Fachbeitrag

Felix Kirchner, Marco Dittrich, Phillip Beckenbauer und Maximilian Nöth

OCR bei Inkunabeln – Offizinspezifischer Ansatz der Universitätsbibliothek Würzburg

DOI 10.1515/abitech-2016-0036

Zusammenfassung: Im Rahmen des BMBF-geförderten Projekts KALLIMACHOS an der Universität Würzburg soll unter anderem die Textgrundlage für digitale Editionen per OCR gewonnen werden. Das Bearbeitungskorpus besteht aus deutschen, französischen und lateinischen Inkunabeln. Dieser Artikel zeigt, wie man mit bereits heute existierenden Methoden und Programmen den Problemen bei der OCR von Inkunabeln entgegentreten kann. Hierzu wurde an der Universitätsbibliothek Würzburg ein Verfahren erprobt, mit dem auf ausgewählten Werken einer Druckerwerkstatt bereits Zeichengenauigkeiten von bis zu 95 Prozent und Wortgenauigkeiten von bis zu 73 Prozent erzielt werden.

Schlüsselwörter: OCR, Tesseract, Inkunabel

OCR processing of incunabula: printshop-specific approach of the University Library of Würzburg

Abstract: In the context of the Kallimachos project at the University of Würzburg, the textual base for digital editions of incunabula is obtained via OCR. The corpus of texts to be worked on consists of German, French, and Latin opera. This article shows how problems with OCR of incunabula can be tackled with already existing methods and programs. The developed method focuses on setting up a type-specific OCR training to be reused with different medieval printings from one printshop. Following this method we achieved letter accuracies of up to 95 percent and word accuracies of up to 73 percent.

Keywords: OCR, Tesseract, Incunabula

1 Einführung – Herausforderungen bei der OCR von Inkunabeln

Im Rahmen des BMBF-geförderten Projekts *KALLIMACHOS*[1] an der Universität Würzburg entstehen innerhalb des Teilprojektes „Narragonien“[2] digitale Editionen von deutschen, lateinischen und französischen Ausgaben von Sebastian Brants „Narrenschiff“ (1494 bis 1509). Die hierfür benötigten Textgrundlagen sollen weitestgehend per OCR gewonnen werden.

Bisherige OCR Modelle, wie sie zum Beispiel von der Firma *ABBYY*[3] angeboten werden, zielen auf ein möglichst breites Schriftarteninventar ab, um ein großes Spektrum unterschiedlicher Dokumente mit möglichst wenig Trainingsaufwand erfassen zu können. Dieser OCR-Ansatz funktioniert auf modernen Druckvorlagen ab dem 19. Jahrhundert sehr gut, lässt sich aber nicht verlustfrei auf mittelalterliche Druckwerke übertragen. Bei alten Drucken verschiedener Jahrhunderte konnten im Rahmen einiger Projekte, wie IMPACT[4] oder EMOP[5], kleinere Fortschritte erzielt werden, jedoch noch keine allgemeingültige Lösung für Werke vor dem 17. Jahrhundert.[6] Einige OCR-Experten gehen sogar davon aus, dass sich eine OCR von Inkunabeln mit ausreichender Erkennungsgenauigkeit bei vertretbarem Arbeitsaufwand selbst mit aktuellen OCR-Programmen nicht bewerkstelligen lässt.[7]

1 http://www.kallimachos.de (09.05.2016).
2 http://www.presse.uni-wuerzburg.de/einblick/single/artikel/narrenschi/ (23.05.2016); http://www.damals.de/de/8/Narrenschiff-auf-digitalem-Kurs.html?aid=191576&cp=2&action=showDetails (23.05.2016).
3 http://www.frakturschrift.com/ (09.05.2016).
4 http://www.impact-project.eu/ (23.05.2016).
5 http://emop.tamu.edu/ (23.05.2016).
6 Federbusch, Maria, Christian Polzin: *Volltext via OCR – Möglichkeiten und Grenzen*. Berlin, 2013. 10 ff.
7 Rydberg-Cox, Jeffrey A.: „Digitizing Latin Incunabula: Challenges, Methods and Possibilities. “ In *Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure* 3, 1 (2009). (siehe http://www.digitalhumanities.org/dhq/vol/3/1/000027/000027.html#p7, 30.05.2016)

Der offizinspezifische OCR-Ansatz der Universitätsbibliothek Würzburg basiert auf der Tatsache, dass Offizinen[8] ihre Werke über einen langen Zeitraum mit den gleichen Drucktypeninventaren gedruckt haben. Dies liegt an den damals vergleichsweise hohen Herstellungs- bzw. Anschaffungskosten für alle Druckstempel einer Schrifttype. Deshalb wurden Druckstempel über einen längeren Zeitraum verwendet. Abgenutzte Stempel mussten während des Drucks eines Werkes fortlaufend durch neue ersetzt werden. Dies führte zum Gebrauch von Stempeln mit unterschiedlichen Abnutzungsgraden, die wechselnde Druckbilder für ein- und denselben Buchstaben erzeugten. Die Herstellung der Stempel unterlag größeren Fertigungstoleranzen, da sie von unterschiedlichen Personen, unter anderem auch Lehrlingen und Hilfskräften, manuell gefertigt wurden.

Für den Druck wurden die Stempel in hölzerne Druckstöcke gesetzt. Diese verzogen sich im Laufe der Zeit, da das Papier beim Druck befeuchtet werden musste. Die gesetzten Texte wurden per Hand in die Presse montiert, die Druckfarbe aufgetragen, das Papier eingelegt und manuell angepresst. Bedingt durch dieses Herstellungsverfahren erscheinen die Texblöcke einer Seite oftmals gegeneinander verdreht oder verzerrt.

Schmutzeinschlüsse und Wellen im feuchten Papier verschlechtern das Druckbild zusätzlich. Auswirkungen hiervon sind beispielsweise unscharfe Abdrücke, Schattenbildungen oder Druckfarbenflecke.

Durch die manuelle Drucktechnik des Mittelalters besitzen Inkunabeln also kein homogenes Druckbild. Ein Einsatz von OCR bei diesen Werken bedeutet eine automatische Verarbeitung von individuell gefertigten Werken durch regelbasiert agierende Maschinen. Der offizinspezifische OCR-Ansatz der Universitätsbibliothek Würzburg beruht auf der Annahme, dass zumindest bei Werken einer Offizin Regelmäßigkeiten in diesen Individualismen gefunden werden können.

Der im Folgenden vorgestellte Trainingsansatz versucht durch ein offizin- und drucktypenspezifisches Training von *Tesseract*[9] diese Probleme zu kompensieren.

2 Qualitätsanforderungen an Digitalisate

Obwohl inzwischen viele Inkunabeln und frühe Drucke digitalisiert sind und im Netz frei heruntergeladen werden können, reicht die hierbei zur Verfügung gestellte Qualität der Scans bzw. Bilddateien für ein hinreichend gutes OCR-Ergebnis oftmals nicht aus. Die Gründe hierfür reichen von schlechten Aufnahmebedingungen bis hin zu mangelnder Bildauflösung der frei zugänglichen Bilddateien.

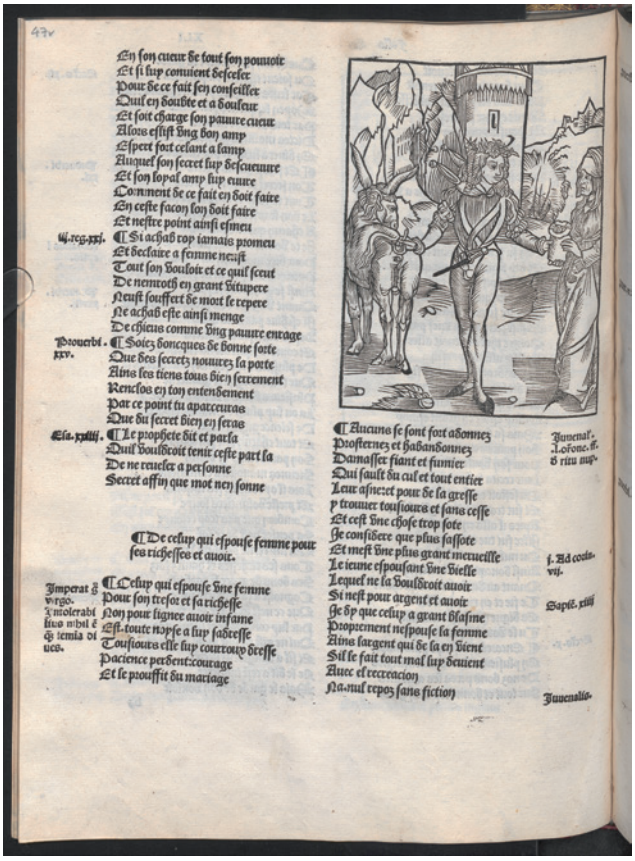

Abb. 1: Beispiel für Umbug im Buchfalz in GW 5058[10].

Ungeeignete Aufnahmebedingungen resultieren zum Beispiel in Verzerrungen im Buchfalz (vgl. Abb. 1) und nicht korrekt ausgerichteten Buchseiten. Eine weitere wichtige Voraussetzung für ein gutes OCR-Training ist die Auflösung der verwendeten Digitalisate. Bei im Netz frei verfügbaren Bildern liegt diese häufig unter 300 DPI, was die Erkennungsrate stark negativ beeinflusst.

8 Als eine Offizin bezeichnet man seit dem späten Mittelalter eine Werkstatt, die hochwertige Waren, wie zum Beispiel Bücher, produzierte. An die Werkstatt war häufig auch ein Verkaufsraum angeschlossen. (vgl. https://de.wikipedia.org/wiki/Offizin).

9 https://github.com/tesseract-ocr (09.05.2016).

10 Brant, Sebastian: *Das Narrenschiff*. Franz. von Pierre Rivière. Paris: [Jean Lambert für] Geoffroi de Marnef und Johann Philippi, 1497. 47v. (Österreichische Nationalbibliothek Wien, Ink 8.E.26 – GW 5058)

3 Bildvorverarbeitung

Ein wichtiger Bestandteil eines OCR-Workflows ist die Aufbereitung der Digitalisate. Da die Bilddateien die Grundlage jedes weiteren Folgeschrittes sind, hat jeder Verarbeitungsschritt direkte Auswirkungen auf das OCR-Ergebnis.

Auf Grund des relativ hohen Alters von Inkunabeln kann das Papier zahlreiche Gebrauchsspuren und Verschmutzungen aufweisen. Dies sind unter anderem Markierungen und Notizen, verursacht durch die intensive Nutzung, aber auch Flecken, verursacht durch Flüssigkeiten oder Schimmelbefall.

Die Binarisierung der Digitalisate ist ein wichtiger Schritt der Bildvorverarbeitung. Eine automatische Binarisierung ist in den meisten OCR-Programmen bereits integriert, kann jedoch auch als eigenständiger Vorgang durchgeführt werden. Für eine optimale Binarisierung bietet es sich an, diese als separaten Einzelschritt durchzuführen. Hierbei kann der Nutzer Einfluss auf das gewünschte Binarisierungsverfahren nehmen und das jeweils optimale Verfahren wählen. Zwei bekannte Vefahren sind Otsu[11] und Sauvola[12], wobei es inzwischen eine Vielzahl unterschiedlicher Methoden zur Binarisierung von Bildern gibt. Das jeweils optimale Verfahren muss durch Testen ausgewählter Seiten eines Werkes ermittelt werden. Im Workflow der Universitätsbibliothek Würzburg hat sich das Binarisierungsverfahren nach Sauvola bewährt und lieferte bei vielen Werken gute Ergebnisse.

Für optimale OCR-Ergebnisse muss der Text verzerrungsfrei und korrekt ausgerichtet sein. Dies kann durch Programme wie zum Beispiel *ScanTailor*[13] oder *Abbyy FineReader*[14] bewerkstelligt werden (siehe Abb. 2 und 3).

Eine große Sorgfalt bei der Beseitigung der oben genannten Bildfehler führt in der Regel zu deutlich besseren OCR-Ergebnissen und spart Zeit bei den folgenden Arbeitsschritten. Programme wie zum Beispiel *ScanTailor* unterstützen den Nutzer bei der Bildvorverarbeitung.

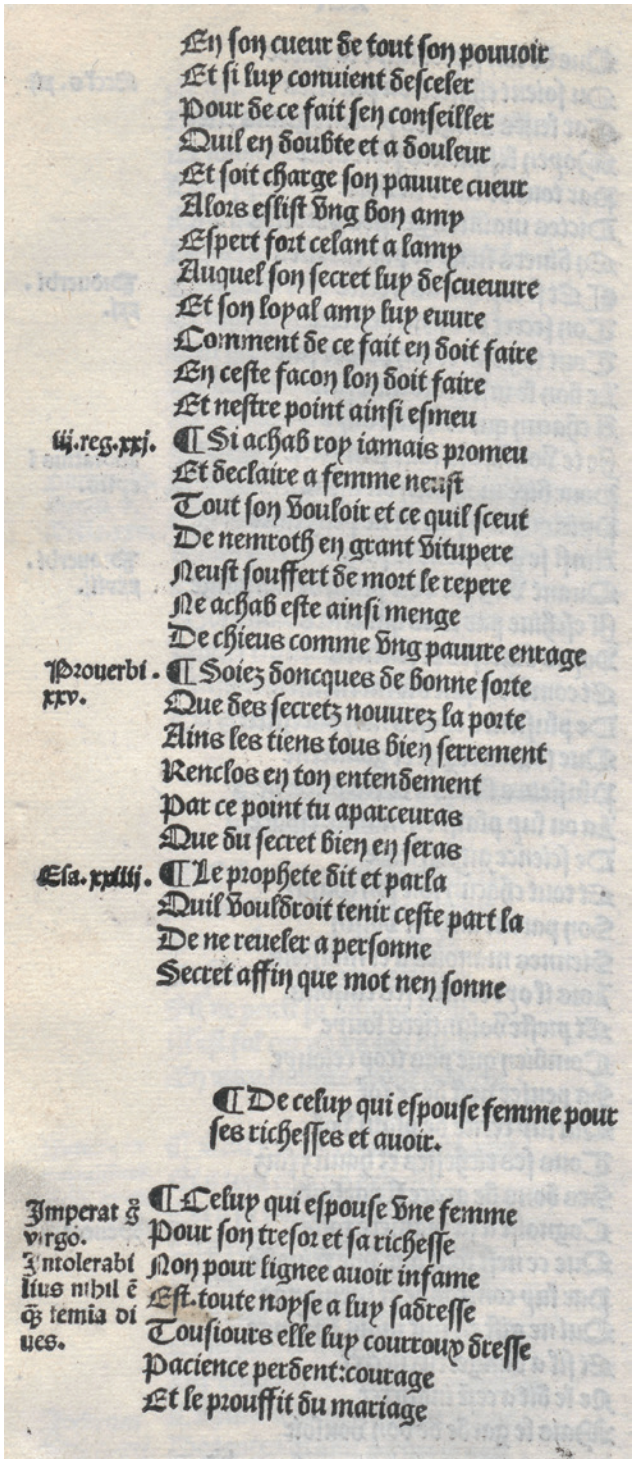

Abb. 2: Originalbild mit Bildstörungen[15].

11 Otsu, N.: „A threshold selection method from gray-level histograms." In *IEEE Transactions on Systems, Man and Cybernetics* 9, 1 (1979): 62–66.

12 Sauvola, J., M.Pietikainen: „Adaptive document image binarization." In: *Pattern Recognition* 33, 2 (2000): 225–236.

13 http://scantailor.org/ (09.05.2016).

14 https://www.abbyy.com/de-de/finereader/ (25.05.2016).

15 https://github.com/tesseract-ocr (09.05.2016).

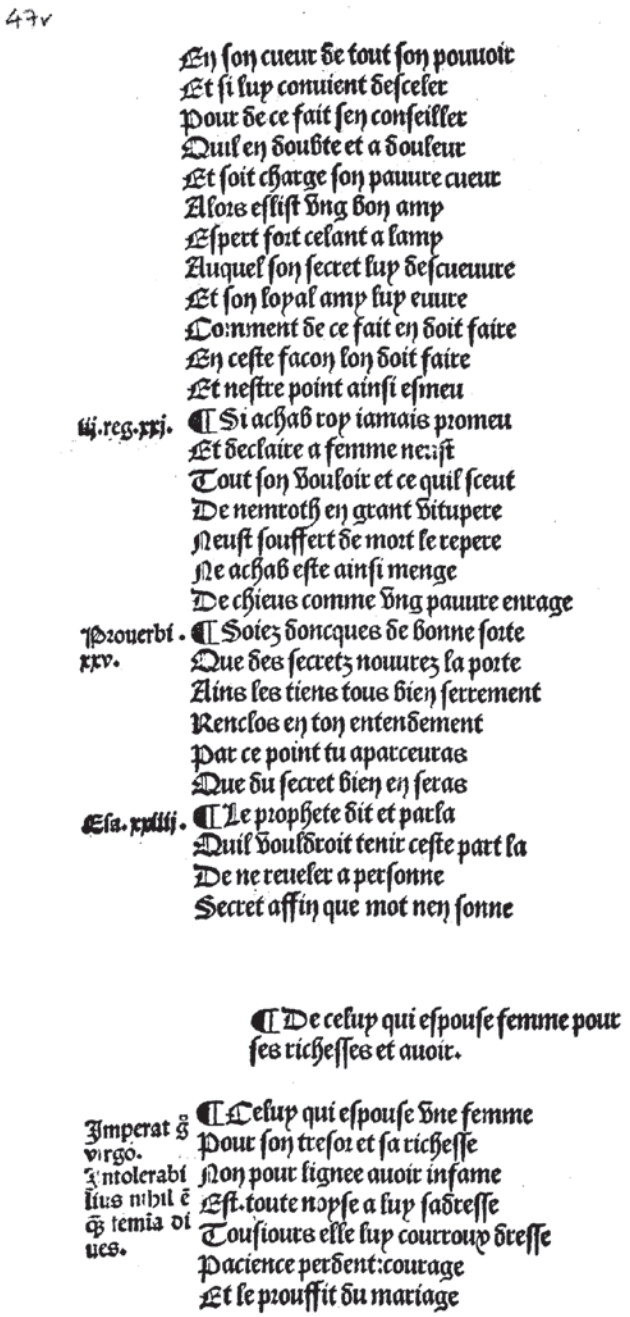

Abb. 3: Vorverarbeitetes Bild zu Abb. 2.

4 Erstellung von offizin- und schriftartspezifischen Typentabellen

Um möglichst alle verwendeten Zeichen einer Offizin nach und nach erfassen und trainieren zu können, ist es von Vorteil, offizin- und schriftartspezifische Typentabellen zu erstellen. So gewinnt man im Laufe der Zeit das vollständige Glypheninventar einer Druckerei. Als Hilfestellung kann das *Typenrepertorium der Wiegendrucke von der Staatsbibliothek zu Berlin*[16] dienen. Bei den im Teilprojekt „Narragonien“ untersuchten Expemplaren von Sebastian Brants „Narrschiff“ erwies sich dieses jedoch als unvollständig, so dass die Glypheninventare neu erstellt werden mussten.

Inkunabeln, insbesondere lateinische, enthalten in der Regel eine Vielzahl an Sonderzeichen, die in modernen Standard-Schriftarten nicht enthalten sind und somit auf Computern zunächst nicht angezeigt werden können.

Abb. 4: Beispiele für Sonderzeichen.

Zur Lösung des Anzeigeproblems bieten sich zwei Möglichkeiten an:

a) Substitution der Sonderzeichen: Das heißt, alle nicht im Standardzeichensatz enthaltenen Sonderzeichen werden durch ein Metazeichen ersetzt:

   #q für q̃

b) Verwendung einer speziellen Schriftart, die zum Beispiel nach den Vorgaben der *Medieval Unicode Font Initiative*[17] designt wurde und auf einem erweiterten Unicode-Zeichensatz basiert.

Die Entscheidung für eine der beiden Alternativen hängt stark vom jeweiligen Einsatzszenario ab.

An der Universitätsbibliothek Würzburg wurde die Entscheidung zu Gunsten des *MUFI*-Ansatzes getroffen. Ziel war eine möglichst exakte Darstellung des Originaltextes sowohl auf Webseiten als auch in den Texteditoren der Anwender. Dies führt dazu, dass zu den OCR-Ergebnissen auch immer die benötigte Schriftart mit geliefert werden muss, damit die Anzeige der Sonderzeichen korrekt erfolgt.

Damit beim Trainieren der OCR den jeweiligen Glyphen der richtige Unicode aus dem *MUFI*-Zeichensatz zugeordnet wird, werden die Typentabellen um eine Spalte erweitert. Diese Spalte weist dann jedem Glyphen exakt eine Unicode Entsprechung zu (vgl. Abb. 5). Die Zuordnung sollte hierbei von paläographisch geschulten Personen vorgenommen werden. Um die Erstellung der Ty-

16 http://tw.staatsbibliothek-berlin.de/ (09.05.2016).
17 http://folk.uib.no/hnooh/mufi/ (09.05.2016).

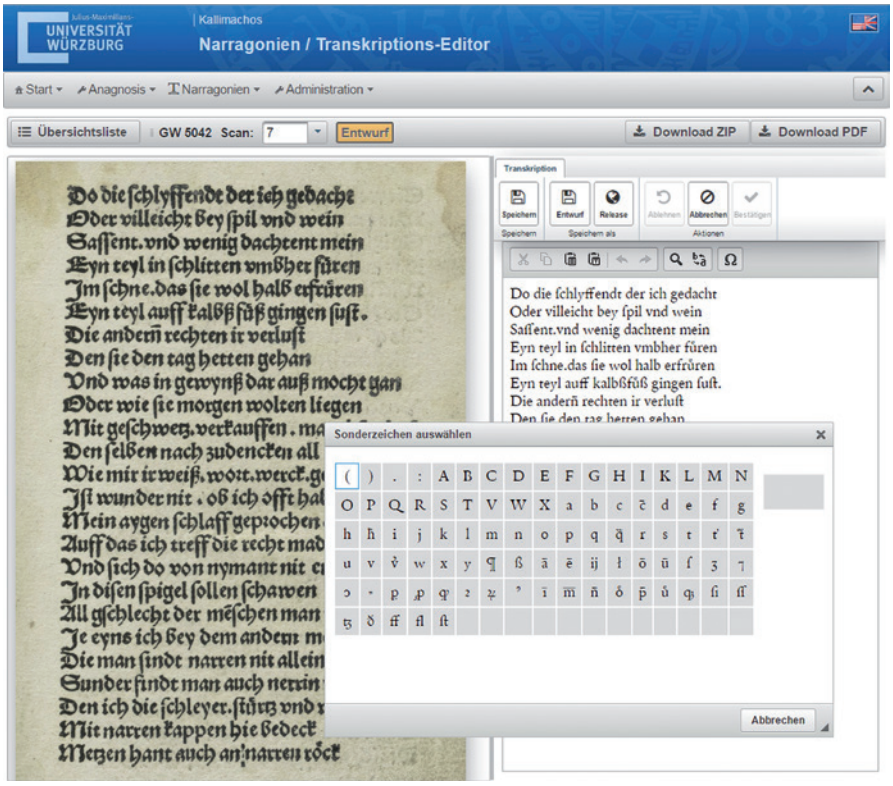

Abb. 5: Transkriptions-Editor der Universitätsbibliothek Würzbug mit werkspezifischem Sonderzeichen-Zeichensatz.

peninventare möglichst effizient zu gestalten, wird an der Universitätsbibliothek Würzburg das Programm *Glyph Miner*[18] entwickelt. Das System erlaubt es, eine große Anzahl an Glyphen aus Scans von alten Drucken zu extrahieren und diese samt zugehörigen Unicodes in eine Glyphenbibliothek zu überführen. Ein Prototyp wurde im Rahmen einer Benutzerstudie während des Workshops *philtag 2016*[19] an der Universitätsbibliothek Würzburg evaluiert. Dabei wurde das Programm von den teilnehmenden Geisteswissenschaftlern und Digitalisierungsexperten ausführlich getestet und hinsichtlich Arbeitsaufwand und Qualität der Resultate positiv bewertet.

MUFI Zeichen	MUFI Unicode
n̄	E5DC
o	006F
ō	014D
œ	0153
o̊	E644
p	0070
ꝓ	A753
p̄	E665
ꝑ	A751
q	0071
q̄	0071+0304

Abb. 6: Ausschnitt aus einer Typentabelle.

18 https://github.com/benedikt-budig/glyph-miner (02.06.2016); Budig, Benedikt, Thomas C. Van Dijk, Felix Kirchner: „Glyph Miner: A System for Efficiently Extracting Glyphs from Early Prints in the Context of OCR." In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '16) 2016:31–34.

19 http://kallimachos.de/kallimachos/index.php/Philtag (02.06.2016).

5 Schriftartspezifische Segmentierung

Inkunabeln besitzen oft ein komplexes Layout, was eine sorgfältige Segmentierung erforderlich macht. Die einzelnen Segmenttypen, wie zum Beispiel Überschriften, Marginalien und Fließtexte, sind häufig in unterschiedlichen Schriftarten gedruckt. Deshalb bietet sich hier eine nach Typ bzw. Schriftart getrennte Segmentierung an, wodurch beim Erkennungsprozess mit *Tesseract* deutlich bessere Ergebnisse erzielt werden können.

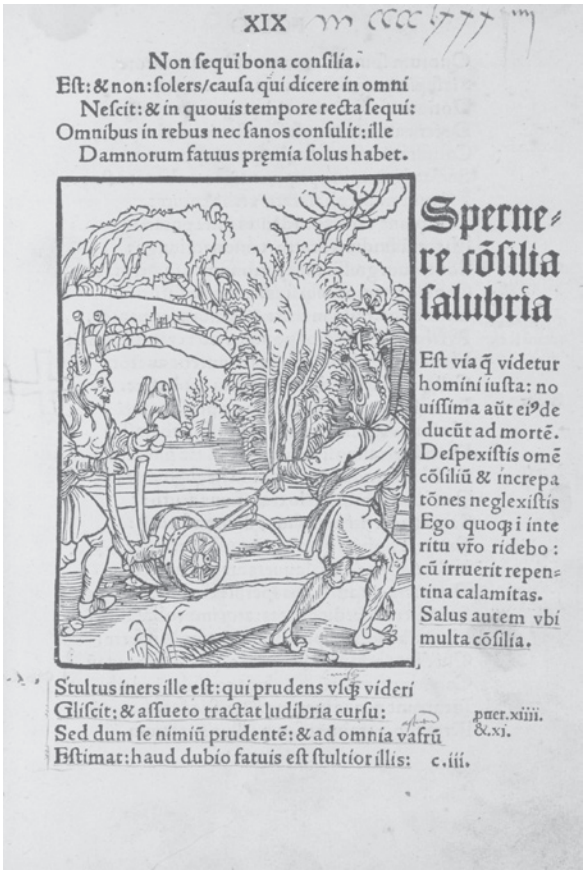

Abb. 7: Beispiel unterschiedlicher Schriftarten auf einer Seite[20].

Abbildung 7 zeigt eine ausgewählte Seite des Inkunabel GW 5061, die drei verschiedene Schriftarten enthält: Überschrift, Haupttext, Marginalie. Die Segmentierung kann manuell oder halbautomatisch durchgeführt werden. An der UB Würzburg wurde das Programm *Aletheia*[21] von *PrimaResearch*[22] verwendet. Die nachfolgende Beschreibung des Segmentierungsvorgangs kann auch auf andere Anwendungen übertragen werden.

Zunächst werden die einzelnen Bildbereiche durch das Aufziehen von Polygonen markiert. Anschließend erfolgt eine Typisierung der Bereiche. Hierzu werden den Bildausschnitten Bezeichner zugewiesen. Der Bezeichner bestimmt im nachfolgenden OCR-Prozess die für diesen Abschnitt verwendete *Tesseract*-Trainingsdatei. Aus diesem Grund ist es sinnvoll, die Bezeichner der Bildregionen und die Trainingsdateien entsprechend der verwendeten Schriftart zu benennen. Abschließend erfolgt die Festlegung der Lesereihenfolge. Um eine maschinelle Weiterverarbeitung durchführen zu können, müssen die in diesem Arbeitsschritt gewonnenen Informationen in einem XML gespeichert werden. Hierfür hat sich der *PAGE-XML-Standard*[23] etabliert. Die PAGE-XML-Datei wird genauso wie die dazugehörige Bilddatei benannt.

Abb. 8: Beispiel einer Segmentierung[24].

20 Brant, Sebastian: *Das Narrenschiff*. Lat. von Jacobus Locher Philomusus. Mit Beig. von Thomas Beccadelli. Basel: Johann Bergmann von Olpe, 01.08.1497. 19r. (Universitätsbibliothek Würzburg, Inc.q.32 – GW 5061).

21 http://www.primaresearch.org/tools/Aletheia (09.05.2016).

22 http://www.primaresearch.org/ (09.05.2016).

23 http://www.primaresearch.org/schema/PAGE/gts/pagecontent/2013-07-15/ (09.05.2016).

24 Brant, Sebastian: *Das Narrenschiff*. Lat. von Jacobus Locher Philomusus. Mit Beig. von Thomas Beccadelli. Basel: Johann Bergmann von Olpe, 01.08.1497. 29r. (Universitätsbibliothek Würzburg, Inc.q.32 – GW 5061).

Die so entstandenen Bild-/XML-Pärchen werden anschließend mit einem eigens an der Universitätsbibliothek Würzburg entwickelten Programm weiter verarbeitet. Hierbei werden die Bilder an Hand der im XML gespeicherten Informationen in Einzeldateien zerlegt und entsprechend ihres Labels und ihrer Position innerhalb der Lesereihenfolge benannt.

6 Schriftartspezifisches Trainieren von *Tesseract*

Da das Training von *Tesseract* über die Kommandozeile sehr anspruchsvoll ist, wurde im Rahmen des *EMOP*[25] Projektes die Anwendung *Franken+*[26] entwickelt. Diese bietet eine grafische Benutzeroberfläche zum Trainieren von *Tesseract*.

Franken+ benötigt im Wesentlichen zwei Komponenten zur Erstellung von Trainingsdatensätzen. Zum einen bedarf es mehrerer Repräsentanten und Varianten für jedes Zeichen in den Glyphentabellen. Die Glyphen können manuell oder automatisch gewonnen werden. Für die automatische Gewinnung bietet sich abermals *Aletheia* an, wie dies auch schon im *EMOP*-Projekt erprobt wurde. Zum anderen wird ein Trainingstext benötigt, der idealerweise alle im Werk vorkommenden Zeichen enthält.[27]

Die Erstellung der Trainingsdatensätze erfolgt schriftartspezifisch, ähnlich der Segmentierung. Das bedeutet, dass für jede Schriftart genau eine Trainingsdatei erstellt wird. Dies bietet den Vorteil, dass die Anzahl der gespeicherten Bildmuster pro Zeichen gering gehalten wird, was zu einer deutlich höheren Trennschärfe bei der Erkennung der Zeichen führt und somit die OCR Qualität verbessert.

Tab. 1: Vergleich der Erkennungsgenauigkeit auf 100 Seiten aus GW 5061[28].

	Separate Trainingsdateien	Kombinierte Trainingsdatei
Zeichengenauigkeit	90,63 %	93,21 %

Da das Druckbild bei Inkunabeln auf Grund von Beschädigungen und Abnutzungen der Druckstempel variiert, ist es notwendig, für jeden Buchstaben die Repräsentanten sorgfältig zu wählen. Stark beschädigte oder zerfallene Buchstaben sollten für das Training ausgespart werden. Andererseits sollten häufig vorkommende Varianten ein und desselben Zeichens in das Training aufgenommen werden.

Für Inkunabeln existieren in den meisten Fällen keine glyphengenauen Transkriptionen, weshalb an der Universitätsbibliothek Würzburg der Ansatz verfolgt wurde, ein *Tesseract*-Training basierend auf bedeutunglosen Zeichenfolgen durchzuführen. In diesen künstlichen Trainingstexten sind alle erfassten Zeichen einer Schrifttype in ausreichendem Maße enthalten. Das heißt, dass auch selten vorkommende Zeichen trainiert werden. Mit dieser Trainingsmethode konnten auf Inkunabeln der Offizin Bergmann von Olpe bereits gute OCR-Ergebnisse erzielt werden. Die OCR-Qualität konnte anschließend mit der Hinzunahme einiger weniger Seiten transkribierten Originaltextes weiter gesteigert werden. In Tests hat sich gezeigt, dass circa 3 Seiten, abhängig von deren Umfang, als Zusatz zum bedeutunglosen Trainingstext ausreichend sind. Zu beachten ist, dass beim Abschreiben der Texte die entsprechenden Sonderzeichen aus dem *MUFI*-Zeichensatz verwendet werden.

In der Praxis hat herausgestellt, dass die Kombination von künstlichen, bedeutungslosen Trainingstexten zusammen mit ausgewählten transkribierten Textseiten beim Trainieren von Tesseract die besten OCR Ergebnisse liefert.

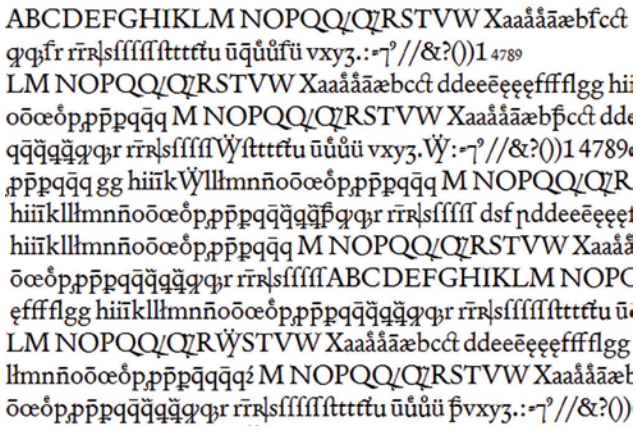

Abb. 9: Ausschnitt der bedeutungslosen Zeichenfolge für das Training von Tesseract.

7 Schriftartspezifische OCR

Der eigentliche OCR-Prozess erfolgt nun auf den einzelnen Segmentdateien. Mit einem an der Universitätsbibliothek Würzburg entwickelten Steuerskript wird der im Dateina-

25 http://emop.tamu.edu/ (09.05.2016).
26 http://emop.tamu.edu/outcomes/Franken-Plus (09.05.2016).
27 Eine ausführliche Anleitung zum Trainieren von *Tesseract* mit *Franken+* in Kombination mit *Aletheia* befindet sich auf der Webseite des *EMOP*-Projekts: http://emop.tamu.edu/Aletheia-Franken%2BQuickStart (25.05.2016).
28 https://github.com/benedikt-budig/glyph-miner (02.06.2016).

men kodierte Segmenttyp ausgewertet und die entsprechende Trainingsdatei beim automatisierten Aufruf von *Tesseract* angegeben. Somit wird jedes Segment entsprechend seiner Schriftart erkannt und als Textdatei neben der Bilddatei gespeichert.

Bei der Texterkennung durch *Tesseract* bereitet der automatische *Page-Segmentation-Mode* vor allem bei einzeiligen Segmenten häufig Probleme. Die OCR-Ausgabe von Tesseract ist in diesen Fällen meist leer. Dies gilt insbesondere für Segmente, die nur aus einem einzelnen Wort bestehen. Diese Probleme werden durch das Steuerskript gesondert behandelt. Das Skript wurde dahingehend erweitert, dass die Ausgabe von *Tesseract* auf ihren Inhalt hin überprüft wird. Sollte der Inhalt einer Ausgabedatei leer sein, so erfolgt ein erneuter Aufruf von *Tesseract* mit dem zusätzlichen Parameter *-psm 7*. Dieser Parameter erzwingt einen speziellen Segmentierungsmodus für einzeilige Eingaben. Sollte das Ergebnis dieses OCR-Vorgangs ebenfalls leer sein, so wird Tesseract erneut mit dem Parameter *-psm 8* aufgerufen. Dieser Modus ist speziell für Eingaben bestimmt, die nur aus einem einzigen Wort bestehen.

Das hier beschriebene Vorgehen erwies sich insbesondere bei Marginalien, die in Inkunabeln häufig auftreten, als sehr hilfreich (siehe Abb. 8).

8 Nachverarbeitung: Wörterbücher

In der Nachverarbeitung können die per OCR gewonnenen Texte halbautomatisch aufbereitet werden. Hierbei sind zeitlich und sprachlich passende Wörterbücher von starkem Nutzen. Allerdings existieren bislang kaum elektronische Wörterbücher für die Zeit vor 1500. Ein Grund ist die starke dialektale Prägung vor allem im Deutschen. Ein weiterer Grund ist die häufig kontextbezogene Verwendung von Abkürzungen, insbesondere in lateinischen Texten. Im Vordergrund stand zur damaligen Zeit das optimierte Druckbild. Rechtschreibung und Verwendung der Abkürzungen wurden dem Layout untergeordnet.[29] Die Abbildungen 10 und 11 zeigen zwei unterschiedliche Schreibweisen des Wortes *monumenta*, einmal mit und einmal ohne Abkürzung. Da die entsprechende Zeile in Abbildung 10 bereits „voll“ war, wurden hier verstärkt Abkürzungen verwendet, um das Gesamtlayout zu wahren.

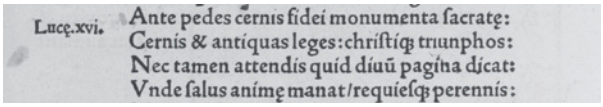

Abb. 10: Schreibweise von „monumenta“ ohne Abkürzung in GW 5061[30].

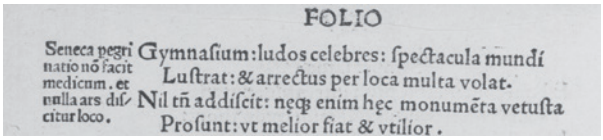

Abb. 11: Schreibweise von „monumenta“ mit Abkürzung in GW 5061[31].

Ein weiteres Beispiel zur Verwendung von Abkürzungen mit dem Ziel der Einhaltung des Layouts wird in Abbildung 12 gezeigt. Die herausgestellte und vergrößerte Zeile enthält beispielsweise ‚i‘ als Abkürzung für ‚ihn‘. Das sehr kompakte Schriftbild erlaubt an dieser Stelle keine weiteren ausgeschriebenen Worte.

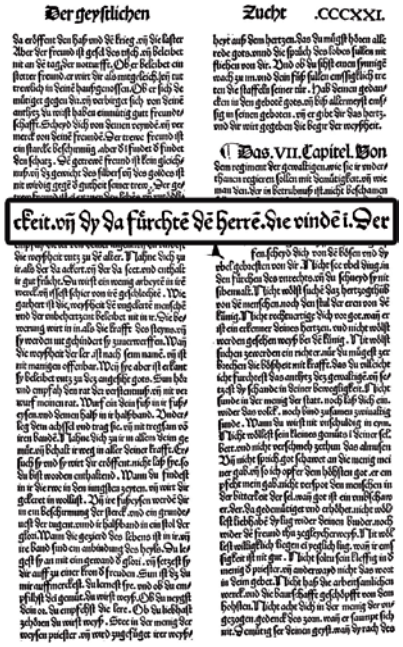

Abb. 12: Verwendung von Abkürzungen in der Koberger-Bibel[32].

29 Geldner, Ferdinand: *Inkunabelkunde: eine Einführung in die Welt des frühesten Buchdrucks*. Wiesbaden 1978. 2.

30 Brant, Sebastian: *Das Narrenschiff*. Lat. von Jacobus Locher Philomusus. Mit Beig. von Thomas Beccadelli. Basel: Johann Bergmann von Olpe, 01.08.1497. 22v. (Universitätsbibliothek Würzburg, Inc.q.32 – GW 5061).

31 Brant, Sebastian von Olpe: *Das Narrenschiff*. Lat. von Jacobus Locher Philomusus. Mit Beig. von Thomas Beccadelli. Basel: Johann Bergmann von Olpe, 01.08.1497. 45v. (Universitätsbibliothek Würzburg, Inc.q.32 – GW 5061).

32 Biblia. Nürnberg: Anton Koberger, 17.02.1483. S. 321r. (Universitätsbibliothek Würzburg, I.t.f.326-2 – GW 4303)

9 Zusammensetzen der erkannten Textsegmente

Das Zusammensetzen der durch die OCR gewonnen Textsegmente erfolgt an Hand der in den jeweiligen PAGE-XML-Dateien gespeicherten Informationen. Hierbei spielt die Einhaltung der Lesereihenfolge eine wesentliche Rolle. Um die abschließende manuelle Korrektur der OCR-Ergebnisse zu vereinfachen, werden die Textsegmente in Lesereihenfolge untereinander zusammengefügt. Für Präsentationszwecke kann die Darstellung der Textsegmente allerdings auch originalgetreu wiedergegeben werden.

10 Validierung von OCR-Ergebnissen

Die Validierung der OCR-Ergebnisse erfordert zunächst die Erstellung einer glyphengenauen Transliteration von mindestens 10 Seiten eines Werkes, um repräsentative und verwertbare Ergebnisse zu erhalten. Im Idealfall werden die Seiten zufällig von einem Skript aus der Gesamtmenge aller Seiten ausgewählt.

Der so gewonnene *Ground Truth* muss mit der Unicode-Codierung der OCR-Ergebnisse übereinstimmen. Für die Validierung wird an der Universitätsbibliothek Würzburg das Programm *Text Evaluation*[33] von *PrimaResearch* verwendet. Mit ihm lassen sich Zeichengenauigkeit, Wortgenauigkeit und andere Ergebnisbewertungen bestimmen.

Mit dem in diesem Artikel vorgestellten Verfahren konnten auf Inkunabeln aus dem Bestand der Universitätsbibliothek Würzburg aus der Offizin Bergmann von Olpe (Basel) nachfolgende Ergebnisse erzielt werden.

Tab. 2: OCR Ergebnisse für Offizin Bergmann von Olpe.

Signatur	Zeichen-genauigkeit	Wortgenauigkeit	Validierungs-seitenzahl
Inc.q.32[34]	92,99 %	67,19 %	98
Inc.q.12[35]	94,03 %	65,07 %	30
Inc.q.12 angeb.1[36]	95,66 %	73,60 %	16, komplett
Inc.q.12 angeb.2[37]	93,54 %	63,07 %	30
Inc.q.12 angeb.3[38]	91,78 %	55,43 %	30
Inc.q.12 angeb.4[39]	94,19 %	64,37 %	28, komplett
Inc.q.27 angeb.1[40]	95,27 %	67,99 %	30

Die hierbei verwendeten Tesseract-Trainingsdatensätze wurden ausschließlich auf Grundlage des Werkes GW 5061 erstellt und anschließend auf die anderen Werke, die von derselben Offizin mit identischer Type gedruckt wurden, angewendet. Die Ergebnisse wurden mit jeweils 50 Textseiten validiert. Damit konnte der Beweis angetreten werden, dass einmal sorgfältig erstellte Trainingsdaten offizinspezifisch „recycled“ und ohne weiteres Training damit Zeichen-Erkennungsgenauigkeiten von knapp 92 Prozent bis zu über 95 Prozent erzielt werden können (ohne Nachverarbeitung über Wörterbücher oder dgl.).

Eine hohe Zeichengenauigkeit muss nicht zwangsweise eine hohe Wortgenauigkeit bedingen. Die Wortgenauig-

33 http://www.primaresearch.org/tools/PerformanceEvaluation (10.05.2016).

34 Brant, Sebastian: *Das Narrenschiff*. Lat. von Jacobus Locher Philomusus. Mit Beig. von Thomas Beccadelli. Basel: Johann Bergmann von Olpe, 01.08.1497. (Universitätsbibliothek Würzburg, Inc.q.32 – GW 5061).

35 Brant, Sebastian: *Carmina in laudem virginis Mariae multorumque sanctorum*. [Basel]: [Johann Bergmann von Olpe], [nicht vor 1494]. (Universitätsbibliothek Würzburg, Inc.q.12 – GW 5067).

36 Wimpfeling, Jakob: *De nuntio angelico carmen*. Basel: [Johann Bergmann von Olpe], 1494. (Universitätsbibliothek Würzburg, Inc.q.12 angeb.1 – GW M51660).

37 Wimpfeling, Jakob: *De conceptu et triplici Mariae virginis candore carmen*. Basel: [Johann Bergmann von Olpe], 1494. (Universitätsbibliothek Würzburg, Inc.q.12 angeb.2 – GW M51598).

38 Brant, Sebastian: *De origine et conversatione bonorum regum et de laude civitatis Hierosolymae*. Basel: Johann Bergmann von Olpe, 01.03.1495. (Universitätsbibliothek Würzburg, Inc.q.12 angeb.3 – GW 5072).

39 Dal Maino, Giasone: *Epithalamium in nuptiis Maximiliani regis et Blancae Mariae*. Basel: [Johann Bergmann von Olpe], [nicht vor 1494]. (Universitätsbibliothek Würzburg, Inc.q.12 angeb.4 – GW M22387).

40 Brant, Sebastian: *Carmina in laudem virginis Mariae multorumque sanctorum*. [Basel]: [Johann Bergmann von Olpe], [nicht vor 1494]. (Universitätsbibliothek Würzburg, Inc.q.27 angeb.1 – GW 5067).

keit ist in der Regel höher zu werten, da ein Text bei der Nachkorrektur leichter zu korrigieren ist, je mehr Wörter als Ganzes bereits korrekt erkannt wurden. So kann es sein, dass der Text eines OCR-Verfahrens mit hoher Zeichengenauigkeit schwerer zu korrigieren ist als ein Text mit geringerer Zeichengenauigkeit, bei dem die meisten Worte richtig erkannt wurden und die meisten Fehler zum Beispiel in die Wortzwischenräume fallen.

Im Gegensatz zu der Aussage von Jeffrey A. Rydberg-Cox: „Because of the prevalence of these glyphs, incunabula cannot be processed using OCR software. Commercial OCR programs produce almost no recognizable character strings, let alone searchable text. ...“[41], zeigen die hier vorgestellten Ergebnisse, dass OCR auf Inkunabeln durchaus verwertbare Ergebnisse liefern kann und die in den DFG Praxisregeln verlangte Mindesterkennungsgenauigkeit von 90 Prozent[42] oder besser erreicht werden kann.

11 Fazit

Bereits mit heute zur Verfügung stehenden OCR-Verfahren und -Programmen sind Inkunabeln kein „hoffnungsloser Fall“ mehr. Grundvoraussetzung ist jedoch eine sorgfältige Arbeitsweise. Hochwertige verzerrungsfreie Aufnahmen, die Entfernung von Flecken, das Ausrichten und Geradestellen von Textblöcken und Zeilen sowie die Wahl eines für das jeweilige Werk optimalen Binarisierungsverfahrens bilden das Fundament für gute OCR-Ergebnisse. Ein weiterer elementarer Baustein für ein verwertbares OCR-Ergebnis bei Inkunabeln ist die vollständige Erfassung aller vorkommenden Schriftarten und deren Glyphen. Auch selten auftretende Zeichen sollten erfasst werden, um die Wiederverwendbarkeit bei Werken der gleichen Offizin sicherzustellen. Ein sorgfältiges schriftartspezifisches Training der Glyphen ist Bedingung für bestmögliche Erkennungsgenauigkeiten. Hierbei hat sich die Komabination aus künstlichen bedeutunglosen Texten und zeichengenauer Transkription bewährt. Diese ermöglicht erst das hinreichende Training seltener Zeichen.

Nachlässigkeiten bei den oben genannten Arbeitsschritten summieren sich über den Trainingsprozess hinweg auf und können auch durch Sorgfalt in den darauffolgenden Arbeitsschritten nicht mehr ausgeglichen werden. Sie führen am Ende zu Erkennungsraten von deutlich unter 90 Prozent, was für eine Nachnutzung der Volltexte nicht mehr ausreichend ist. Eine echte Erleichterung bei der Nachnutzung der Volltexte für Editoren ist erst ab einer Erkennungsgenauigkeit von 93–95 Prozent zu erwarten. Dies wird auch in den DFG Praxisregeln[43] gefordert. Fortschritte bei der OCR-Qualität sind durch weitere Nachverarbeitungsschritte zu erwarten. Zum Beispiel kann die Verwendung dialektaler und zeitlich eingeordnerter Wörterbücher eine halbautomatische Korrektur der OCR Texte vereinfachen und beschleunigen. Diese und weitere Möglichkeiten werden im Rahmen des KALLIMACHOS Projekts weiter erforscht. Bei ausreichender Trainingsdaten- und Gutmustermenge ist insbesondere mit LSTM-basierten OCR-Programmen[44] eine weitere Steigerung der Erkennungsraten möglich. Ein Merkmal der Workflowkette der Universitätsbibliothek Würzburg ist, dass sie modular aufgebaut ist. So werden die Ausgaben der Einzelschritte nachvollziehbar und mögliche Fehler können bereits vor dem nächsten Arbeitsschritt erkannt und korrigiert werden. Zudem sind einzelne Abschnitte beziehungsweise Programme der Kette austauschbar, wodurch mögliche Kooperationen mit anderen Institutionen möglich werden. Der Austausch von Wissen und Technik im Sinne eines Open-Source-Gedankens ist erstrebenswert und soll zu einer klaren Win-Win-Situation für alle führen.

41 Rydberg-Cox, Jeffrey A.: „Digitizing Latin Incunabula: Challenges, Methods and Possibilities.“ In *Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure* 3, 1 (2009) (siehe http://www.digitalhumanities.org/dhq/vol/3/1/000027/000027.html#p7, 30.05.2016).

42 DFG Praxisregeln „Digitalisierung“: 12.151 – 02/13, 30 ff., http://www.dfg.de/formulare/12_151/12_151_de.pdf (30.05.2016).

43 DFG Praxisregeln „Digitalisierung“: 12.151 – 02/13, http://www.dfg.de/formulare/12_151/12_151_de.pdf (30.05.2016).

44 OCRopus: Open Source Document Analysis and OCR System. (https://github.com/tmbdev/ocropy, 31.05.2016); OCROCIS: Ludwig-Maximilians-Universität Münschen, Centrum für Informations- und Sprachverarbeitung (http://cistern.cis.lmu.de/ocrocis/, 30.05.2016); anyOCR: Deutsches Forschungszentrum für Künstliche Intelligenz. (https://www.dfki.de/web/presse/pressemitteilungen_intern/2015/anyocr-2013-intelligente-texterkennung-steuert-das-201enarrenschiff201c-ins-digitale-zeitalter/, 03.06.2016); Ul-Hasan, Adnan; Syed Saqib Bukhari, Andreas Dengel: „OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters.“ 2016. (https://www.researchgate.net/publication/294575734_OCRoRACT_A_Sequence_Learning_OCR_System_Trained_on_Isolated_Characters, 03.06.2016).

Autoreninformationen

Felix Kirchner
Universitätsbibliothek Würzburg
Am Hubland
97074 Würzburg
felix.kirchner@bibliothek.uni-wuerzburg.de
orcid.org/0000-0002-2653-6554

Marco Dittrich
Universitätsbibliothek Würzburg
Am Hubland
97074 Würzburg
marco.dittrich@bibliothek.uni-wuerzburg.de
orcid.org/0000-0002-8681-6443

Phillip Beckenbauer
Universitätsbibliothek Würzburg
Am Hubland
97074 Würzburg
phillip.beckenbauer@stud-mail.uni-wuerzburg.de
orcid.org/0000-0003-4379-1669

Maximilian Nöth
Universitätsbibliothek Würzburg
Am Hubland
97074 Würzburg
maximilian.noeth@stud-mail.uni-wuerzburg.de
orcid.org/0000-0002-5048-7238