<sup>4</sup> die Transformation der Drucke in strukturierten Volltext konzeptuell und prototypisch vorbereitet. Dazu werden im Rahmen von OCR-D

Anwendungen, Adaptionen und Weiterentwicklungen von Verfahren der Optical Character Recognition (OCR) für historische Drucke geprüft bzw. implementiert und in einer finalen, prototypischen Produktionsumgebung kombiniert. Eine zentrale Aufgabe von OCR-D besteht dabei in der Bereitstellung eines umfassenden Ground-Truth-Korpus, das sowohl Referenz- als auch Trainingsdaten sowie Richtlinien zur Transkription von Texten für deren Verwendung als Ground Truth umfasst. Damit können sowohl Texte bezüglich ihrer Zeichengenauigkeit transparent geprüft als auch spezielle statistische Modelle für die Text- und Strukturerkennung trainiert werden.

## Diskussion

1. Gegenstand der Digital Humanities ist das digitale Objekt, beispielsweise digitaler Text.<sup>5</sup> Vergleicht man den Bereich der automatischen Texterfassung mit einem in der empirischen Sozialforschung etablierten Modell des Forschungsprozesses, wird deutlich, dass die Referenzdaten in beiden Prozessen, im Besonderen in der Phase der Evaluation nach Abschluss der Datenanalyse, eine bedeutende Rolle spielen. In dieser Phase wird auf entsprechende Referenzdaten oder -systematiken zurückgegriffen, die in der Phase der Theoriebildung identifiziert und angesammelt wurden. Anzumerken ist, dass der Forschungsprozess nicht isoliert zu betrachten ist, da er schon zu Beginn bei Formulierung und Auswahl des Forschungsproblems auf vorhandene Forschungsdaten zurückgreift. Der traditionelle Forschungsprozess hat durch ein System von Referenzdaten (u. a. Wörterbücher, Editionen, Nachschlage- und Quellenwerke) ein System der *referenzbasierten* Evaluation geschaffen. Dieses System wird gestützt durch Konventionen der Zitierung, Dokumentation, Verzeichnung und Aufbewahrung in Institutionen sowie spezifischen Publikationsformen. Der Forschungsprozess in den Digital Humanities kann auf diesen Hintergrund nur teilweise bzw. gar nicht zurückgreifen, da bisher zu wenige digitale Daten vorliegen.

Auf der anderen Seite haben die Digital Humanities bezüglich der Verfügbarmachung von Quellcode und Forschungsdaten einen großen Vorsprung gegenüber anderen datenbasiert arbeitenden Disziplinen (etwa der kognitiven Psychologie). *Reproducible Science* wird sowohl gefördert als auch propagiert. Problematisch sind aber die fehlende Vereinheitlichung und Transparenz bei der Datenerhebung sowie der Einsatz mangelhaft erfasster Daten bzw. deren ad-hoc Surrogaten bedingt durch die mangelnde Verfügbarkeit an (insbesondere annotierten) Forschungsdaten.

Die Arbeitsweise mit Referenzdaten, wie sie in den Naturwissenschaften und Teilen der Geisteswissenschaft Anwendung findet, möchte neben der Evaluation, Verifikation gerade die Vergleichbarkeit der Forschungsergebnisse stützen. Das setzt voraus, dass diese Daten in ihrer Qualität diesen Ansprüchen

genügen müssen und durch entsprechende Normungen Interpretationsspielräume definiert sind.

Trotz des wissenschaftlichen Anspruches der Digital Humanities auf Objektivität und dem Bemühen Referenzdaten zu schaffen, werden unterschiedliche Interpretationen, die auf Grundlage von mangelhaften Referenz- und Trainingsdaten entstehen, möglich. Solch ein „hinzunehmendes Übel“ wird erkannt und mit „Pragmatismus“, mit der „Flexibilität“ oder „Austauschbarkeit von Konzepten im Konkreten der Texte“ gerechtfertigt<sup>6</sup>, eine Vergleichbarkeit der so entstandenen Texte bzw. Forschungsergebnisse damit aber wesentlich behindert. Um eine Vergleichbarkeit im Sinne von *Reproducible Science* zu erreichen, ist somit der Gegenstand der Digital Humanities um die Fragen der Objekterstellung zu erweitern.

2. Mit dem Begriff OCR wird üblicherweise der Gesamtprozess der automatischen Texterfassung bestehend aus den Teilaufgaben Bildvorverarbeitung, Struktur- und Texterkennung sowie gegebenenfalls (automatisierter) Nachkorrektur bezeichnet. Damit eine OCR vorgenommen werden kann, sind entsprechende Erkennungsmodelle notwendig. Diese werden durch sogenanntes Training unter Nutzung von Ground Truth induziert. Generische Modelle werden vom Anbieter der OCR-Software zur Verfügung gestellt, domänenspezifische Modelle müssen trainiert werden. Das Training dient immer der Verbesserung der Endergebnisse des Erkennungsprozesses. Somit sollten bei einem universellen Anspruch Ground-Truth-Daten sowohl spezifische als auch allgemeine Normungen enthalten, damit sowohl generische als auch spezifische Modelle trainiert werden können. Ziel ist es, die unterschiedlichen Bedürfnisse und Schwerpunkte in den Digital Humanities zu bedienen.


3. Um diesem sehr breiten Anspruch gerecht zu werden, reicht eine Akklamation, dass diese oder jene Sammlung von Daten als Ground Truth bezeichnet und genutzt werden kann, nicht aus. Es bedarf hingegen eines Ground-Truth-Konzepts. Dieses Konzept dokumentiert sowohl dessen inneren Aufbau als auch dessen Erstellung. Damit können im Bereich der Texterfassung beispielsweise folgende Punkte im Umgang mit Ground Truth erreicht werden:

1. (weitere) Reduktion bzw. Standardisierung der Freiheitsgrade bei der Transkription (z. B. langes s, Ligaturen, Zeilenumbruch)
2. weitergehende Operationalisierung der Überprüfbarkeit der Validität der Transkription
3. Ergänzung von (weiteren) Anweisungen zum Umgang mit koordinatenbasierten Phänomenen

Illustriert werden kann ein solches Konzept am Beispiel der *OCR-D Ground-Truth-Guidelines*. Um die Freiheitsgrade von Interpretationen (Punkt A) zu normieren werden drei Level von Erfassungsgraden angeboten. Die einzelnen Level sollen nachvollziehbare Interpretationsentscheidungen sowohl festlegen als auch

dokumentieren und damit die Möglichkeit der maschinellen Überprüfbarkeit eröffnen.

Tabelle : Beispiel der Anwendung der Level bei der Ligatur ct.

Zeichen	Level 1	Level 2	Level 3
	ct Die Ligatur wird in zwei einzelne Zeichen aufgespalten.	ct Die Ligatur wird aufgespalten und mit einer zusätzlichen Annotation, dass es sich um eine Ligatur handelt, im PAGE-Format versehen. textStyle{offset:0; length:2;ligatur:true;}	&#xEEEC5; Die Ligatur wird als ein Zeichen interpretiert und mit dem entsprechenden Unicode-Zeichen wiedergegeben.

Damit werden Anweisung und Richtlinien weitgehend operationalisiert und eine computergestützte Validierung umsetzbar (Punkt B). Das entsprechende Ground-Truth-Korpus liegt im XML-basierten PAGE-Format<sup>7</sup> vor. Das PAGE-Format hat sich im Rahmen des EU-Projekts IMPACT<sup>8</sup> sowie durch seine Verbreitung im Rahmen von Wettbewerben bei wissenschaftlichen Konferenzen (z.B. ICDAR, ICFHR, DAS) als de-facto Standard für XML-basierter Ground Truth etabliert. Mit Hilfe von Schematron<sup>9</sup>-Regeln kann nun, wie das Beispiel zeigt, geprüft werden, nach welchem Level die „ct“-Ligatur kodiert ist:

```
<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>
```

Die Wahl des PAGE-Formates ermöglicht im Unterschied zum TEI-basierten Basisformat des DTA (DTABf)<sup>10</sup> eine unkomplizierte Lösung für die Repräsentation von koordinatenbasierten Phänomenen (Punkt C).

```
<Word id="word_1479724691818_218" language="German" custom="ze
structure {type:signature-mark;} textStyle {offset:0; length:
<Coords points="1111,2184 1037,2184 1037,2123 1111,2123"/>
<TextEquiv>
<Unicode>a</Unicode>
</TextEquiv>
<TextStyle fontFamily="Antiqua" fontSize="26.0"/>
</Word>
```

## Fazit

Ground Truth stellt im Texterkennungsprozess und im dazu betrachteten Forschungsprozess in den Digital Humanities eine entscheidende Rolle dar. Ergebnis dieser beiden Prozesse sind immer Publikationen, die als Forschungsdaten in neuen Forschungszusammenhängen nachgenutzt werden. Gelingt es den Digital Humanities nicht, ein Referenzsystem mit Konventionen zur Prüfung, Dokumentation, Verzeichnung und Aufbewahrung ihrer Daten in Institutionen sowie spezifischen Publikationsformen aufzubauen, ist die Vergleichbarkeit der Forschungsergebnisse nicht immer gegeben. Dabei bietet der Aufbau so genannter Forschungsdatenrepositorien erste positive Entwicklungen in Richtung dokumentierter Forschungsprozesse.<sup>11</sup>

Der Gegenstand der Digital Humanities ist nicht nur auf das digitale Objekt zu beschränken, sondern auch auf dessen Erstellung zu erweitern. Die Erstellung ist ein Prozess, der von den Digital Humanities zu dokumentieren ist, da nur so eine Referenzierbarkeit und Reproduzierbarkeit der Forschungsergebnisse möglich wird. Der kritische Umgang mit diesen Daten stellt eine Aufgabe der Wissenschaft dar.

Daher täten die Digital Humanities gut daran, in einen aktiven Dialog mit digitalisierenden Einrichtungen und Förderern einzutreten um gemeinsame Standards und Richtlinien zur Dokumentation zu etablieren, die transparent Auskunft über die Provenienz eines digitalen Objekts geben sowie klar über Möglichkeiten sowie Einschränkungen zu dessen Nachnutzbarkeit informieren. Entscheidungen, Kriterien, Daten und Ergebnisse sollten, im Unterschied zur bisherigen, in diesem Beitrag kritisierten Praxis, zukünftig möglichst transparent, operationalisierbar sowie validierbar digital zur Verfügung gestellt werden.

## Fußnoten

1. Vgl. dazu These 1.1: „Die Digital Humanities bereichern die traditionellen Geisteswissenschaften konzeptionell und methodisch - ihre Werkzeuge und Verfahren ergänzen das „Wie“ unserer Praxis um eine empirisch ausgerichtete Epistemologie.“ [Thesenpapier des Fachverbands „Digital Humanities im deutschsprachigen Raum“ 2014]
2. siehe [Schnell, Hill, Esser 2011: 4]
3. siehe Fußnote 2
4. OCR-D: Koordinierungsprojekt zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR), <http://www.ocr-d.de/>.
5. „Es werden weitere korpusbasierte Analysen angestrebt. Diese erfordern aber eine Voraussetzung: Im Zuge der ‚digitale[n] Wende‘ [Schöch 2014: 130] ist es weiterhin wünschenswert und erforderlich, dass immer mehr literarische Texte digital zur Verfügung stehen

oder diese durch leichte und praktikable Verfahren der Texterkennung (OCR, optical character recognition) digitalisierbar gemacht werden können.“ [Mihm 2016: 200]

6. „Dass die von uns einbezogenen Online-Repositorien hinsichtlich der editionsphilologischen Textqualität variieren ist ein hinzunehmendes Übel, dem wir zum einen pragmatisch (Wahl der bestmöglichen verfügbaren Ausgabe; Ziel, die Fehlermarge unter 2% zu halten), zum anderen unter Hinweis auf die flexible Struktur des Korpus (Austausch durch eine qualitativ hochwertigere Version ist möglich) begegnen. Durch die nahtlose Dokumentation des Korpus wird zudem die nötige Transparenz gewährleistet um auch Nachnutzern flexible Kontrolle der Daten zu ermöglichen.“ [Herrmann-Wolf, Lauer 2016: 159]

7. Page Analysis and Ground Truth Elements, siehe [http://www.primaresearch.org/publications/ICPR2010\\_Pletschacher\\_PAGE](http://www.primaresearch.org/publications/ICPR2010_Pletschacher_PAGE) und <https://github.com/PRImA-Research-Lab/PAGE-XML>.

8. Vgl. <https://www.digitisation.eu/tools-resources/image-and-ground-truth-resources/>.

9. ISO/IEC-Standard 19757-3:2006

10. Deutsches Textarchiv, DTA-Basisformat, <http://www.deustextarchiv.de/doku/basisformat/>.

11. Vgl. z.B. <https://rdmorganiser.github.io/>

Jahrestagung des Verbandes in Passau. <http://dig-hum.de/digital-humanities-2020>.

## Bibliographie

**Herrmann-Wolf, J. Berenike / Lauer, Gerhard** (2016): „Aufbau und Annotation des Kafka/Referenzkorpus“ in: Burr Elisabeth (ed): *DHd 2016 : Modellierung, Vernetzung, Visualisierung : die Digital Humanities als fächerübergreifendes Forschungsparadigma* : Konferenzabstracts : Universität Leipzig 7. bis 12. März 2016. [Duisburg]: nisaba 158-160.

**Mihm, Melanie** (2016): „Weibliches Erzählen im Expressionismus? Eine Stilometrie von Mela Hartwigs Prosa“ in: Burr Elisabeth (ed): *DHd 2016 : Modellierung, Vernetzung, Visualisierung : die Digital Humanities als fächerübergreifendes Forschungsparadigma* : Konferenzabstracts : Universität Leipzig 7. bis 12. März 2016. [Duisburg]: nisaba 198-200.

**Schöch, Christof** (2014): „Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik“, in: Schöch, Christof / Schneider, Lars (eds.): *Literaturwissenschaft im digitalen Medienwandel* (=Philologie im Netz Beiheft 7) 130-157 <http://web.fu-berlin.de/phn/beiheft7/b7t08.pdf> [letzter Zugriff 25.September 2017].

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

**Thesenpapier des Fachverbands „Digital Humanities im deutschsprachigen Raum“ (DHd)** (2014): „Digital Humanities 2020“, vorgestellt im März 2014 auf der ersten