Alexander Gehrke*, Nico Balbach, Yong-Mi Rauch, Andreas Degkwitz und Frank Puppe

Erkennung von handschriftlichen Unterstreichungen in Alten Drucken

https://doi.org/10.1515/bfp-2019-2083

Zusammenfassung: Die Erkennung handschriftlicher Artefakte wie Unterstreichungen in Buchdrucken ermöglicht Rückschlüsse auf das Rezeptionsverhalten und die Provenienzgeschichte und wird auch für eine OCR benötigt. Dabei soll zwischen handschriftlichen Unterstreichungen und waagerechten Linien im Druck (z. B. Trennlinien usw.) unterschieden werden, da letztere nicht ausgezeichnet werden sollen. Im Beitrag wird ein Ansatz basierend auf einem auf Unterstreichungen trainierten Neuronalen Netz gemäß der U-Net Architektur vorgestellt, dessen Ergebnisse in einem zweiten Schritt mit heuristischen Regeln nachbearbeitet werden. Die Evaluationen zeigen, dass Unterstreichungen sehr gut erkannt werden, wenn bei der Binarisierung der Scans nicht zu viele Pixel der Unterstreichung wegen geringem Kontrast verloren gehen. Zukünftig sollen die Worte oberhalb der Unterstreichung mit OCR transkribiert werden und auch andere Artefakte wie handschriftliche Notizen in alten Drucken erkannt werden.

Schlüsselwörter: Brüder Grimm Privatbibliothek; Erkennung handschriftlicher Artefakte; Convolutional Neural Network; regelbasierte Nachbearbeitung

Recognition of Handwritten Underlines in Historical Printings

Abstract: The recognition of handwritten artefacts like underlines in historical printings allows inference on the reception and provenance history and is necessary for OCR (optical character recognition). In this context it is important to differentiate between handwritten and printed lines, since the latter are common in printings, but should be ignored. We present an approach based on neural nets with the U-Net architecture, whose segmentation results are post processed with heuristic rules. The evaluations show that handwritten underlines are very well recognized if the binarisation of the scans is adequate. Future work includes transcription of the underlined words with OCR and recognition of other artefacts like handwritten notes in historical printings.

Keywords: Grimm brothers personal library; handwritten artefact recognition; convolutional neural network; rule based post processing

***Kontaktperson: Alexander Gehrke,**
Alexander.Gehrke@uni-wuerzburg.de
Nico Balbach, Nico.Balbach@uni-wuerzburg.de
Yong-Mi Rauch, Yong-Mi.Rauch@ub.hu-berlin.de
Andreas Degkwitz, andreas.degkwitz@ub.hu-berlin.de
Frank Puppe, Frank.Puppe@uni-wuerzburg.de

1 Einleitung

Die Privat- und Arbeitsbibliothek von Jacob und Wilhelm Grimm stellte ein zentrales Arbeitsinstrument der Philologen dar. In den Büchern sind zahlreiche handschriftliche Eintragungen, Notizen und Unterstreichungen zu finden. Dies stellt eine typische Arbeitsweise in der historischen Wissenschaft dar. Die Werke mitsamt Bearbeitungsspuren erlauben Rückschlüsse auf die Arbeitsvorhaben und -weise der Brüder Grimm, wobei besonders die *Kinder- und Hausmärchen (KHM)* oder das *Deutsche Wörterbuch* zu nennen sind. Buchspiegel, Vorsatzblätter und Vacat-Seiten sind häufig dicht mit Exzerpten beschrieben. Die Bibliothek enthält zudem Merkmale und Notizen, die über die Geschichte der Bücher Auskunft geben (z. B. Stempel und Kaufvermerke), und Geschenk- und Widmungsexemplare, welche das wissenschaftliche Netzwerk der Brüder Grimm belegen. Durch diese individuellen Merkmale werden viele Grimm-Bücher eindeutig zu Unikaten. Da nur wenige bürgerliche Privatbibliotheken vor 1900, zumal von dieser Bedeutung, heute geschlossen überliefert sind, ist diese universelle, fachkundig aufgebaute Büchersammlung eine bedeutende Quelle für wissenschafts- und bildungsgeschichtliche Fragen. Die Bibliothek ist zwar ausführlich dokumentiert, aber auch im Bereich der Grimm-Forschung nur unvollständig berücksichtigt.[1]

Erhalten sind insgesamt mehrere tausend Bände, die heute in verschiedenen Institutionen verwahrt werden, der bei weitem größte und heute geschlossen aufgestellte Bestand in der Universitätsbibliothek der Humboldt-Universität zu Berlin mit knapp 6 000 Bänden. Die manuelle Auswertung dieser Sammlung ist zu aufwändig; dies gilt auch

1 Denecke und Teitge (1989), Friemel (2005), Rauch (2018).

für kleinere thematische Segmente der Bibliothek wie der Handbibliothek zu den KHM, quasi dem Quellen- und Arbeitsapparat der beiden Philologen zu diesem zentralen Werk. Dieser umfasst etwa ein Zehntel des Gesamtbestands und ist Gegenstand eines Referenzprojekts zur Digitalisierung und Erschließung der Bibliothek, in dessen Rahmen dieser Beitrag entstand.

Die Auswertung der Annotationen als wissenschaftliche Quelle wird durch die ungleichmäßige Verteilung erschwert: Während Teile der Bände recht dicht bearbeitet sind, sind andere Bereiche oder ganze Bände kaum benutzt, oder die Spuren sind erst während der Jahrzehnte nach 1865 entstanden. Denn die Bücher wurden in der Berliner Universitätsbibliothek bis weit ins 20. Jahrhundert in der regulären Ausleihe genutzt. Die Suche nach Arbeitsspuren und deren Bewertung sowohl im materiellen Band als auch im Digitalisat ist deshalb mühsam und zeitaufwendig. Insofern ist eine automatische Erkennung, Auszeichnung der Provenienzspuren im Digitalisat und eine Navigation in diesen Elementen ein Desiderat.

In diesem Beitrag werden erste Ergebnisse für eine wichtige Teilaufgabe der Annotation der Arbeitsspuren vorgestellt, nämlich ein Ansatz zur automatischen Erkennung von handschriftlichen Unterstreichungen als Voraussetzung für die Erkennung der unterstrichenen Wörter im Drucktext. Die Erkennung erfolgt durch Segmentierung der Scans der Druckseiten mit Neuronalen Netzen sowie einer Nachbearbeitung der gefundenen Linien-Fragmente mit heuristischen Regeln. Da für die Neuronalen Netze weit mehr Trainingsmaterial benötigt wird als verfügbar war, wurde das Trainingsmaterial automatisch generiert und die verfügbaren Dokumente mit manuell annotierten Unterstreichungen nur zur Evaluation verwendet.

Im nächsten Abschnitt wird über verwandte Arbeiten berichtet. Danach werden im dritten und vierten Abschnitt die Methoden und die verwendeten Daten vorgestellt. Abschnitt fünf und sechs enthalten die Ergebnisse und eine Fehleranalyse. Im letzten Abschnitt werden die Ergebnisse zusammengefasst und weitere Ansätze zur Verbesserung vorgestellt.

2 Related Work

Das Erkennen und Entfernen und Unterstreichungen in Texten hat eine lange Tradition, da es Voraussetzung für die OCR der Texte ist.[2] Allerdings ist das Ziel dabei nicht, die unterstrichenen Wörter zu finden, sondern die Unterstreichungen aus dem Scan zu entfernen, da die anschließende OCR unterstrichene Wörter häufig falsch transkribiert. Das betrifft im Prinzip alle Linien, nicht nur die manuell hinzugefügten. Wichtige Ansätze sind:[3]

- Kantenfilter (z. B. Sobel-Operator, Canny-Edge-Algorithmus, Aktive Konturen, Laplace-Filter usw.), ggf. mit anschließender Hough-Transformation
- Zweidimensionale Gabor Filter: Gabor Filter sind in der Bildverarbeitung weit verbreitet und eignen sich zur Erkennung von Ecken bzw. Kanten in konfigurierbaren Orientierungen.
- Analyse von Connected Components, wobei für die Linienerkennung längliche Komponenten gesucht werden.
- Convolutional Neural Networks (CNN)

Im Ansatz von Pratihar et al. (2012) werden sehr gute Ergebnisse für die Erkennung und Entfernung von handgeschriebenen Unterstrichen in Texten mit Connected Components berichtet (100 % Linien-Erkennung, 98,7 % Linien-Entfernung). Dabei wird von modernen Texten mit korrekt binarisierten Linien ausgegangen. In Kaur und Mittal (2016) wird ein Ansatz basierend auf Neuronalen Netzen vorgestellt, bei dem die anschließende OCR mit und ohne entfernte Linien verglichen wird. In den vorgestellten Beispielen wird jeweils ein Ergebnis von 100 % erzielt, allerdings fehlt eine systematische Evaluation.

Im Kontext der Detektion von Arbeitsspuren in alten Drucken ist die Aufgabe komplexer, da zwischen handschriftlichen Unterstreichungen und gedruckten Linien unterschieden werden muss und das Ziel nicht das Entfernen der Linien ist, sondern die Extraktion der unterstrichenen Wörter. Dazu muss der Kontext einer Unterstreichung mit ausgewertet werden: Der Abstand zum Drucktext ist oberhalb von Unterstreichungen im Allgemeinen wesentlich kleiner als oberhalb von gedruckten Linien.

3 Methoden

Für die Erkennung von Linien wurde ein Deep Fully Convolutional Neuronal Net (FCN) auf binarisierten Bildern trainiert, dessen Architektur auf der U-Net Architektur mit Skip-Connections von Long et al. (2015) beruht, die in Wick und Puppe (2018) weiterentwickelt wurde. Zur Binarisierung wurde ein einfaches Schwellwertverfahren verwendet. Da das manuelle Auszeichnen von Unterstreichungen sehr aufwändig ist und sehr viele Trainingsdaten erforder-

2 Bai und Huo (2004).

3 Vgl. Rovina et al. (2015).

Abb. 1: Das mittlere Bild zeigt eine Druckseite mit generierten Unterstreichungen, die durch Überlagerung des linken und des rechten Bildes erzeugt wurden (mit Beachtung, dass der Text nicht durchgestrichen wird).

lich sind, wurden die Trainingsdaten generiert. Dazu wurden Druckseiten durch generierte, zufällig platzierte Unterstreichungen angereichert (s. Abb. 1). Die Unterstreichungen wurden auf einem Tablet manuell erstellt (rechtes Teilbild) und zufällig in die Zeilenzwischenräume eines Original-Scans (linkes Teilbild) eingesetzt, so dass der Text zwar berührt werden kann, aber nicht durchgestrichen ist (mittleres Teilbild).

Das trainierte FCN sagt für jeden Pixel in einem neuen Bild dessen Klasse voraus, in diesem Fall Drucktext, Unterstreichung oder Sonstiges.

Ein Beispiel für einen Teil eines Original-Scans mit echten Unterstreichungen zeigt Abb. 2, die zwei zentrale Probleme verdeutlicht: Zum einen ähneln die Linien links und rechts der obigen Seitenzahl den Unterstreichungen im Text und zum anderen ist die Unterstreichung bei „Man verwies mich zur Ruhe“ wesentlich dezenter als die übrigen Unterstreichungen. Letzteres Problem führt dazu, dass manche Unterstreichungen nach der Binarisierung der Bilder (Umwandlung in ein Binärbild, das nur schwarze oder weiße Pixel enthält) nur durch wenige schwarze Pixel oder gar nicht mehr sichtbar sind. Ersteres Problem führt dazu, dass zu viele Linien als Unterstreichungen erkannt werden. Daher sind Nachbearbeitungsschritte erforderlich, die mit heuristischen Regeln unter Nutzung empirisch ermittelter Schwellwerte umgesetzt werden. Dazu gehören:

1. Oberhalb einer Unterstreichung muss sich in geringem Abstand eine Mindestmenge an schwarzen Pixel befinden, nämlich der unterstrichene Text (jeweils definiert durch empirisch ermittelte Schwellwerte), was bei Drucklinien im Allgemeinen nicht der Fall ist.
2. Die erkannte Linie darf aber nicht direkt den Text berühren, d. h. unmittelbar oberhalb der Linie dürfen sich nur wenige schwarze Pixel befinden (vgl. Abb. 3; s. aber auch Abb. 1 für Unterstreichungen, die sich mit einzelnen Buchstaben überschneiden).
3. Erkannte Fragmente von Unterstreichungen, die sich auf gleicher Höhe in einem nicht zu großen Abstand befinden, werden zu einer Linie verbunden.

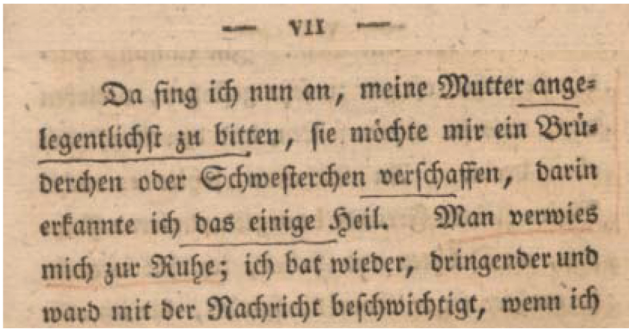

Abb. 2: Beispiel eines Teils eines echten Scans mit handschriftlichen Unterstreichungen. Die letzte Unterstreichung ist wesentlich dezenter und kann bei der Binarisierung verloren gehen.

Abb. 3: Vom FCN zunächst falsch erkannte Unterstreichungen unter den Buchstaben m, n, u (dicke Linien), die durch die dritte Nachbearbeitungsregel zu einer langen Unterstreichung verbunden würden (dünne Linie), aber durch die zweite Nachbearbeitungsregel (Unterstreichung berührt in hohem Maße den Text) eliminiert wird.

Das FCN klassifiziert jedes einzelne Pixel, ob es zur Klasse Unterstreichung gehört oder zu einer anderen Klasse. Im Rahmen dieser Arbeit sind nur schwarze Pixel (Drucktext oder Unterstreichungen) interessant, die sogenannten Vordergrund-Pixel (foreground). Aus diesen werden zusammenhängende Komponenten berechnet. Wenn für eine erkannte Unterstreichung mindestens 50 % der Vordergrund-Pixel mit der Ground Truth, d. h. der manuell annotierten Unterstreichung, übereinstimmen, wird diese als korrekt gewertet, d. h. True Positive (TP). Zu viel erkannte Unterstreichungen werden als False Positive (FP) und nicht erkannte Unterstreichungen als False Negative (FN) bewertet. Da es auch viele Drucklinien im Text gibt, wird das korrekte Nicht-Erkennen dieser Linien als Unterstreichung als True Negative (TN) gewertet.

4 Daten

Während für das Training Unterstreichungen in vorhandene Drucktexte künstlich hineingeneriert wurden, wurden für die Evaluation echte Unterstreichungen in Drucktexten manuell markiert und als Ground Truth verwendet. Die Markierung besteht aus rechteckigen Regionen, die die Unterstreichung einschließen (dies kann bei Berührung der Unterstreichungen mit dem unterstrichenen Text zu Ungenauigkeiten führen, die korrigiert wurden). Die Daten wurden in zwei Datensätze aufgeteilt: Der erste Datensatz-1 enthält sehr viele Unterstreichungen und nur wenige Drucklinien. Er besteht aus 42 Scan-Seiten mit insgesamt 196 Unterstreichungen. Der zweite Datensatz enthält viele Drucklinien und nur wenige Unterstreichungen. Er besteht aus 386 Scans mit 2016 Drucklinien und nur 59 Unterstreichungen. Mit Datensatz-1 sollte getestet werden, ob Unterstreichungen gut erkannt werden. Mit Datensatz-2 sollte darüber hinaus getestet werden, ob Unterstreichungen und Drucklinien unterschieden werden können.

Ein Problem ist der geringe Kontrast mancher Unterstreichungen zum Hintergrund (vgl. Abb. 2), da diese nach der Binarisierung in seltenen Fällen komplett verloren gehen oder nur durch relativ wenige, nicht zusammenhängende Pixel repräsentiert werden: In Datensatz-1 gingen von 196 Unterstreichungen zehn komplett verloren, in Datensatz-2 vier von 59.

5 Ergebnisse

Die Ergebnisse des FCN ohne Nacharbeitung auf den beiden Datensätzen zeigten wie erwartet keine gute Ergebnisse, die sich aber durch Nachbearbeitung deutlich verbesserten. Auf dem Datensatz-1 wurden von 186 Unterstreichungen 167 gefunden (Recall: ca. 90 %), wobei das Kriterium war, dass mindestens die Hälfte der Pixel einer Unterstreichung überdeckt wurden. Dabei waren von den 169 vorhergesagten Unterstreichungen nur 2 falsch (Precision ca. 99 %). Das ergibt einen F1-Wert von ca. 94 % (s. Abb. 4).

TP	FN	FP	Precision	Recall	**F1**
167	19	2	98,8%	89,8%	**94,1%**

Abb. 4: Evaluationsergebnisse auf dem Datensatz-1. TP: True Positive (korrekt erkannte Unterstreichungen), FN: False Negative (nicht erkannte Unterstreichungen), FP: False Positive (fälschlicherweise erkannte Unter-streichungen), Precision = TP / (TP+FP), Recall = TP / (TP + FN), F1 = 2 * Precision * Recall / (Precision + Recall).

Auf dem Datensatz-2 wurden von den 59 Unterstreichungen 50 erkannt und von den 2016 Drucklinien 1957 korrekt als Nicht-Unterstreichungen erkannt, und nur 7 mit Unterstreichungen verwechselt. Das entspricht einer Genauigkeit (Accuracy) von ca. 99 % (s. Abb. 5).

TP	TN	FN	FP	Precison	Recall	F1	**Accuracy**
50	1957	9	7	87,7%	84,7%	86,2%	**99,2%**

Abb. 5: Evaluationsergebnisse auf dem Datensatz-2. TP: True Positive (korrekt erkannte Unterstreichungen), TN: True Negative (korrekt erkannte Drucklinien), FN: False Negative (nicht erkannte Unterstreichungen), FP: False Positive (fälschlicherweise erkannte Unterstreichungen), Accuracy = (TP + TN) / (TP + TN + FP + FN).

6 Fehleranalyse

Die weitaus meisten Fehler der nicht erkannten Unterstreichungen (FN) beim Datensatz-1 und auch beim Datensatz-2 sind durch eine unzureichende Binarisierung bedingt, wie Abb. 6 zeigt. Abb. 7 enthält Beispiele, bei denen zu kleine Teile der Unterstreichungen erkannt wurden. Abb. 8 zeigt als Unterstreichungen erkannte Linien, oberhalb derer sich schwarze Pixel befinden, die aber keinen Text darstellen (was derzeit nicht überprüft wird, aber geplant ist, s. Ausblick).

Abb. 6: Die beiden Unterstreichungen in der obersten Zeile wurden nicht erkannt, weil aufgrund der Binarisierung des Bildes die relativ wenigen und nicht zusammenhängenden Pixel vom FCN nicht als Unterstreichung identifiziert wurden. Das kann daher durch die Nachkorrektur nicht verbessert werden.

Abb. 7: Hier wurde zwar eine Unterstreichung am Anfang der beiden Zeilen im linken Bild bzw. ein Teil der Unterstreichungen im rechten Bild gefunden. Die Unterstreichung im linken Bild erstreckt sich bis zum Ende der Zeilen, was durch die schlechte Binarisierung (es sind nur einzelne Punkte übrig geblieben) nicht erkannt wurde und deswegen einen Fehler (False Negative) darstellt, da für ein True Positive mindestens 50 % einer Unterstreichung abgedeckt werden muss. Im rechten Bild wurden die beiden gefundenen Unterstreichungen ebenfalls als Fehler gewertet, da weniger als 50 % der Pixel der Unterstreichung überdeckt sind. Allerdings würden im rechten Bild die darüberstehenden Wörter als unterstrichen markiert, weswegen diese Fehler im geplanten nächsten Schritt (der Erkennung der Wörter mittels OCR) korrigiert werden könnten.

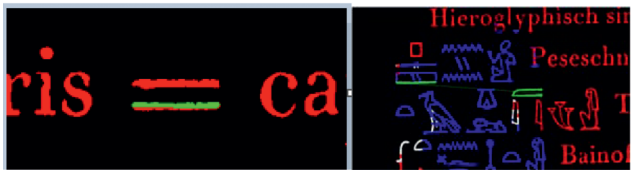

Abb. 8: Als Unterstreichungen erkannte Linien (jeweils grün), die nicht mit der ersten Regel verworfen wurden, da sich dicht oberhalb der Linien schwarze Pixel befinden.

7 Zusammenfassung und Ausblick

Es wurde ein vielversprechender Ansatz zur Erkennung unterstrichener Zeilen in binarisierten Scans von alten Drucken vorgestellt, der aus zwei Teilen besteht: Zunächst werden mit einem auf generierten Daten trainierten Fully Convolutional Neural Net (FCN) Unterstreichungen erkannt und im zweiten Schritt werden diese mit heuristischen Regeln nachbearbeitet, um erkannte Teillinien zu verbinden und Unterstreichungen von Linien im Drucktext zu unterscheiden. Der Ansatz hat in einem manuell erstellten Evaluationsdatensatz von 186 vorhandenen Unterstreichungen 167 gefunden und nur 2 zu viel vorhergesagt (F1 von 94 %) und in einem zweiten, größeren Datensatz gezeigt, dass Drucklinien gut von manuellen Unterstreichungen unterschieden werden können.

Es ist geplant, die Arbeiten weiterzuführen und im nächsten Schritt die Drucktexte oberhalb der Unterstreichungen automatisch mittels OCR zu ermitteln. Wie in Abb. 7 und 8 angedeutet, kann das bewirken, dass dadurch Fehler der Verwechselung von Drucklinien und Unterstreichungen verringert werden, da direkt oberhalb dieser Linien kein Drucktext vorhanden ist. Die meisten Fehler sind aber eine Folge der bisherigen einfachen, schwellwertbasierten Binarisierung. Daher sollen Techniken der adaptiven Binarisierung eingesetzt werden, um kontrastarme handschriftliche Unterstreichungen besser zu transformieren. Weiterhin sollen zukünftig auch andere Artefakte in den Scans der alten Drucke wie handschriftli-

che Notizen mit dem Neuronalen Netz und Nachbearbeitung erkannt werden.

Literaturverzeichnis

Bai, Z.-L.; Huo, Q. (2004): Underline Detection and Removal in a Document Image Using Multiple Strategies, In: *Proc. of the 17th int. conf. on pattern recognition* (ICPR04), 2, 578–81.

Denecke, L.; Teitge, I. (1989): Die Bibliothek der Brüder Grimm. Annotiertes Verzeichnis des festgestellten Bestandes. Weimar/Stuttgart.

Friemel, B. (2005): Die Grimm-Bibliothek. In: *Die Brüder Grimm in Berlin. Bilder, Studien, Dokumente*. [Ausstellungskatalog], 97–114.

Kaur, T.; Mittal, R. (2016): Hand-Drawn Annotation and Underline Detection and Removal in Scanned Documents Using Artificial Neural Network & Fuzzy C-Means Clustering. In: *European Journal of Advances in Engineering and Technology*, 3 (1), 12–20.

Long, J.; Shelhamer, E.; Darrell, T. (2015): Fully convolutional networks for semantic segmentation. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–40.

Pratihar, S.; Bhowmick, P.; Sural, S.; Mukhopadhyay, J. (2012): Detection and removal of hand-drawn underlines in a document image using approximate digital straightness. In: *Proc. Workshop on Document Analysis and Recognition* (DAR '12), 124–31.

Rauch, Y. (2018): Verborgene, verteilte und rekonstruierte Büchersammlungen: Gelehrtenbibliotheken an der Friedrich-Wilhelms-Universität zu Berlin. In: *Autorschaft und Bibliothek*, Göttingen, 62–81.

Rovina, Bahila, S.; Kumar, S. (2015): A Review: Detection And Removal Of Hand-Drawn Annotation Lines. In: *Int. Journal of Advances in Science Engineering and Technology* (IJASEAT), 3 (3), 136–40.

Wick, C.; Puppe, F. (2018): Fully Convolutional Neural Networks for Page Cegmentation of Historical Document Images. In: *13th IAPR Int. Workshop on Document Analysis Systems* (DAS), 287–92.

Alexander Gehrke
Universität Würzburg
Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik
Am Hubland
D-97074 Würzburg
Alexander.Gehrke@uni-wuerzburg.de

Nico Balbach
Universität Würzburg
Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik
Am Hubland
D-97074 Würzburg
Nico.Balbach@uni-wuerzburg.de

Yong-Mi Rauch
Humboldt Universität Berlin
Universitätsbibliothek
Unter den Linden 6
D-10099 Berlin
Yong-Mi.Rauch@ub.hu-berlin.de

Andreas Degkwitz
Humboldt Universität Berlin
Universitätsbibliothek
Unter den Linden 6
D-10099 Berlin
andreas.degkwitz@ub.hu-berlin.de

Frank Puppe
Universität Würzburg
Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik
Am Hubland
D-97074 Würzburg
Frank.Puppe@uni-wuerzburg.de