Zusatzinformation

Skalierbare Verfahren der Text- und Strukturerkennung für die Volltextdigitalisierung historischer Drucke

Zusatzinformationen zur Ausschreibung „Skalierbare Verfahren der Text- und Strukturerkennung für die Volltextdigitalisierung historischer Drucke“ mit detaillierten Anforderungen an die einzelnen Module sowie Vorschlägen zu Lösungen

Der Ausschreibungstext ist unter folgender URL abrufbar:

http://www.dfg.de/download/pdf/foerderung/programme/lis/170306_ausschreibung_verfahren_volldigitalisierung.pdf

Deutsche Forschungsgemeinschaft
Kennedyallee 40 · 53175 Bonn · Postanschrift: 53170 Bonn
Telefon: + 49 228 885-1 · Telefax: + 49 228 885-2777 · postmaster@dfg.de · www.dfg.de

Vorbemerkung

Das vorliegende Dokument bezieht sich auf die DFG-Ausschreibung „Skalierbare Verfahren der Text- und Strukturerkennung für die Volltextdigitalisierung historischer Drucke“ und informiert über inhaltliche Aspekte der dort aufgeführten Module. Es ist als Ergänzung zum Ausschreibungstext zu verstehen und enthält detaillierte Anforderungen an die einzelnen Module. Die hier beschriebenen Aspekte sowie teils auch Lösungsvorschläge für Module sind beispielhaften Charakters und sollen Umsetzungsvorschläge der Modulprojekte keinesfalls einschränken. Am Ende des Dokumentes werden „Allgemeine inhaltliche und technische Rahmenbedingungen für die Modulprojekte“ formuliert, die bei der Antragstellung zu beachten sind.

Ausgewählte Literaturhinweise können dem Technology Watch der projekteigenen Zotero-Gruppe entnommen werden (https://www.zotero.org/groups/ocr-d).

1. Modul 1: Bildvorverarbeitung

1.1. Hintergrund

Am Beginn der automatischen Texterkennung mit *Optical Character Recognition* (OCR) stehen Verfahren, die die Digitalisate für die Layout- und Texterkennung charakterisieren und optimieren. Trotz der normierten Digitalisierung und der enormen technischen Entwicklung der Scanner ist aufgrund der speziellen Eigenschaften historischer Vorlagen eine große Heterogenität bzgl. Qualität und spezifischer Charakteristika der einzelnen Digitalisate zu verzeichnen. Durch eine Bildcharakterisierung nach optischen Kriterien unter Anwendung von Methoden der Bildähnlichkeitssuche können Metadaten erfasst werden, die den nachfolgenden Arbeitsschritten zu einem optimierten Einsatz verhelfen und damit qualitativ bessere Ergebnisse bei der nachfolgenden *Optical Layout Recognition* (OLR) und OCR erwarten lassen. So können beispielsweise Spaltendruck oder Marginalien frühzeitig im Prozess der Volltextdigitalisierung oberflächlich identifiziert werden und so die Layouterkennung bei deren Lokalisierung unterstützen. Werden Spalten oder Marginalien nicht als solche erkannt, erfolgt die Texterkennung über die Spalte hinaus oder in Marginalien hinein, entsteht fehlerhafter Text (Leseflussfehler), selbst wenn die Buchstaben- und Worterkennung korrekt ausgeführt worden sein sollte.

Die Bilddigitalisierung der fragilen und oftmals einzigartigen Materialien unterliegt außerdem Kompromissen zwischen technischen und konservatorischen Erfordernissen, die nicht immer zu optimalen Abbildungsergebnissen führen. Aufgrund dieser Realitäten ist der Einsatz adaptiver und flexibler Bildvorverarbeitungsverfahren oftmals von entscheidender Bedeutung für die Sicherstellung eines verwertbaren Erfassungsergebnisses. Diese Verfahren umfassen u. a. den Zuschnitt des Digitalisats (*cropping*), dessen Bereinigung von Artefakten (*despeckling*), eine Korrektur eventuell vorhandener Schrägstellungen oder Versatze auf Bild- (*deskewing*) bzw. Zeilenebene (*dewarping*) sowie, in Abhängigkeit vom gewählten Texterfassungsverfahren, eine Binarisierung (*binarization*).

Die Vorverarbeitung der Digitalisate hat erheblichen Einfluss auf die Genauigkeit des Gesamtergebnisses im Texterfassungsprozess sowohl auf struktureller als auch textueller Ebene. Selbst einfache prozedurale Anpassungen wie etwa die Auswahl des Binarisierungsalgorithmus oder die konkrete Parameterbelegung beim Bildzuschnitt können die Text- bzw. Strukturerkennungsgenauigkeit positiv beeinflussen. Das Vorhandensein und der korrekte Einsatz entsprechender hochqualitativer, modularer Verfahren stellt somit eine der wichtigsten Voraussetzungen für eine optimale Texterfassung mit Methoden der OCR dar.

1.2. Ziele

Innerhalb dieses Moduls sollen Lösungen erarbeitet werden, die in der Lage sind, den gesamten Bestand der Digitalisate (VD-Drucke des 16., 17. und 18. Jh. sowie Werke des 19. Jh.) optimal für die anschließende Volltextdigitalisierung mit OCR-Verfahren vorzuverarbeiten.

Bei der **Bildcharakterisierung** sind alle vorhandenen und zu ermittelnden Informationen gezielt auszuwerten. Unter anderem können an Hand der intrinsischen und extrinsischen Problemmatrix (s.u.) entsprechende Seiten identifiziert, Metadaten der Objekte gezielt interpretiert sowie automatische Verfahren wie Ähnlichkeits- und Vergleichsanalysen genutzt werden.

Bei der **Bildoptimierung** sind sowohl vollautomatische als auch individuelle Verfahren in die Überlegungen einzuschließen. Um die verfahrenstechnischen Veränderungen am zu bearbeitenden Digitalisat nachzuvollziehen, ist deren Dokumentation in Form von standardisierten Metadaten unumgänglich.

Da die Bilddigitalisate aufgrund der stark heterogenen Vorlagen trotz der standardisierten Digitalisierung teilweise sehr unterschiedliche Charakteristiken aufweisen, wird empfohlen, im Sinne einer optimalen Adaptivität, für jede Teilaufgabe mehrere Algorithmen bzw. Parametrisierungsmöglichkeiten zu implementieren. Das gilt insbesondere für die Teilaufgabe Binarization. Die Anwendung der einzelnen Bildvorverarbeitungsschritte soll auf Seitenebene aber auch auf kleineren Einheiten wie Absatz oder Zeile möglich sein. Folgende Teilaufgaben der Bildvorverarbeitung sind dabei mindestens zu bearbeiten:

1.3 Teilaufgabe 1.A: Bildcharakterisierung

Das zu erkennende Material zeigt ein stark heterogenes Erscheinungsbild. Dabei kann zwischen intrinsischen und extrinsischen Problemen unterschieden werden. Bei der Anwendung von Algorithmen geleiteten Verfahren zur Ermittlung von Seiten- (z. B. Grauwert der Seite) bzw. Segmenteigenschaften (z. B. Verhältnis Abbildung – Text) können spezifische Charakteristika ermittelt werden. In der Interpretation, die auf Ähnlichkeitsverfahren beruhen könnte, werden diese Eigenschaften den intrinsischen und extrinsischen Charakteristika zugeordnet. Im Folgenden werden in einem übersichtlichen Raster beispielhaft typische Probleme aufgelistet. Die bei der Bildcharakterisierung gewonnenen Informationen sind als Metadaten zum Objekt zu speichern.

Intrinsisch:
- Typographie/Layout
  - Fontmix (Absatzebene)
  - Fontmix (Wortebene)
  - Fontmix (Zeichenebene)
  - Schriftgrößenmix
    - vor allem zu finden auf Titelblättern
  - Sprachmix
  - Langes s, rundes r
  - komplexes Layout
    - Drucke die Spaltendruck, Tabellen, geschachtelte Listen, verschiedene eingerückte Absätze (z.B. Funeraldrucke) aufweisen
  - Abbildung
    - alle graphischen Elemente, ausgenommen sind Buchschmuck
  - Fußnoten/Marginalien
  - Sonderzeichen
    - Zeichen die nicht einem Alphabet zuzuordnen sind z.B. Sternkreiszeichen
  - Ligaturen
  - Abbreviaturen
  - Zierbuchstaben/Versalia
  - Fettschrift/Normschrift-Wechsel
  - Zahlen
    - größere Zahlenwerke (z. B. astronomische Tafeln)

Extrinsisch:
- Produktionsfehler
  - Tintendurchdruck
- Benutzung
  - Flecken/Annotationen und Papierqualität
  - Knicke/Falten
- Digitalisierung
  - Verzerrung
  - Farbsprung
  - Variierende Kontrastverhältnisse
  - Gespiegelte Buchseite

Abb. 1: Beispielmatrix zur Bildcharakterisierung

1.4 Teilaufgabe 1.B: Bildoptimierung

Im Folgenden werden die wichtigsten fünf Bildvorverarbeitungsstufen (*Cropping, Deskewing, Binarization, Despeckling, Dewarping*) anhand von Beispielen illustriert. Nicht immer ist die Anwendung aller Vorverarbeitungsschritte notwendig bzw. zielführend. Im Einzelfall hängt dies vom spezifischen Layout- bzw. Texterkennungsalgorithmus ab.

Abb. 2: **Dewarping** – Begradigung von Wellen auf Zeilenebene
Der Seelen Wurzgarten, Augsburg, 1504 [VD16 S 5276]
Permalink: http://www.mdz-nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:bvb:12-bsb10943350-8

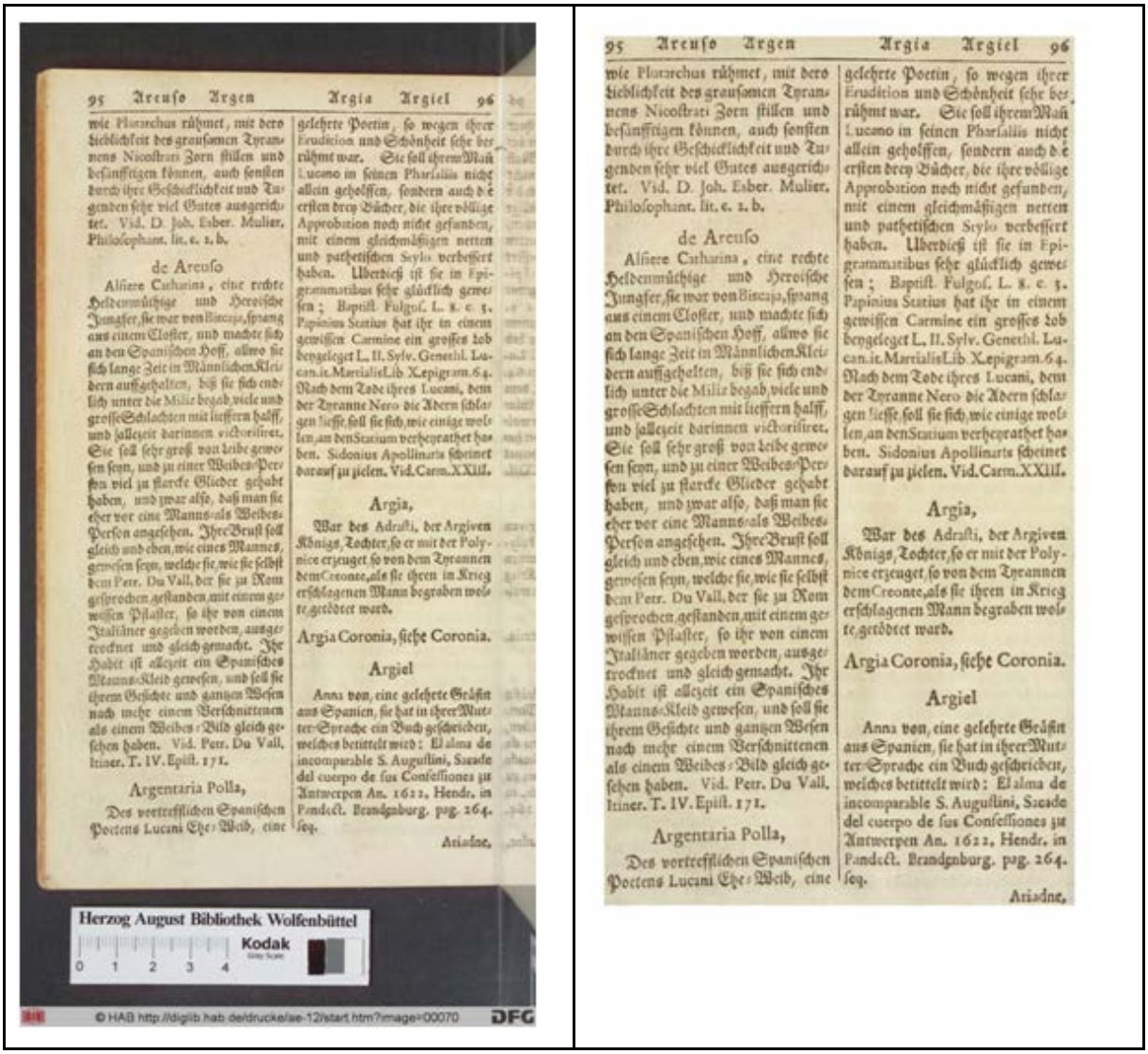


Abb. 3: **Cropping** – Beschneiden des Digitalisats auf den Druckbereich
Gottlieb Siegmund Corvinus: Nutzbares, galantes und curiöses Frauenzimmer-Lexicon : Worinnen nicht nur Der Frauenzimmer geistlich- und weltliche Orden, Aemter, Würden, Ehren-Stellen, Professionen und Gewerbe, ... Nahmen und Thaten der Göttinnen, ... gelehrter Weibes-Bilder …,.../. Leipzig, 1715. Image 70.
Permalink: http://diglib.hab.de/drucke/ae-12/start.htm?image=00070

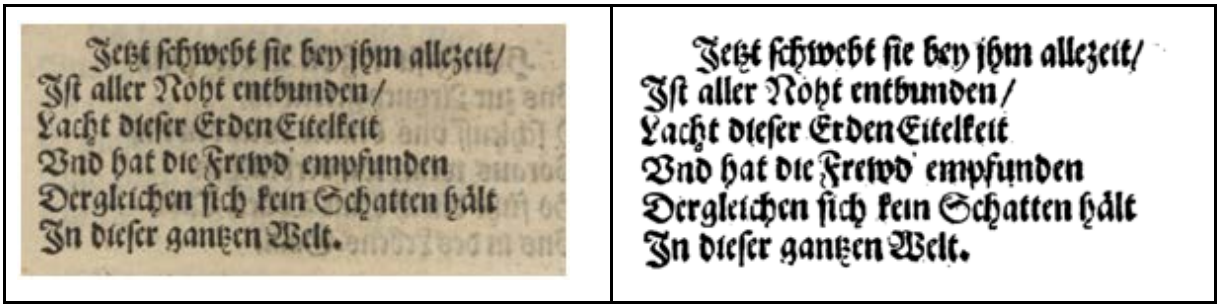


Abb. 4: **Binarization** – Binäre Kodierung der Pixel
(bedruckte Bereiche schwarz, nicht-bedruckte Bereiche weiß)
Simon Dach: Einfältige Leich-Reime Der ... Catharinen Pohlinn/ Seel. Herrn Cornelius Cronen etc. Hinterlassenen Witwen. S. 12, Königsberg, 1653. Permalink: http://resolver.staatsbibliothek-berlin.de/SBB000032AA00000012

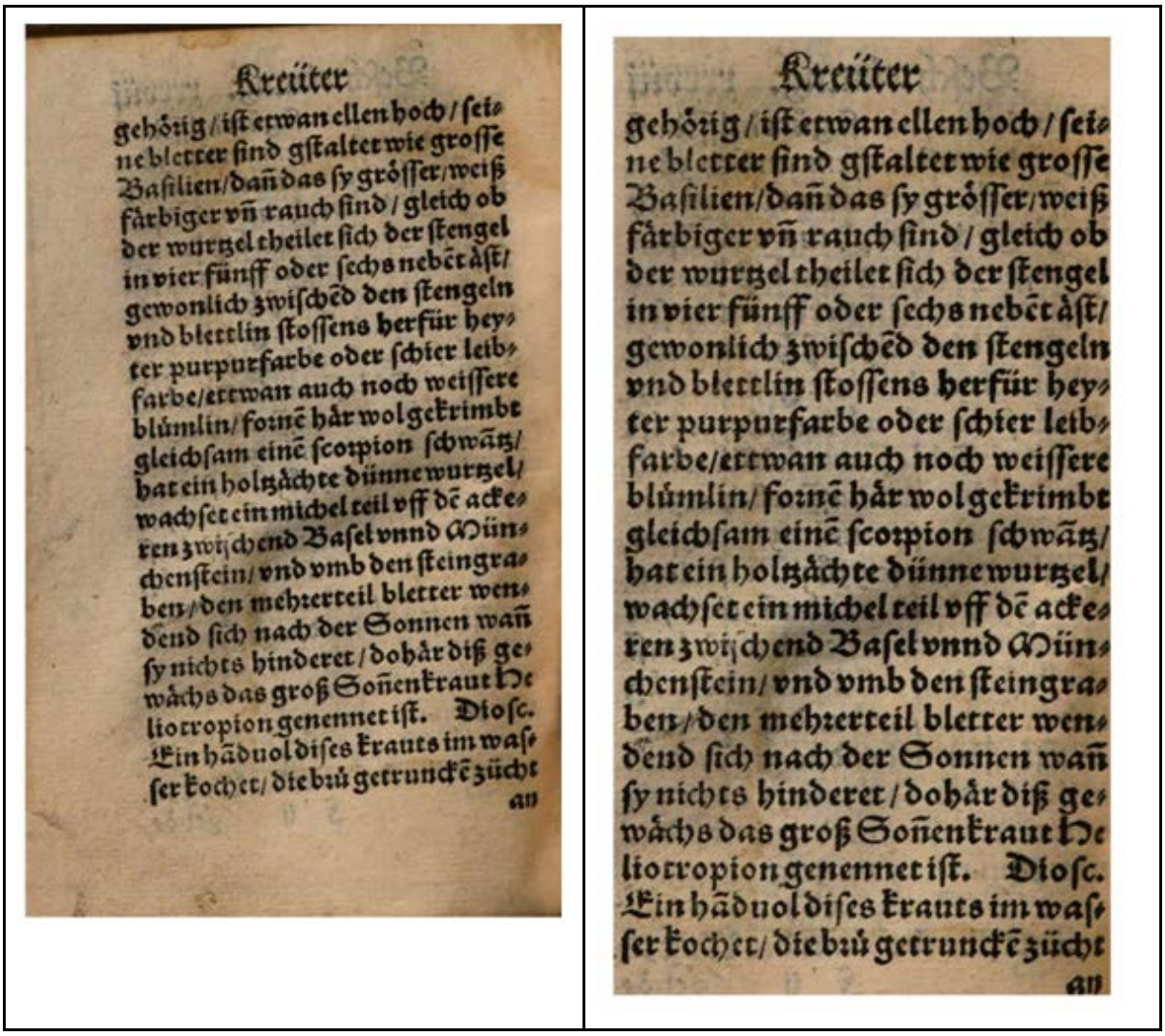


Abb. 5: **Deskewing** – Rotation des Digitalisats zur Begradigung von Schrägstellungen
Adam von Bodenstein: Wie sich meniglich vor dem Cyperlin, Podagra genennet, waffnen solle unnd Bericht diser Kreüter, so den himmelischen Zeichen Zodiaci zugeachtet. S. XXXVIII, Basel, 1557.
Permalink: http://www.mdz-nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:bvb:12-bsb11106588-9

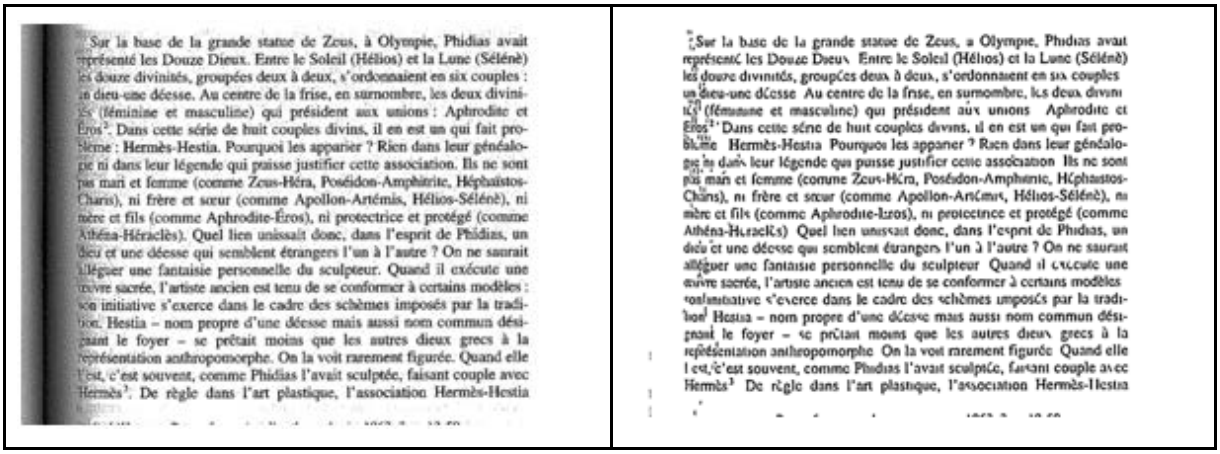


Abb. 6: **Despeckling** - Entfernung von Bildartefakten (Verschmutzungen, etc.)
Jean-Pierre Vernant: Hestia-Hermès: Sur l'expression religieuse de l'espace et du mouvement chez les Grecs.
Link: http://imgur.com/lm9luhE (Stand 15.04.2016)

2. Modul 2: Layouterkennung

2.1. Hintergrund

Die Erfassung des korrekten, d.h. die Wiedergabe des logisch strukturell auf sich beziehenden Textflusses ist neben einer möglichst hohen Textgenauigkeit als wichtigstes Ziel der Volltextdigitalisierung zu betrachten. Im Allgemeinen reiht sich in einem Text Wort an Wort und Absatz an Absatz. Jedoch wird der Textfluss durch metatextliche Informationen wie z.B. Seitenzahlen, Kustoden, Marginalien, Kolumnentitel, Fußnoten, Spalten, Tabellen, Abbildungen unterbrochen.

Menschliche Leser sind in der Lage, den logischen Textfluss trotz derartiger Unterbrechungen auf einen Blick zu erkennen. Im Falle einer maschinellen Auswertung müssen Informationen vorliegen, die z.B. Seitenzahlen als nicht Bestandteil des Haupttextes ausweisen. So sollte beim maschinellen Verarbeiten des gewonnenen OCR-Volltextes die Logik des Textes weiterhin erfassbar sein. Dies ist eine Voraussetzung für die wissenschaftliche Recherche im Volltext, deren Grundlage u. a. die korrekte Indizierung der textuellen Segmente voraussetzt, sowie für die Verarbeitung des Textes in digitalen Forschungsumgebungen.

Software-Lösungen im Bereich der OCR enthalten Verarbeitungsschritte, die dazu führen, dass abgrenzbare Blöcke im Quelldigitalisat lokalisiert werden. Dabei werden vor allem drei Typen von Blöcken unterschieden: (reiner) Text, Abbildungen und Tabellen. Dies ist jedoch für eine tiefergehende Erschließung der Texte nicht ausreichend. Hier setzt die automatische Layouterkennung (*Optical Layout Recognition*, OLR) an. Auf Basis der Seitensegmentierung werden explizite Metainformationen über die layout-strukturelle Funktion des erkannten Zeichens oder der Zeichenkette erzeugt und können in datentechnisch höheren Ausgabeformaten (OCR-Ausgabeformate und Meta[daten]formate) gespeichert werden.

2.2. Ziele

Ziel der Lösungen der Teilaufgaben in diesem Modul ist, dass alle auf einer Seite erkennbaren Regionen lokalisiert und entsprechend den Anforderungen klassifiziert werden. In die Überlegungen ist mit einzubeziehen, dass Regionen nicht immer in Form von Rechtecken umrissen werden können sondern oftmals Polygone höherer Ordnung dafür nötig sind. Die Lösungen sollen dazu beitragen, dass die Volltextdigitalisierung der Drucke als Massenanwendung realisierbar wird.

Die folgenden Vorgaben an die Strukturierung der Volltexte schließen an die Strukturerfassung (vgl. z.B. http://dfg-viewer.de/strukturdatenset/), wie sie in den DFG-Praxisregeln „Digitalisierung" (siehe http://www.dfg.de/formulare/12_151/12_151_de.pdf) formuliert wurden, an. Somit sollen nicht nur die Dokumente durch einen hochstrukturierten Metadatensatz repräsentiert werden, sondern die logische Textstruktur des Dokumentes soll im Dokument selbst ablesbar werden. Die OLR ist als ein Bearbeitungsschritt im Prozess der OCR zu verstehen. Metadaten die in vorangegangenen Bearbeitungs- und Erfassungsschritten automatisch oder manuell erfasst wurden sind in die OLR mit einzubeziehen. Metadaten, die das Textergebnis beschreiben, sind gegebenenfalls bei einer wiederholten Anwendung der OLR auszuwerten.

2.3. Teilaufgabe 2.A: Seitensegmentierung

Alle Regionen einer Seite sind automatisch zu erkennen und entsprechend der untenstehenden beispielhaft aufgeführten Typologie zu kennzeichnen. Es reicht nicht aus, den kompletten Satzspiegel als einen Textblock zu identifizieren sondern gefordert ist die Lokalisierung einzeln abgrenzbarer Blöcke. Übergeordnetes Ziel ist die Trennung von textuellen und nicht-textuellen Segmenten.

Die minimalen Erfassungsergebnisse beinhalten:

- (reiner) Text
- Abbildung
- Tabelle
- Separator
- (Sonstiges)

Ein maximales Ergebnis der entwickelten Lösung liegt vor, wenn die folgenden Regionen wie folgt feintypisiert werden:

- Text
- Abbildung (Zeichnung, Grafik)
- Vignette (Buchschmuck)
- Tabelle
- Diagramm
- Separator
- Mathematische Formel
- Chemische Formel
- Noten
- Werbung
- Rauschen (u.a. Verschmutzungen, Stempel)
- Handschriftliche Anmerkungen
- Sonstiges

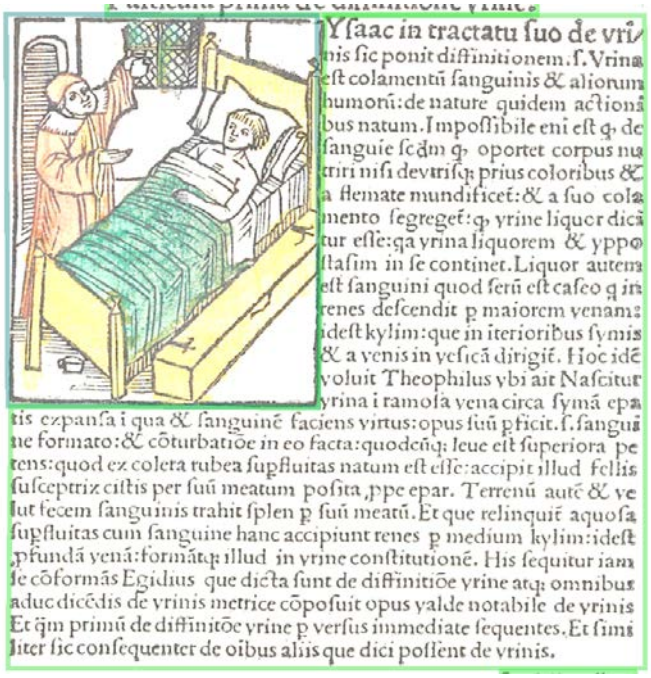


Abb. 7: Polygone Segmente
Pinder, Ulrich: Epyphani medicorum, [Nürnberg], 1506.
Permalink: http://resolver.staatsbibliothek-berlin.de/SBB0001C4CF00000009

Abb. 8: Polygone Segmente

Berliner Lokal Anzeiger, 14 Mai 1910.

Link: http://zefys.staatsbibliothek-berlin.de/dfg-viewer/?no_cache=1&set[image]=8&set[zoom]=default&set[debug]=0&set[double]=0&set[mets]=http%3A%2F%2Fzefys.staatsbibliothek-berlin.de%2Foai%2F%3Ftx_zefysoai_pi1%255Bidentifier%255D%3D2e9ef054-0947-4b14-b45e-1fb41a76a170

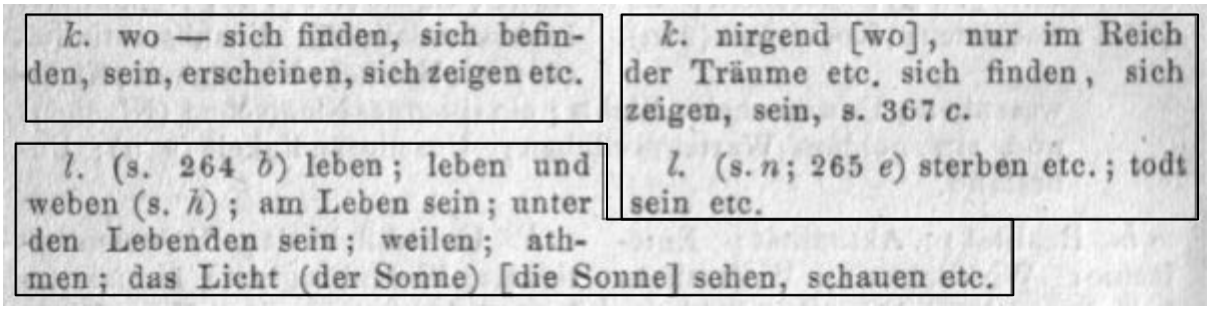


Abb. 9: Polygone Segmente

Deutscher Sprachschatz geordnet nach Begriffen zur leichten Auffindung und Auswahl des passenden Ausdrucks. Ein stilistisches Hülfsbuch für jeden Deutsch Schreibenden. Bd. 1. Hamburg, 1873.

Link: https://archive.org/details/deutschersprachs01sanduoft, Seite 2

2.3. Teilaufgabe 2.B: Textzeilenerkennung

Innerhalb der zuvor lokalisierten Textblöcke sind die einzelnen Zeilen zu verorten und anhand ihrer Koordinaten zu repräsentieren. Dieser Aufgabe kommt innerhalb des skizzierten Funktionsmodells eine besondere Bedeutung zu, da die Zeile im Rahmen der Projektarbeit von

OCR-D als atomare Einheit für die nachfolgende Texterkennung betrachtet wird (vgl. Modul 4). Fehler innerhalb der Textzeilenerkennung wirken sich somit direkt negativ auf das Textergebnis aus. Eine besondere Herausforderung stellen dabei die hier adressierten historischen Vorlagen dar, die häufig Digitalisierungsartefakte wie schräggestellte oder gewellte Zeilen enthalten. Hier ist eine enge Abstimmung mit Modul 1 sinnvoll. Die zu entwickelnden Lösungen sollen sich außerdem möglichst robust gegenüber Fehlern in der Seitensegmentierung verhalten bzw. auch auf Seitenebene anwendbar sein.

2.3. Teilaufgabe 2.C: Segmentklassifizierung

Auf Grundlage der identifizierten Blöcke der Seitensegmentierung im Zusammenspiel mit deren Koordinaten kann eine semantische Zuordnung von bestimmten Regionen auf dem Digitalisat erfolgen. Möglichkeiten der automatischen Segmentklassifikation u. a. auf Grundlage von Ground-Truth-Daten sowie semiautomatische Methoden auf Basis von nutzerorientierten Konfigurationen sind in die angestrebten Lösungen mit einzubeziehen. Die Segmentklassifizierung bildet zum einen die Grundlage für die nachfolgende Extraktion der logischen Dokumentstruktur und zum anderen beeinflusst sie die Rekonstruktion des Textflusses (s.u.). Folgende Regionen sind spezifisch zu bezeichnen.

Die minimalen Anforderungen umfassen:

- Absatz
- Überschrift der Ordnung n
- Beschriftung
- Kolumnentitel
- Fußzeile
- Seitenzahl
- (Schmuck-)Initiale
- schwimmende Elemente im Satzspiegel
- Bogensignatur
- Kustode
- Marginalie
- Fußnote
- Anderes

Ein maximales Ergebnis der entwickelten Lösung liegt vor, wenn zu den minimalen Anforderungen die folgenden Typisierungen hinzukommen:

- fortgesetzter Absatz (eingeschlossen sind alle sich fortsetzende Regionen)
- fortgesetzte Fußnote und Endnote
- Endnote
- Inhaltsverzeichniseintrag

Nachfolgend findet sich eine Übersicht der häufigsten Elemente, die den linearen Textfluss unterbrechen können.

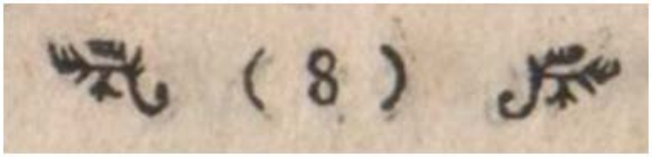
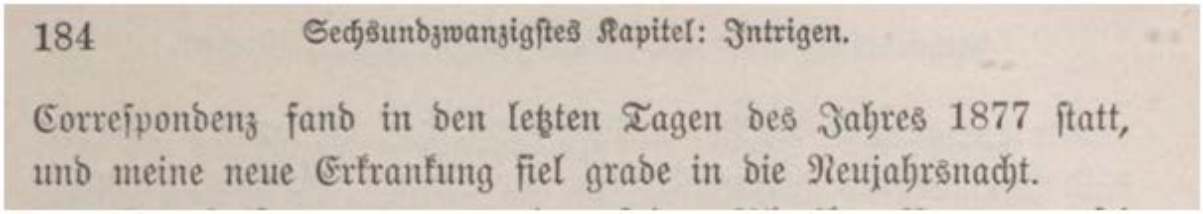


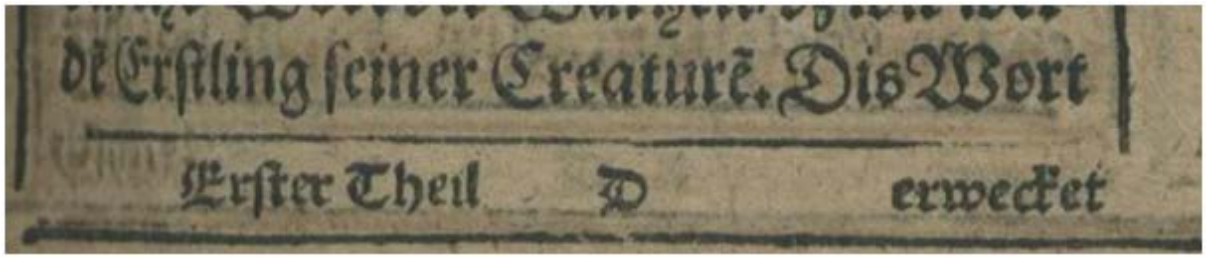
Abb. 10: **Seitenzahl** – Angabe der aktuellen Seite unter- oder oberhalb des Satzspiegels
Campe, Joachim Heinrich: Robinson der Jüngere. Bd. 2. Hamburg, 1780, [Faksimile 14].
Permalink: http://www.deutschestextarchiv.de/book/view/campe_robinson02_1780?p=14
URN: urn:nbn:de:kobv:b4-20090519755


184 Sechsundzwanzigstes Kapitel: Intrigen.

Correspondenz fand in den letzten Tagen des Jahres 1877 statt, und meine neue Erkrankung fiel grade in die Neujahrsnacht.

Abb. 11: **Lebende Kolumnentitel** – befinden sich in den Vorlagen jeweils zu Beginn einer Seite in der Kopfzeile

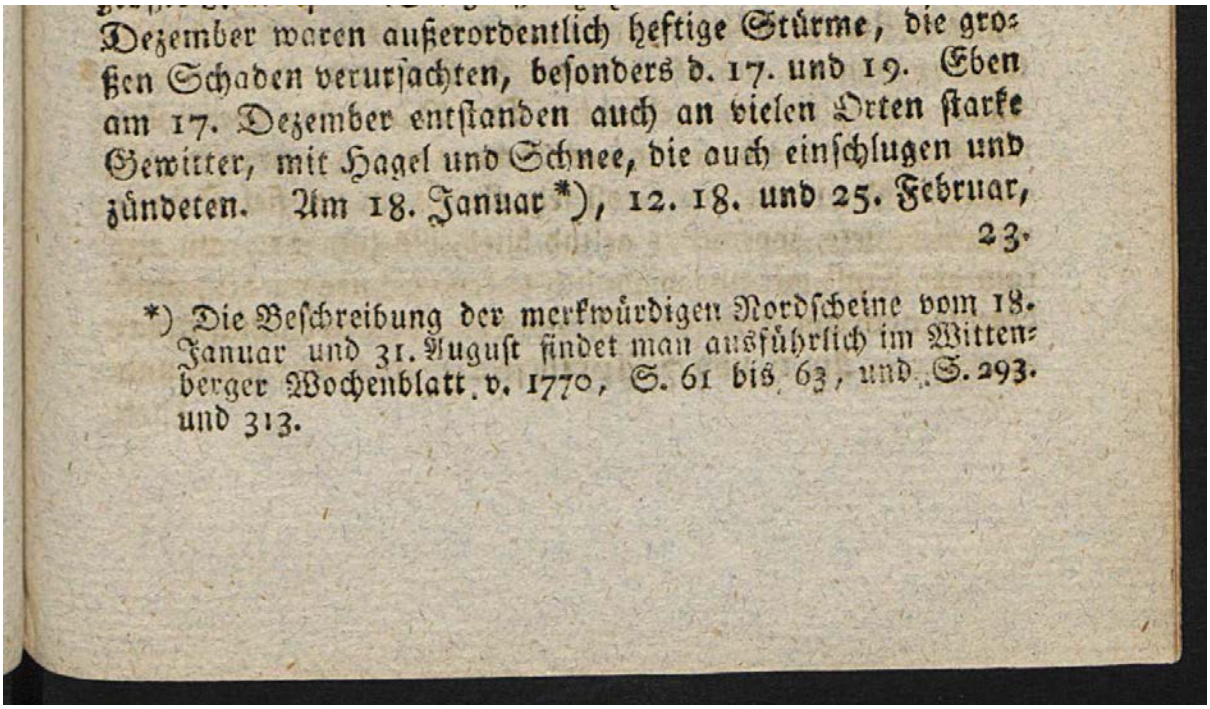
Bismarck, Otto von: Gedanken und Erinnerungen. Bd. 2. Stuttgart, 1898. [Faksimile 208].
Permalink: http://www.deutschestextarchiv.de/book/view/bismarck_erinnerungen02_1898?p=208
URN: urn:nbn:de:kobv:b4-20090519401


dt Erſtling ſeiner Creatur. Dis Wort

Erſter Theil D erwecket

Abb. 12: **Bogensignatur** – Rohbogenbezeichnung in sehr kleinem Schriftsatz, die sich unterhalb oder seltener oberhalb des Satzspiegels befindet; **Kustode** - In der rechten unteren Ecke der Seite angebrachte Angabe der Anfangssilbe oder des ersten Worts der Folgeseite

Arndt, Johann: Von wahrem Christenthumb. Bd. 1. Magdeburg, 1610. [Faksimile 55].
Permalink: http://www.deutschestextarchiv.de/book/view/arndt_christentum01_1610?p=55
URN: urn:nbn:de:kobv:b4-200905199938


Dezember waren außerordentlich heftige Stürme, die groſſen Schaden verurſachten, beſonders d. 17. und 19. Eben am 17. Dezember entſtanden auch an vielen Orten ſtarke Gewitter, mit Hagel und Schnee, die auch einſchlugen und zündeten. Am 18. Januar *), 12. 18. und 25. Februar, 23.

*) Die Beſchreibung der merkwürdigen Nordſcheine vom 18. Januar und 31. Auguſt findet man ausführlich im Wittenberger Wochenblatt, v. 1770, S. 61 bis 63, und S. 293. und 313.

Abb. 13: **Fußnote** (unten) – In kleine(re)m Schriftsatz gedruckte Zusatzinformation am unteren Ende einer Seite.

Gronau, Karl Ludwig: Versuch einiger Beobachtungen über die Witterung der Mark Brandenburg, besonders in der Gegend um Berlin. Berlin, 1794, S. 219.
Permalink: http://resolver.staatsbibliothek-berlin.de/SBB000064A700010227

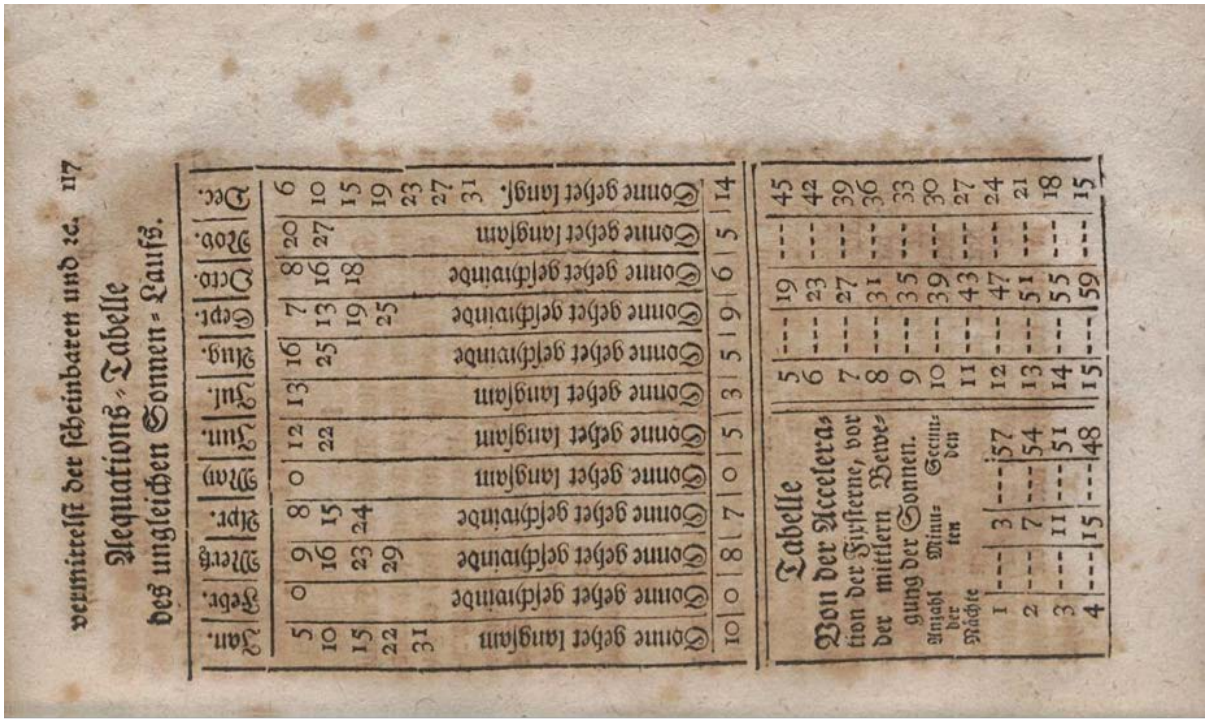


Abb. 14: **Gleitobjekt – Abbildungen und Tabellen** die aus ökonomischen Gründen nicht an der Stelle gesetzt sind, an der sie referenziert werden
Heinrich Sully Uhrmachers des Herzogs von Orleans, und Mitgliedes der Societät der Künste zu Paris Kunstmässiger Unterricht Von der Eintheilung der Zeit, Worinnen gehandelt wird von der natürlichen und künstlichen Eintheilungen der Zeit; von den verschiedenen Einrichtungen grosser und kleiner Uhren [...]. Lemgo 1746.
Permalink: http://resolver.staatsbibliothek-berlin.de/SBB0001B2DD00000145

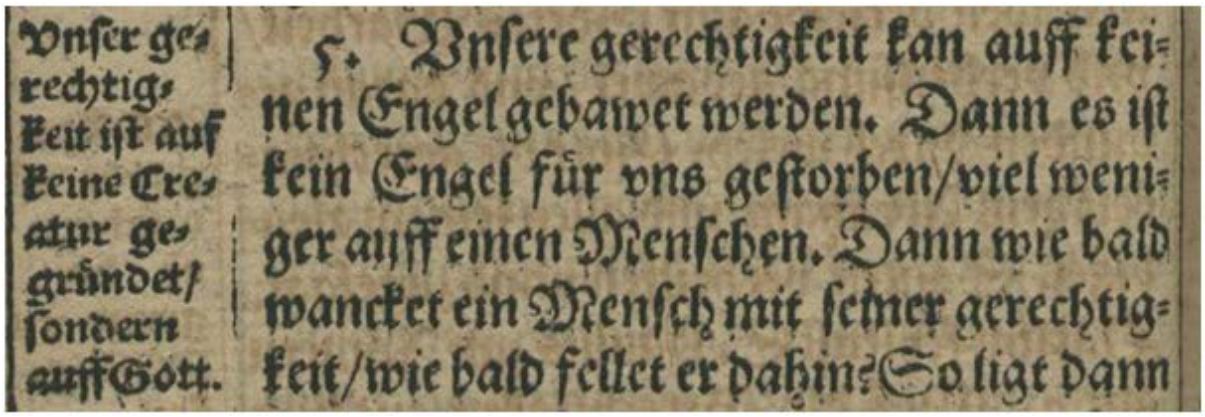


Abb. 15: **Marginalie** – Auf dem Rand einer Seite platzierte Bemerkung, die einen Kommentar zu einer Stelle des Textes bietet
Arndt, Johann: Vom wahren Christenthumb. Bd. 2. Magdeburg, 1610. [Faksimile 64].
Permalink: http://www.deutschestextarchiv.de/book/view/arndt_christentum02_1610?p=64
URN: urn:nbn:de:kobv:b4-200905199920

Weiterhin gibt es in Abhängigkeit der Textsorte viele weitere weniger verbreitete Phänomene wie etwa Bühnen- oder Regieanweisungen in Dramen oder Libretti, Gleittext in Zeitungen, Textboxen in Lehrbüchern oder Schmucktext, die hier nicht vollständig aufgeführt werden können, bei der Segmentklassifizierung aber behandelt werden sollten.

2.3. Teilaufgabe 2.D: Dokumentanalyse

Da OCR-Programme zumeist auf Seitenebene agieren, findet eine Ermittlung der logischen Textstruktur des gesamten Dokumentes in diesem Zusammenhang nicht statt. Die Herausforderung bei der automatischen Dokumentanalyse besteht in der Aggregation der isolierten, seitenspezifischen Informationen als Ergebnis der Segmentklassifizierung zu einer den Textfluss abbildenden, in den Kontext des Dokuments eingebetteten Struktur.

Eine Seite kann dabei über mehrere Ebenen von Text verfügen. Neben dem Haupttext auf der Seite werden spezifische Informationen z.B. in Marginalien, Seitenzahlen, sowie Fußnoten festgehalten. Diese spezifischen Informationen bilden eigene Textebenen über Seitengrenzen hinweg und werden im Vergleich zum Haupttext auf der Seite in diesem Zusammenhang als nachgeordnete Strukturen gesehen. Aus pragmatischen Gründen ist die Ermittlung des Haupttextes mit einem korrekten Textfluss, der bspw. die Satzrekonstruktion ermöglicht, zu priorisieren. Ein maximales Ergebnis beinhaltet die Ermittlung und Verknüpfung aller Textebenen im gesamten Dokument.

Entsprechende Untersuchungen bzw. Wettbewerbe (vgl. ICDAR Competition on Book Structure Extraction, http://pageperso.univ-lr.fr/antoine.doucet/StructureExtraction/2013/) in diesem Feld zeigen, dass eine verlässliche Segmentklassifizierung (insbesondere die korrekte Identifikation der Überschriften) sowohl die Voraussetzung als auch die Herausforderung für eine hochqualitative Aggregation auf Dokumentebene darstellen. Die Erfassung der Dokumentenstruktur stellt eine zeit- und kostenintensive Aufgabe in der Bibliothek und den Archiven dar. Mit Hilfe von Methoden aus der computergestützen Dokumentanalyse auf Basis der vorklassifizierten layoutsemantischen Einheiten soll dieser Prozess automatisiert werden. Die Speicherung der Dokumentstruktur setzt voraus, dass die Dokumentenanalyse in den dafür vorhandenen Formaten (z. B. METS-Metadaten, vgl. z.B. http://dfg-viewer.de/strukturdatenset/) erfolgt.

3. Modul 3: Textoptimierung

3.1. Hintergrund

Aktuelle Untersuchungen zum Vergleich von OCR-Verfahren machen deutlich, dass verschiedene Ansätze unterschiedliche Stärken und Schwächen haben und somit unterschiedliche Volltextausgaben für das gleiche Digitalisat produzieren. Dabei macht ein Vergleich der einzelnen Textversionen deutlich, dass es unter geeigneter Parameterwahl für die OCR bzw. die Bildvorverarbeitung möglich ist, praktisch jedes Zeichen korrekt zu erkennen. Die Auswahl der für ein Digitalisat bzw. für einen Teil eines Digitalisats optimalen Erkennungsroutine sowie die perfekte Parameterbelegung sind jedoch bisher nicht prozedural zu treffen.

Dies legt ein Vorgehen nahe, das anstatt sich auf die Optimierung eines einzelnen OCR-Verfahrens zu konzentrieren, den Einsatz verschiedener OCR-Verfahren optimiert. Eine weitere Option besteht in der Kombination mehrerer OCR-Engines bzw. Textergebnisse[1]. Pilothafte Studien in diesem Bereich zeigen das Potential, aber auch den Entwicklungsbedarf in dieser Frage.

Tesseract	OCRopus	Vereinigter Volltext
Es koſtet jhm kein zeitlich Gut	Es koſlet jhm kein zeitlich Gut	Es koſtet jhm kein zeitlich Gut
Vns wieder zu erwerben/	Vns wieder zu terwerben/	Vns wieder zu erwerben/
Es ihaietz?i1ichi der Opffer Blui/	Es that es nicht der Opffer Blut/	Es that es nicht der Opffer Blut/
Er muſte ſelber ſterben	Er muſte ſelber ſterben	Er muſte ſelber ſterben
Vnd einenTod zwar / welcher gar	Vnd einen Tod zwar/ welcher gar	Vnd einen Tod zwar/ welcher gar
EinFliich'vud Grewel war.	Ein Fluch vnd Grewel war.	Ein Fluch vnd Grewel war.

Abb. 16: Strophe aus Simon Dach „Einfältige Leichreime“ (1653) samt OCR-Ergebnis bei Erfassung mit Tesseract und OCRopus (jeweils mit angepassten OCR-Modellen) unter Hervorhebung der Erkennungsvorteile bei Tesseract und einem ideal vereinigten Volltext
Dach, Simon: Einfältige Leich-Reime Der ... Catharinen Pohlinn/ Seel. Herrn ..., 1653.
Permalink: http://resolver.staatsbibliothek-berlin.de/SBB000032AA00000000, Image 10

Um historische Drucke uneingeschränkt für eine Volltextrecherche sowie die sofortige Verwendung für Forschung und Wissenschaft zur Verfügung stellen zu können, muss trotz der Nützlichkeit auch weniger genauer Texte mittel- bis langfristig eine hohe Textgenauigkeit erreicht werden. Auch die derzeit besten Verfahren für historische OCR erreichen die notwendige Qualität ohne zusätzliche Bearbeitung oftmals nicht. Eine Nachkorrektur ist daher unumgänglich. Es existiert bereits eine Reihe kommerzieller und Open-Source-Lösungen für die Nachkorrektur von OCR-generierten Texten. Oftmals handelt es sich hierbei jedoch um interaktive oder rein manuelle Verfahren. Vollautomatische Verfahren, (auch *unsupervised post-correction* genannt), korrigieren den Text ohne menschliches Einwirken, während halbautomatische Verfahren die menschliche Arbeit, z.B. durch die Anzeige möglicher Fehler, unterstützen.

3.2. Teilaufgabe 3.A: Optimierter Einsatz von OCR-Verfahren

Es sind Methoden zu entwickeln, die die automatische Texterkennung auf den hier adressierten Materialien verbessern. Dabei sollen aktuelle Forschungsergebnisse geprüft und gegebenenfalls in den Zustand der Praxisreife gebracht werden. Dabei sind vornehmlich Verfahren zu betrachten, die direkt die Erkennung verbessern (vgl. Teilaufgabe Nachkorrektur für anschließende Verbesserungen des Textergebnisses), etwa durch die Optimierung und Ergänzung vorhandener Ansätze oder deren Kombination (zum Beispiel durch Votingverfahren). Gemäß der Projektprämissen ist ein spezielles Augenmerk auf Verfahren auf Basis neuronaler Netze zu legen. Dies kann zum Beispiel durch die Weiterentwicklung und Adaption von Open-Source-OCR-Engines wie OCRopus und Tesseract, die beide in ihren aktuellen Versionen sog. rekurrente neuronale Netze einsetzen, geschehen. Auch die bedeutenden Entwicklungen im Bereich der Handschriftenerkennung (*Handwritten Text Recognition*, HTR) der letzten Jahre nutzen vergleichbare Verfahren, deren Potentiale für die Erkennung gedruckter, historischer Vorlagen es ebenfalls systematisch auszuloten gilt.

Betrachtet seien in der Bearbeitung der Aufgabe auch Teilprobleme wie die Auswahl von OCR-Methoden, deren Parametrisierung sowie der Einsatz eines geeigneten OCR-Modells im konkreten Anwendungsfall, da diese Entscheidungen einen maßgeblichen Einfluss auf das Textergebnis haben, meist aber nicht einfach zu treffen sind. Insbesondere vor dem Hintergrund der Massenvolltextdigitalisierung sind dabei automatische Prozeduren bzw. Empfehlungen notwendig. Den speziellen Charakteristika der historischen Vorlagen – Sprach-, Schriftart- und Alphabetwechsel oder eng bedruckte Seiten und geringe Spationierung – sind bei der OCR-Optimierung besondere Aufmerksamkeit zu schenken.

3.3. Teilaufgabe 3.B: Nachkorrektur

Ziel dieser Teilaufgabe ist die Entwicklung einer Software, die die Möglichkeit bietet, OCR-Texte nachzukorrigieren. Aufgrund der zu bearbeitenden Textmenge ist eine Option zur vollautomatischen Nachkorrektur das primäre Ziel. Semiautomatische Lösungen (*interactive post-correction*) sowie Umgebungen zur manuellen Nachkorrektur sind jedoch ebenfalls in die Überlegungen einzubeziehen.

Ein Hauptaugenmerk soll dabei der Lokalisierung möglicher OCR-Fehler gelten. Bei der Entwicklung entsprechender Verfahren sollen dabei eine Vielzahl möglicher Metadaten auf ihre Eignung als Fehler- oder Korrekturhinweis überprüft werden.

Die zu entwickelnden Lösungen können dabei unter anderem auf fehlerfreien Volltext, historische Lexika und andere linguistische Ressourcen zurückgreifen. Hierbei haben sich Verfahren, die morphologische oder lexikalische Ressourcen einbeziehen sowie historische Sprachmodelle (*language models*), als sehr erfolgreich erwiesen. Darüber hinaus stehen Metadaten aus den vorangegangen Arbeitsschritten, wie etwa die Eigeneinschätzung der OCR-Engine, zur Verfügung. Im Falle von OCR-Verfahren auf Basis neuronaler Netze kann außerdem auf einfache Art und Weise an jeder Position des Textes die Wahrscheinlichkeitsverteilung über das gesamte Alphabet konsultiert werden. Zusätzlich sind für einen Teil des bislang im Koordinierungsprojekt zusammengestellten Ground-Truth OCR-Ausgangsdaten vorhanden, die z.B. der Entwicklung von Fehlermodellen dienen können.

In die Entwicklung und Anpassung von Korrekturverfahren sollen auch Überlegungen zum Zusammenhang zwischen Ausgangsqualität des OCR-Volltextes und „Korrekturpotential“ einfließen, um das in diesem Bereich häufige Problem der *false negatives*, also Veränderung des eigentlich korrekt erkannten Text zu vermeiden.

Die zu entwickelnden Verfahren sollen es darüber hinaus ermöglichen, mehrere OCR-Ergebnisse für den gleichen Text zu vereinigen und dabei durch Auswahl der geeigneten Zeichen und Wörter aus der einen oder anderen Textversion einen möglichst korrekten Text zu konstruieren. Die Länge der zu vereinigenden Texte soll dabei variabel sein, so dass eine Anwendung der entwickelten Methoden auf Seiten-, Absatz- oder Zeilenebene möglich ist. Es ist zu beachten, dass die zu entwickelnden Verfahren auch auf mehr als zwei Textversionen anwendbar sein sollen.

Eine weitere Anwendungsperspektive der innerhalb des Modulprojekts entwickelten Lösungen besteht im Spannungsfeld zwischen manueller (community-basierter) Nachkorrektur, Langzeitarchivierung und Reprozessierung von OCR. Hier kann die Textvereinigung wertvolle Beiträge zur Zusammenstellung bzw. Erhaltung der optimalen Textfassung leisten. Die Problemstellung der Textvereinigung kann dabei als Instanz des Alignierungsproblems von Zeichenketten bzw. allgemein der Sequenzalignierung betrachtet werden. Neben den Eingabetexten sollen dabei auch Metadaten wie z. B. die verwendete OCR-Software bzw. das OCR-Modell, Eigeneinschätzung der OCR-Software zur Erkennungsgüte oder Angaben zu Sprache, Schriftart und Entstehungszeitraum des zugrundeliegenden Digitalisats zu Rate gezogen werden. Der Einsatz von Sprachmodellen, (historischen) Wörterbüchern und anderen externen Ressourcen bei der Umsetzung der Zielvorgabe wird empfohlen. Die Leistungsfähigkeit der entwickelten Verfahren ist anhand von repräsentativen Beispielen im Vergleich zu fehlerfreiem Volltext zu demonstrieren.

4. Modul 4: Textoptimierung

4.1. Hintergrund

Heutige OCR-Lösungen können für Vorlagen aus unterschiedlichen Jahrhunderten universell eingesetzt werden. Dafür sind sie mit verschiedenen statistischen Modellen für unterschiedliche Anwendungsszenarien ausgestattet. Darüber hinaus ist es möglich, auf Basis vorhandener Modelle und/oder speziellem Trainingsmaterial neue OCR-Modelle anzulegen, mit deren Hilfe eine signifikant höhere Textgenauigkeit bei der automatischen Texterkennung erreicht werden kann als mit Standardmodellen.[2] Der Aufwand solche Modelle zu erstellen hängt von verschiedenen Faktoren ab: Diese sind u. a. die verwendete OCR-Software, Umfang des Trainingsmaterials, Spezifika des zu erkennenden Textmaterials sowie Anspruch und Vorgaben für die Textgenauigkeit. Der zeitliche Aufwand des Trainings kann als nicht unerheblich festgestellt werden.

Für die Erkennung historischen Textmaterials sind die den OCR-Programmen beigegebenen Modelle oftmals nicht geeignet, da diese nicht auf entsprechendem Material trainiert werden. Dieser Effekt verstärkt sich mit zunehmendem Alter der zu erfassenden Drucke. Dennoch ist eine hochqualitative OCR auch für frühe Druckerzeugnisse des 15. (Inkunabeln), 16. und 17. Jahrhunderts, unter der Voraussetzung, dass geeignete Texterkennungsmodelle vorliegen, möglich.

Verschiedene OCR-Implementierungen können in zwei generelle Ansätze unterschieden werden. Zeichenorientierte Ansätze fokussieren die Erkennung der einzelnen Zeichen durch Abgleich mit einem im Modell repräsentierten Glypheninventar während segmentierungsfreie Ansätze auf wesentlich kleineren Einheiten basieren, die aber auf Zeilenebene klassifiziert werden. Für beide Vorgehensweisen bedarf es unterschiedlicher Trainingsmodelle, deren Erstellung aber vergleichbaren Prozeduren folgt.

Mit der Erfindung des Buchdruckes mit beweglichen Lettern nach Gutenberg hat sich in den darauffolgenden Jahrhunderten die Drucktechnik bei gleichzeitiger Standardisierung wesentlich weiterentwickelt. Trotz dieser fortwährenden Normierung gerade bei den verwendeten Schrifttypen liegt eine Vielzahl an unterschiedlichen Formen vor, die gerade zeichenorientierte OCR-Programme vor Probleme stellen.

Trotz entsprechender Bemühungen beginnend beim Typenrepositorium (http://tw.staatsbibliothek-berlin.de/ – Stand, 15.04.2016) der Wiegendrucke bis hin zur Verzeichnung von Buchdruck und Schriftproben in Wikisource (https://de.wikisource.org/wiki/Buchdruck/Schriftproben – Stand, 15.04.2016) bleibt ein systematisch geordnetes und umfangreiches Verzeichnis von Schriftarten, Glyphen- und Druckereikatalogen bis heute ein Desiderat. Ein solches Verzeichnis bildet vor dem Hintergrund der anstehenden Herausforderungen bei der Massenvolltextdigitalisierung jedoch die Grundlage für ein konzertiertes, standardisiertes und damit nachnutzbares Modelltraining.

4.2. Teilaufgabe 4.A: Trainingsinfrastruktur

Mit diesem Modul soll die Erstellung und Verbreitung spezifischer Trainingsmodelle vereinfacht und standardisiert werden. Dazu soll ein modulares Client-Server-basiertes Serviceframework implementiert werden, dass die zum OCR-Training notwendigen Arbeitsschritte auf einfache Art und Weise zugänglich macht. Dabei sind folgende Funktionalitäten umzusetzen:

- Hilfestellungen für die Erstellung und Pflege,
- Hilfestellungen für Sammlung, Erfassung und Speicherung sowie
- Hilfestellungen für Verbreitung und Verfügbarkeit von spezifischen Trainingsmodellen.

Der zu entwickelnde Service muss in der Lage sein, das OCR-Training mit seinen notwendigen Arbeitsschritten auf einfache Art und Weise zu unterstützen.

Die Grundlagen des Trainings der verschiedenen verfügbaren OCR-Programme sind sehr ähnlich und bestehen im Kern aus der Alignierung von manuell erfasstem Text und dazugehörigem Bildmaterial. Der Trainingsprozess besteht aus mehreren Schritten von der Bereitstellung der Bilder und Texte über die Alignierung auf Zeilen- oder Zeichenebene bis hin zur eigentlichen Modellinduktion. Diese Schritte sind teilweise zeit- und rechenintensiv und weichen für die einzelnen OCR-Programme in Details voneinander ab.

Das Training innerhalb dieses Dienstes soll als ein abstrakter, universeller Workflow verstanden werden, der am Beispiel verfügbarer OCR-Programme instanziiert wird.

Innerhalb von OCR-D erfolgt dabei eine Konzentration auf sog. segmentierungsfreie Ansätze auf Basis neuronaler Netze, da aktuelle Forschungsergebnisse und Erkenntnisse aus Projekten zum Thema OCR einen Paradigmenwechsel bei den der Texterfassung zugrundeliegenden Verfahren konstatieren: Erkennungsroutinen auf Basis künstlicher, neuronaler Netze, die die Texterfassung als Sequenzklassifizierungsaufgabe modellieren, erzielen im Vergleich zu zeichenfokussierten Ansätzen eine deutlich höhere Textgenauigkeit insbesondere auf schwierigen Vorlagen wie historischem Material oder Handschriften.

Zwei der populärsten OCR-Programme, OCRopus und Tesseract sind Vertreter dieses Paradigmas und sind vorrangig in der Trainingsinfrastruktur zu adressieren. Über wohldefinierte Schnittstellen sollen aber auch weitere OCR-Lösungen einfach integrierbar sein. So ergibt sich die Möglichkeit OCR-Modelle für verschiedene OCR-Programme auf der gleichen Materialbasis zu erzeugen. Optionen einer zeitlichen Optimierung bspw. durch Vorschlagsfunktionen bei der Segmentierung oder Parallelisierung bei der Modellinduktion sollen eruiert werden. Neben dem Modelltraining mit manuell oder halbautomatisch erstellten Text-Bild-Alignierungen werden heute viele OCR-Modelle mit Hilfe synthetischen Trainingsmaterials auf Basis von Volltext und Schriftartdateien trainiert. Da der Aufwand der Modellerstellung dadurch signifikant reduziert wird, ist auch diese Variante des Modelltrainings zu implementieren. Dazu sind innerhalb der Trainingsinfrastruktur Werkzeuge zur Synthese von Schriftarten, Texten, Digitalisten bereitzustellen.

4.3. Teilaufgabe 4.B: Mikrotypographisches Formeninventar

Ziel dieser Teilaufgabe ist es, im ersten Schritt ein mikrotypographisches Formeninventar auf Grundlage des digitalen Bestandes der Drucke des 16. - 18. Jahrhunderts zu entwickeln, auf deren Grundlage ein planmäßiges OCR-Training realisiert werden kann. Der Schwerpunkt der Entwicklung ist dabei nicht auf buchwissenschaftliche Fragestellungen wie die Identifikation von einzelnen Druckern oder ihren Werkstätten sowie druckgeschichtliche Ausarbeitungen zu legen, sondern vorrangig auf Aspekte, die vor allem für die automatische Texterkennung relevante Form- und Typenunterscheidungen bzw. -gemeinsamkeiten herausarbeitet.

Für diese Art der Zusammenstellung ist ein Überblick über die einzelnen Anteile der Schriftarten am Gesamtumfang der VD-Drucke zu ermitteln. Das Ergebnis der Erfassung ist in einer Typ-Formen-Datenbank mit entsprechenden Metadaten formal zu beschreiben und mit Verweisen auf entsprechende beispielhafte VD-Digitalisate zu versehen.

Im zweiten Schritt bildet das mikrotypographische Formeninventar die Grundlage für technische Lösungen, die in der Lage sind, automatische Vergleiche (Vgl. z.B. https://www.fontspring.com/matcherator – Stand, 15.04.2016) zwischen den Schriftarten zu realisieren sowie die Möglichkeit bieten, aus den als Bild übergebenen Schriftartenbeispielen maschinenlesbare Fontdateien zu erzeugen und mit den ermittelten Metadaten zu versehen. Diese Aufgabe sollte in enger Zusammenarbeit mit der Teilaufgabe Trainingsinfrastruktur realisiert werden, da sie die Grundlagen für das sog. synthetische Training der OCR-Programme darstellen kann.

4.4. Teilaufgabe 4.C: Modellrepositorium

Um die tatsächliche Produktivität der in OCR-D zu entwickelnden Lösungen für die Massenvolltextdigitalisierung von VD-Titeln sicherzustellen, sind verfügbare, geeignete OCR-Modelle eine der wichtigsten Voraussetzungen. Innerhalb dieser Teilaufgabe sollen diese auf Basis der vom Koordinierungsgremium bereitgestellten Ground-Truth-Daten und unter Verwendung der Ergebnisse der Teilaufgabe Mikrotypographisches Formeninventar erstellt werden.

Eine wichtige Teilaufgabe besteht dabei in der Ermittlung eines sinnvollen Verhältnisses zwischen praktikabler Überschaubarkeit der Modellmenge und notwendiger Spezifizität sowie passender Parameterbelegungen für verschiedene OCR-Programme.

Das "Typenrepertorium der Wiegendrucke" (TW) verzeichnet derzeit allein für die Inkunabelzeit, ca. 6000 verschiedene Drucktypen, darunter ca. 4700 gebrochene Schriften und ca. 1000 Antiquatypen (vgl. http://tw.staatsbibliothek-berlin.de/html/alltypes.xql). Dabei versteht das TW unter "Type" jeweils individualisierbare, d.h. in einer bestimmten Offizin nachweisbare Lettersätze (engl. *founts*). Diese können auf den gleichen Schriftschnitten (engl. *cut*) beruhen und unterscheiden sich daher z.T. nur durch kleinste Details (einzelne Glyphen, Interpunktionszeichen usw.). Die Anzahl der im 15. Jh. angefertigten Schriftschnitte kann mit Hilfsmitteln wie dem TW nicht ermittelt werden. Zurzeit liegen nur Schätzungen vor, dass im 15. Jh. maximal 1000 verschiedene Schriftschnitte verwendet wurden. Diese wiederum sind in vielen Fällen sehr ähnlich. Ab dem 16. Jh. verringert sich die Diversität der Schriftschnitte durch den sich etablierenden Schriftenhandel. So werden zwar mehr Schriftenlettern gegossen, die aber von wenigen Schriftschnitten abstammen. Für das 17. bis 18. Jahrhundert ist eine ähnliche Situation festzustellen. Im 19. Jh. nimmt die Zahl der Schnitte durch neue Produktionstechniken wieder zu, das betrifft aber v.a. Auszeichnungs- und Zierschriften, die nach historischen Vorbildern nachgeschnitten werden.

Daraus ergibt sich, dass für das Modell-Training nicht für jede Schriftart bzw. jede Gruppe von Schriftarten ein eigenes Modell trainiert werden muss sondern der Ansatz eines Modell-Trainings auf Grundlage von gemischte Schriftvorlagen vorzuziehen ist. Eine abschließende Güteprüfung der generierten Modelle soll deren Funktionalität sicherstellen.

Die Erfassung der Metadaten für die Modelle erfolgt auf der Grundlage von Trainingsparametern und verwendetem Ground Truth. Um die Nachnutzung von OCR-Trainingsergebnissen zu erhöhen, soll mit Unterstützung des Koordinierungsgremiums ein Repositorium zur Archivierung, Versionierung und Veröffentlichung der OCR-Modelle angelegt und die Daten eingespeist werden. Hierfür soll ein Metadatenschema zur Beschreibung und Nachnutzung der Modelle entwickelt werden. Trainingsinfrastruktur und Modellrepositorium müssen in geeigneter Art und Weise miteinander interagieren: Generierte Modelle sollen direkt ins Repositorium übertragen und so nachnutzbar veröffentlicht werden.

5. Modul 5: Langzeitarchivierung (LZA) und Persistenz

5.1. Hintergrund

Neben dem generellen Wunsch das digitalisierte Material nachhaltig sicherzustellen, ist die Langzeitverfügbarkeit der gewonnenen Texte zwingende Voraussetzung für die Überprüfbarkeit wissenschaftlicher Ergebnisse, die auf der Auswertung von OCR-Texten beruhen.

Für die Planung und Einrichtung der Langzeitarchivierung (LZA) bedeutet dies, dass neben dem Bilddigitalisat auch der erkannte Text mit seinen Metadaten sowie seinen möglichen Bearbeitungen archiviert werden muss. Darüber hinaus ist eine persistente sowie möglichst granulare Adressierung des OCR-Materials notwendig. Die enge Verknüpfung des erkannten

Textes mit dem Bilddigitalisat, notwendig nicht zuletzt für das sogenannte Highlighting, ist auf Dauer zu sichern. Ebenso ist eine vom Bilddigitalisat unabhängige Verwendung der Textdaten zu gewährleisten.

Anders als das Bilddigitalisat unterliegt der erkannte Text zukünftigen Änderungen. Die maschinelle Texterkennung funktioniert nicht fehlerfrei, so dass Nachkorrekturen erforderlich sind. Maschinelle und/oder intellektuelle Nachkorrekturen sowie mögliche Neuprozessierungen (bei Verbesserung der Trainingsmodelle sowie bei wesentlicher Verbesserung der Erkennungstechnik) sorgen für Änderungen.

Bei der Neuprozessierung von OCR-Text stellt die Übertragung zuvor erfolgter intellektueller Korrekturen, die sich auf den vorausgehenden OCR-Text beziehen, eine große Herausforderung dar. Intellektuelle Korrekturen sind derzeit generell höherrangig zu bewerten als ein neuprozessierter OCR-Text mit verbesserter OCR-Technik.

In der Regel existieren bereits Workflows für die LZA von Bild-Digitalisaten. Evtl. vorhandener OCR-Text wird z.B. als Teil des digitalen Objektes mitgesichert. Veränderungen bzw. Neuprozessierungen spielen bislang eine eher untergeordnete Rolle bzw. ersetzen diese in der Regel dann die älteren Versionen vollständig. Als Lösungen, bei denen Nachlieferungen zu Objekten beschrieben werden, sind z.B. Delta-Pakete im Einsatz.

Die Aufgabe beschreibt zusätzlich zur Langzeitarchivierung Komponenten eines Forschungsdatenmanagement, weswegen die Prüfung von Methoden und Strategien für den Umgang mit Forschungsdaten angeregt wird.

5.2. Ziele

Mit dieser Ausschreibung sollen Lösungen erarbeitet werden, welche die Nachhaltigkeit von OCR-Texten und ihren Bearbeitungen (Korrekturen) sicherstellen und eine persistente Identifikation (mind. bis auf Seitenebene) gewährleisten. Besonderes Augenmerk ist dabei auf die vorkommenden Veränderungen des Materials (hervorgerufen durch Neuprozessierungen sowie Korrekturen) zu richten sowie Antworten auf die Frage nach den damit notwendigen Versionierungen zu finden.

5.3. Teilaufgabe 5.A: LZA/Datenmanagement-Konzept

Klärung der Anforderungen der Wissenschaft an OCR-Texte: Eine grundsätzliche Forderung der Wissenschaft ist der transparente Zugang zum OCR-Text. Dabei ist zu beachten, dass der Text differenziert zu betrachten ist und Aspekte der Erfassung der Textstruktur von zentraler Bedeutung sind. Die akzeptierte Textgenauigkeit ist von den Anwendungsfeldern und spezifischen Fragestellungen abhängig. Bei einer Korrektur des Textes ist es von besonderer Wichtigkeit, dass Veränderungen eines Textes auch rückwirkend verfügbar sein sollten. Dies belegen auch Zwischenergebnisse der laufenden Umfrage des Koordinierungsgremiums von OCR-D “Umfrage zur Verwendung von OCR-Texten” (http://www.ocr-d.de/?q=node/7), in der sich eine Mehrzahl der Befragten dafür ausspricht, dass jegliche Veränderungen des Textes nachvollziehbar sein sollten. Die Veränderungen können dabei auch als einzelne Textversionen gespeichert werden.

Dabei sind folgende Fragen zu beantworten:

- Was sind die Objekte, die eindeutig beschrieben, referenziert und nachvollziehbar bleiben müssen, also archiviert werden müssen?
- Welche Metadaten sind mit zu archivieren?
- Wie ist eine Version eines sich immer wieder ändernden Objektes zu definieren?

Die Lösungen für die Archivierung sollten in enger Abhängigkeit zu den Anforderungen an die Nutzung erarbeitet werden. Das Präsentationssystem hält stets den jeweils aktuellsten Stand der OCR-Texte bereit. Dort erfolgt die Nachkorrektur bzw. dort werden die Ergebnisse der Nachkorrektur bereitgestellt.

Neben der Nutzung einzelner Texte (z.B. für Volltextsuche) ist das Interesse der Forschung an Textkorpora, wie der Zusammenstellung einer Textgruppe aus Einzeltexten zu berücksichtigen. Lösungen zur Bildung und Beschreibung solcher Korpora im Rahmen des LZA-Konzepts sind anzustreben. Eine enge Verzahnung von Präsentation, Korrekturinstanz und LZA erscheint sinnvoll. Neben der reinen Langzeitarchivierung ist die Langzeitverfügbarkeit der verschiedenen Versionen der OCR-Texte und -Korpora für den Nutzer sicherzustellen.

Konzept: Es ist ein organisatorisches und technisches Konzept für die LZA der OCR-Texte zu erarbeiten. Zentrale und dezentrale Ansätze sind dabei zu evaluieren, Lösungen für das Einsammeln der Daten von verschiedenen Orten zu erarbeiten, Fragen nach Konsistenz und Redundanz der Daten in verteilten Systemen zu beantworten. Bedacht werden sollte auch eine automatische Verknüpfung der Texte mit der VD-Nachweisstruktur, vorzugsweise unter Nutzung der bibliographischen Nummern.

Umsetzung: Die Machbarkeit des Konzepts ist in einer prototypischen Umsetzung nachzuweisen. Der technische, personelle und finanzielle Aufwand für die in den Verzeichnissen der im deutschen Sprachraum erschienenen Drucke (VDs) ist abzuschätzen.

5.4. Teilaufgabe 5.B: Persistenz

Für die eindeutige und persistente Identifizierung von Objekten ergeben sich neue Anforderungen, z.B. durch das Zusammenfügen mehrerer bibliographischer Einheiten zu Korpora einerseits, oder durch Adressierung von Bereichen unterhalb von bibliographischen Einheiten, z.B. Kapitel, Seiten, Seitensegmente etc. andererseits. Ausgehend von den in der Regel bereits vorhandenen *persistent identifiers* (meist URNs) auf der Ebene von bibliographischen Einheiten werden Lösungen gesucht sowohl für den Makrobereich darüber (z.B. Korpora), als auch für den Mikrobereich darunter.

Darüber hinaus sind Lösungen für die eindeutige und persistente Identifizierung der durch Korrektur oder Neuprozessierung hervorgerufenen Versionen eines Objektes zu erarbeiten. Die in DARIAH aufgebaute *Collection Registry* und CLARINs *Virtual Language Observatory* (VLO) sind auf ihre Vorbildfunktion dafür ebenso zu prüfen wie bspw. die im Rahmen von URN:NBN, Datacite, DOI vorhandenen Funktionalitäten.

6. Modul 6: Qualitätssicherung

6.1. Hintergrund

Das Funktionsmodell (vgl. Anhang) skizziert einen beispielhaften Ablauf der einzelnen Prozesse der automatischen Layout- und Texterfassung. Das Potential der einzelnen Komponenten wird in wissenschaftlichen Kontexten anhand von Ergebnisvergleichen mit manuell bearbeitetem Material, das nicht zur Entwicklung bzw. zum Training der jeweiligen Verfahren beigetragen hat in gewisser Weise objektiv bestimmt. Im Bereich der Massenvolltextdigitalisierung stellen derartige Messungen jedoch keinesfalls den Erfolg der kombinierten Anwendung also des finalen Struktur- und Textergebnisses sicher. Die Heterogenität des Ausgangsmaterials samt vieler spezifischer, jeweils selten auftretender Degradierungsphänomene, die typischerweise in Forschungsarbeiten nicht betrachtet werden, da dort ja gerade die Leistungsfähigkeit für prototypische Daten optimiert wird, lässt im Gegenteil erwarten, dass einzelne Komponenten mitunter kein sinnvolles Ergebnis erzielen.

Ein Vergleich mit dokumentspezifischen Ground-Truth-Daten ist in der massenhaften Anwendung nur sehr begrenzt einsetzbar.[3] Für eine Abschätzung der Qualität des Struktur- und Textergebnisses ist es daher notwendig, die Qualität der einzelnen Teilergebnisse verlässlich einschätzen zu können. Diese Einschätzung kann darüber hinaus dazu dienen, Fehler im System zu diagnostizieren und zu lokalisieren und gegebenenfalls eine Reprozessierung mit veränderten Parametern auszulösen. Nicht zuletzt können derartige Qualitätsangaben dem Endnutzer der Texte helfen, deren Verwendungsmöglichkeit für die eigene Forschung zu bestimmen.

6.2. Ziele

Die erwarteten Entwicklungen in diesem Modul haben zwei Dimensionen: Zum einen sollen Metriken zur Qualitätseinschätzung ohne dokumentspezifische Ground-Truth-Daten konzipiert werden. Zum anderen sind Verfahren zu entwickeln, die die Qualitätsmessung auf Basis dieser Metriken prozessfähig machen und deren automatische Erhebung ermöglichen. Die Teilaufgaben dieses Moduls ergeben sich somit direkt aus den einzelnen, zum Gesamtergebnis führenden Arbeitsschritten und betreffen somit die Ebenen Vorverarbeitung, Layoutanalyse, Volltexterstellung und Textoptimierung. Neben der Evaluierung der Teilergebnisse soll auch deren Verrechnung zu einem Gesamtergebnis realisiert werden.

Das Modul Qualitätssicherung zeichnet sich durch einen hohen Interaktionsgrad mit den anderen Modulen aus, da die gestellten Aufgaben vermutlich nicht methodenunabhängig umgesetzt werden können. Die vom Koordinierungsgremium zur Verfügung gestellten Ground-Truth-Daten sind in die Überlegungen einzubeziehen. Eine besondere Herausforderung besteht in der Tatsache, dass anders als in den anderen ausgeschriebenen Modulen die Anzahl an verfügbaren Vorarbeiten und wissenschaftlichen Veröffentlichungen im Bereich Ground-Truth-freie Evaluierung äußerst gering ist.

Allgemeine inhaltliche und technische Rahmenbedingungen für die Modulprojekte

Bei der Entwicklung und Realisierung der Lösungen sind einschlägige technische Standards und disziplinspezifische Regelwerke zu berücksichtigen. Im Folgenden wird auf einzelne Aspekte näher eingegangen.

Vorarbeiten

Es wird erwartet, dass sich die geförderten Projekte intensiv und umfänglich mit in der wissenschaftlichen Literatur vorgeschlagenen Verfahren auseinandersetzen und diese für die jeweils beschriebene Problemstellung evaluieren. Funktionale Lösungen sollen in den Status der Praxisreife gebracht und dabei, wo nötig, um neue Methoden ergänzt werden. Die Ausschreibung eröffnet Wissenschaftlerinnen und Wissenschaftlern im Rahmen des Projektes somit sowohl die Möglichkeit, eigene kreative und innovative Ansätze umzusetzen, als auch vorhandene Lösungen bzgl. ihrer Skalierbarkeit und Domänenadaptivität zu optimieren.

Datenformate

Im Folgenden werden In- und Outputdatenformate genannt, die im Schwerpunkt zu unterstützen sind. Nicht für alle in der Ausschreibung adressierten Teilaufgaben existieren spezifische und etablierte Formatstandards. Ein wichtiges Ziel von OCR-D ist die Entwicklung entsprechender Vorschläge auch mit Blick auf die Weiterentwicklung der DFG-Praxisregeln „Digitalisierung“ (http://www.dfg.de/formulare/12_151/12_151_de.pdf). Hier ist eine enge Abstimmung zwischen Modulprojektnehmern und Koordinierungsgremium unabdingbar.

Datenformate der Digitalisate	Gängige Grafikformate sind zu unterstützen, einen Schwerpunkt bilden die Formate (komprimiert/unkomprimiert): TIFF, JPEG2000, PNG, JPEG
Datenformate für Texte	ALTO, PAGE-XML, XML (z. B. TEI), TXT
Font-Formate	PostScript-Fontformate, TTF, OTF
Metadaten	• Bei der Verzeichnung und Dokumentation von Metadaten ist in Abstimmung mit dem Koordinierungsgremium ein geeignetes XML-basiertes Format zu wählen. • Alle Metadatenformate müssen XML-basiert sein. • Entwickelte Metadatenformate sollten in der Regel in das METS-Framework integrierbar sein.

Implementierung

- Die Entwicklung bzw. Implementierung einer entsprechenden Softwarelösung soll in einer gängigen Programmiersprache erfolgen.
- Notwendige Kompilierungs- bzw. Installationshinweise sind zur Verfügung zu stellen.
- Entsprechende externe Programmbibliotheken die als Abhängigkeiten bestehen sind aufzulisten. Es dürfen nur frei nachnutzbare externe Programmbibliotheken verwendet werden.
- Für die Lösungen sind als Zugriffsmethoden eine Programmierschnittstelle (API), ein Kommandozeilenprogramm sowie ein REST-basierter Webservice umzusetzen.
- Zur Sicherung der Interoperabilität der einzelnen Modulprojekte werden den einzelnen Projektnehmern konkrete Anforderungsprofile für Schnittstellenfunktionalitäten vorgelegt.
- Die Gesamtintegration erfolgt innerhalb eines vom Koordinierungsgremium vorgegebenen Frameworks.
- Die Teilaufgaben sollen sowohl einzeln als auch als Gesamtprozess innerhalb des Frameworks ansprechbar sein.
- Als Ein- und Ausgabeformate sind XML-Standardformate vorzusehen.
- Die Lösungen orientieren sich an den Prinzipien der versionierbaren Softwareentwicklung.

Arbeitsversionen und Entwicklertreffen

- Wir definieren vier Meilensteine für die in den Modulprojekten entstehenden Implementierungen:
  (1) Pre-Alpha-Version (Weitgehend funktionsfrei, zur Umsetzung der Schnittstellendefinition)
  (2) Alpha-Version (Umsetzung der Grundfunktionalitäten)
  (3) Beta-Version (vollständiger Funktionsumfang, stabil lauffähig)
  (4) Release Candidate (qualitativ und quantitativ optimierte Version
- Die Modulprojekte werden frühzeitig in eine vom Koordinierungsgremium organisierte, konzertierte Versionsentwicklung eingebunden.
- Die Meilensteine sind mit der Übergabe von Arbeitsversionen an das Koordinierungsgremium verbunden.
- Bereits diese Arbeitsversionen sollen die oben erwähnten Anforderungsprofile umsetzen, so dass die Arbeit an der Implementierung des Gesamtworkflows frühzeitig beginnen kann.
- Die Funktionalität der einzelnen Bausteine soll mit jedem Meilenstein innerhalb der Modulprojekte ausgebaut werden.
- Konkret wird dieser Ablauf anhand der eingehenden Anträge gemeinsam von Koordinierungsgremium und Projektnehmer definiert.
- Nach Übergabe der einzelnen Arbeitsversionen finden Integration und Tests durch das Koordinierungsgremium statt, deren Ergebnisse jeweils bei einem Entwicklertreffen zwei Monate nach Abgabedatum besprochen werden

Lizenzierung

- Alle durch die Modulprojekte zustande gekommenen Ergebnisse sind in der Fachöffentlichkeit bekannt zu machen und kostenlos zur Nachnutzung auch durch Dritte zur Verfügung zu stellen.
- Die Offenlegung der ggf. produzierten Quellcodes ist verpflichtend, die Bereitstellung der Projektergebnisse als „Open Source “ an geeigneter Stelle wird vorausgesetzt.
- Das schließt die umfassende Dokumentation mit ein.
- Mit der Lizenzierung unter der Apache Software License 2.0 (https://www.apache.org/licenses/LICENSE-2.0, Software) bzw. CC-BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/, Dokumentation und Daten) ist die freie Nachnutzbarkeit aller Projektergebnisse zu gewährleisten.

Dokumentation

Softwaredokumentation – Folgenden Grundregeln sind bei der Softwaredokumentation zu beachten. Sowohl die Software als auch die Schnittstellen (API) sind entsprechend zu dokumentieren:

- Die Dokumentation ist in Englisch abzufassen.
- Die Dokumentation ist nach Möglichkeit in den Quellcode einzuarbeiten. Dafür sind Kommentare zu nutzen, die in unmittelbarer Nähe der Programmanweisungen anzubringen sind.
- Zusammenfassungen des Programmcodes sind mit Dokumentationswerkzeugen (bspw. Javadoc oder Doxygen) automatisch zu erstellen.
- Zusätzliche sowie näher erläuternde Dokumentationsteile, die u. a. grafische Strukturbäume, Beschreibungen oder andere Übersichten beinhalten sind der Dokumentation beizugeben.

Benutzerdokumentation

- Die Benutzerdokumentation wendet sich an Endanwender der entwickelten Lösung.
- Die Dokumentation ist in Englisch abzufassen.
- In kurzen Sätzen sind die Funktionalitäten sowie der Umgang formal zu beschreiben.
- Die Dokumentation ist im Format DITA (http://docs.oasis-open.org/dita/dita/v1.3/dita-v1.3-part3-all-inclusive.html) abzufassen. Zur Erstellung von Präsentationsformen kann bspw. das DITA Framework des <oxygen/> xml Editor (http://www.oxygenxml.com/) oder das DITA Open Toolkit (http://www.dita-ot.org/) genutzt werden.

Referenzdaten

Unter Referenzdaten werden Trainings- und Evaluationsdaten verstanden. Für die einzelnen Module werden den Projektnehmern zum Zweck a) der Aufwandsabschätzung, b) des Modell-Trainings und c) der Evaluation der Leistungsfähigkeit der Modul-Ergebnisse entsprechende Daten zur Verfügung gestellt.

Die Evaluation der Leistungsfähigkeit wird nach mit den Projektnehmern abgestimmten, wissenschaftlichen Maßstäben vom Koordinierungsgremium durchgeführt.

Die Referenzdaten umfassen ein Ground-Truth-Korpus und weitere Spezialkorpora. Das Ground-Truth-Korpus umfasst Seiten aus Publikationen aus dem Zeitraum 1500 - 1900. Der Inhalt des Korpus basiert auf einer gezielten Auswahl aus dem Bestand des DFG-Projektes „Deutsches Textarchiv“ (http://www.deutschestextarchiv.de), der Digitalisierten Sammlungen der Staatsbibliothek zu Berlin (http://digital.staatsbibliothek-berlin.de/) und der Wolfenbütteler Digitalen Bibliothek der Herzog August Bibliothek (http://www.hab.de/de/home/bibliothek/digitale-bibliothek-wdb.html). Bestände von Projekten und digitalen Sammlungen anderer Bibliotheken sowie zusätzliche Ground-Truth-Daten, die zusammen mit Modulprojektnehmern erarbeitet werden, können in Abstimmung mit dem Koordinierungsgremium in das Korpus als spezielle Erweiterungen aufgenommen werden. Die Zuständigkeit für die Erstellung der Ground-Truth-Daten liegt beim Koordinierungsgremium. Sollten für einzelne in den Modulprojekten vorgeschlagene bzw. erarbeitete Verfahren zusätzliche Annotationen notwendig sein, werden diese in Abstimmung mit den Projektnehmern im Rahmen des Koordinierungsprojekts erstellt.

Annotationstiefe, Textgenauigkeit und Artefakte

Das Ground-Truth-Korpus bietet drei Annotationstiefen an:

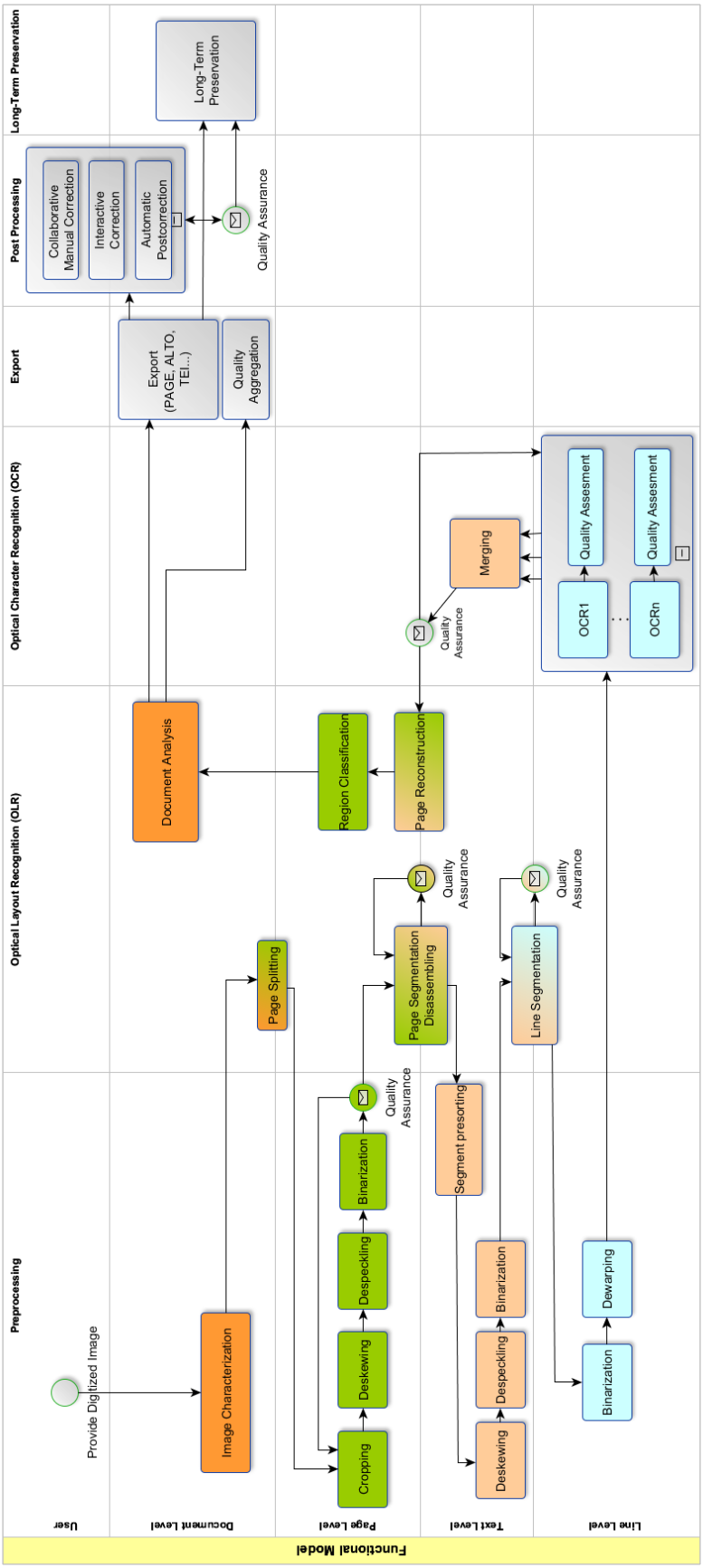
- Strukturregionen, Textzeilen, Wortkoordinaten
- Strukturregionen, Textzeilen
- Textzeilen

Die Spezialkorpora umfassen:

- Spezialkorpus von Daten geringerer Textgenauigkeit (schmutzige OCR), kann für einzelne Vergleiche und Evaluationen herangezogen werden.
- Spezialkorpus Artefakte: Dieses Korpus beinhaltet ausschließlich Objekte die Störungen aufweisen.

Funktionsmodell

Das Funktionsmodell stellt einen beispielhaften, prototypischen Workflow für die Volltextdigitalisierung dar, der alle im OCR-D betrachteten Module beinhaltet und zueinander in Beziehung setzt. Es soll somit vor allem als Schablone zur pragmatischen Einordnung der einzelnen Teilaufgaben dienen und weniger eine konkrete Arbeitsanweisung für die Massenvolltextdigitalisierung darstellen. Die Darstellung, in Anlehnung an ein BPMN-Modell (Business Process Model and Notation), beschreibt in ihren Spalten die einzelnen Prozessebenen und in den Zeilen die jeweilige Bezugsebene im Digitalisat, die zusätzlich durch entsprechende Farbgebung kenntlich gemacht werden. Das Funktionsmodell enthält auch mögliche Ansatzpunkte für die Ground-Truth-freie Qualitätssicherung, die unter bestimmten Umständen Reprozessierungen auslösen kann. Die Mehrfachbearbeitung durch verschiedene Algorithmen ist beispielhaft für den Bereich der Texterkennung ausgeführt. Die tatsächliche funktionelle Implementierung des OCR-D-Gesamtworkflows wird sich am Funktionsmodell orientieren, diese aber an die Gegebenheiten der Modulprojektbeiträge anpassen.


Functional Model
User
Document Level
Page Level
Text Level
Line Level
Preprocessing
Optical Layout Recognition (OLR)
Optical Character Recognition (OCR)
Export
Post Processing
Long-Term Preservation
Provide Digitized Image
Image Characterization
Cropping
Deskewing
Despeckling
Binarization
Quality Assurance
Page Splitting
Segment presorting
Deskewing
Despeckling
Binarization
Page Segmentation Disassembling
Quality Assurance
Line Segmentation
Quality Assurance
Binarization
Dewarping
Region Classification
Page Reconstruction
Quality Assurance
Merging
OCR1
OCRn
Quality Assesment
Quality Assesment
Document Analysis
Export (PAGE, ALTO, TEI...)
Quality Aggregation
Collaborative Manual Correction
Interactive Correction
Automatic Postcorrection
Quality Assurance
Long-Term Preservation

Endnoten

[1] Ein vergleichbares Vorgehen wird beispielsweise bei der manuellen Texterfassung per Double Keying angewendet, wo zwei Bearbeiter den gleichen Text unabhängig voneinander abschreiben, und beide Versionen in einem nachfolgenden Arbeitsschritt vereinigt werden.

[2] Vgl. z.B. Federbusch und Polzin: „Volltext via OCR – Möglichkeiten und Grenzen“; Beiträge aus der Staatsbibliothek zu Berlin, Bd. 43; Berlin 2013, S. 9 oder Springmann, Lüdeling und Schremmer: „Zur OCR frühneuzeitlicher Drucke am Beispiel des RIDGES-Korpus von Kräutertexten“; Poster auf der DHd; Graz 2015.

[3] Die DFG-Praxisregeln „Digitalisierung“ empfehlen das Bernoulli-Experiment zur Feststellung, ob eine vom OCR-Dienstleister angegebene Textgenauigkeit erreicht wurde. Die Umsetzung des Bernoulli-Experiments ist, trotz Einsatzes von interaktiver Software zur Unterstützung der manuellen Auswertung, bislang zeitintensiv und daher für die Massenvolltextdigitalisierung nur bedingt geeignet.

Weitere Informationen

- Ausschreibungstext: http://www.dfg.de/download/pdf/foerderung/programme/lis/170306_ausschreibung_verfahren_volldigitalisierung.pdf
- Merkblätter für die Antragstellung: www.dfg.de/foerderung/formulare
- OCR-D Koordinierungsprojekt: http://ocr-d.de/
- OCR-D Ground-Truth-Daten (Training und Evaluation): http://www.ocr-d.de/daten

Ansprechpersonen

Bei Rückfragen und zur Beratung wenden Sie sich bitte an:

- Förderbedingungen und Förderfragen:
  Dr. Matthias Katerbow: Tel. +49 228 885-2358, Matthias.Katerbow@dfg.de

- Inhaltliche und organisatorische Fragen:
  OCR-D Koordinierungsprojekt
  Elisa Herrmann: Tel. +49 5331 808-306, Elisa.Herrmann@hab.de