Universitäts- und Landesbibliothek | Direktion | Magdalenenstr. 8 | 64289 Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

Deutsche Forschungsgemeinschaft
- elektronisch übermittelt an LIS@dfg.de -

Universitäts- und Landesbibliothek Darmstadt

Leitender Bibliotheksdirektor

ULB

Prof. Dr. Thomas Stäcker

Magdalenenstr. 8
64289 Darmstadt

Tel. +49 6151 16 -76202
Fax +49 6151 16 - 76201
direktion@ulb.tu-darmstadt.de

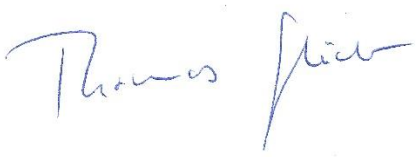
[21.5.2020]

Betrifft: Ausschreibung „Implementierung der OCR-D-Software zurVolltextdigitalisierung historischer Drucke“

Sehr geehrte Dame, sehr geehrter Herr,

nachstehend übersende ich Ihnen fristgerecht unsere Absichtserklärung zur Teilnahme an der Ausschreibung „Implementierung der OCR-D-Software zur Volltextdigitalisierung historischer Drucke“. Für Rückfragen stehe ich jederzeit gern zur Verfügung.

Mit freundlichen Grüßen,
Ihr



Prof. Dr. Thomas Stäcker

Absichtserklärung der ULB Darmstadt zur Teilnahme an der DFG Ausschreibung „Implementierung der OCR-D-Software zur Volltextdigitalisierung historischer Drucke“

Die Universitäts- und Landesbibliothek Darmstadt beabsichtigt die Einreichung eines Förderantrages im Rahmen der Ausschreibung „Implementierung der OCR-D-Software zur Volltextdigitalisierung historischer Drucke“. Bereits in der vorherigen Förderphase des Projektes OCR-D konnten wir das Projekt als Pilotbibliothek begleiten und die entwickelte Software anhand ausgewählter Quellmaterialien testen und dabei Rückmeldungen an die Entwickler geben. Um diese fruchtbare Kooperation fortzusetzen, beabsichtigen wir die Integration von OCR-D in den Regelworkflow der Digitalisierung unserer historischen Bestände.

Bereits begonnen haben wir mit der Nutzung verschiedener Module für die Erkennung von Ausgaben des „Darmstäder Tagblatts“ aus dem 18. und frühen 19. Jahrhundert. Durch die Periodizität dieser Quelle ergeben sich gute Vergleichsmöglichkeiten zur Abhängigkeit der Erkennungsqualität von Schwankungen der Vorlagenqualität, während der geringe Umfang der Einzelausgaben schnell Ergebnisse liefert, die mit anderen Werkzeugen verglichen werden können, die aber auch unmittelbar für die Bibliothek nutzbar sind.
Diese Vergleiche werden wir im Rahmen der Pilotphase fortführen und hierbei erste Messdaten zur Performance verschiedener Module und Modulkombinationen innerhalb eines bereits etablierten Workflows erheben und diese Messdaten mit dem genannten Workflow vergleichen.

Bei der aktuellen Umstellung auf die Software Kitodo werden wir OCR-D als Bestandteil des Regelworkflows integrieren und uns dazu mit weiteren Institutionen mit dieser Software abstimmen. Die Erfahrungen werden dokumentiert und allen interessierten Institutionen zugänglich gemacht. Die Nutzung im Regelworkflow soll es ermöglichen, dass bereits in der Pilotphase alle im Haus digitalisierten Drucke (derzeit rund 1 Mio Images) mittels OCR-D erkannt werden. Auch hierbei wollen wir Messdaten zur Feststellung der Performance erheben und die benötigte Rechenleistung sowie die genau Implementierung dokumentieren. Von besonderem Interesse sind hierbei wiederholte OCR Prozesse auf denselben Materialien.

Neben der Beurteilung der Performance wird auch ein Vergleich der Erkennungsergebnisse nach Qualität und Aufbau erfolgen und zu dokumentieren sein. Hieraus ergeben sich bereits erste qualitative Hinweise auf besonders vielversprechende Werkzeug/Modell-Kombinationen für bestimmte Aufgaben.

Im Falle einer Bewilligung des Antrages werden wir diese Arbeiten fortführen und systematisieren, sodass sich z. B. anhand des Tagblattes umfangreiche und unmittelbar vergleichbare Messdaten ergeben.

Weiterhin wollen wir prüfen, welche Angaben zur Statistik die den Modulen zugrundeliegenden Werkzeuge liefern und inwiefern sich hieraus Metriken zur Beurteilung der Erkennungsqualität ableiten lassen.

Ziel soll es dabei sein, mögliche Probleme (z. B. unpassende Tool/Modell-Kombination) frühzeitig zu erkennen, um so übermäßige Zeitverluste bei umfangreichen Erkennungsjobs zu vermeiden.

Die Vergleiche der Ergebnisse werden wir weiterhin dafür nutzen, ein „Computational Double Keying“ zu erarbeiten: Durch den maschinellen Vergleich der Ergebnisse verschiedener Engines und Modelle kann der Korrekturaufwand merklich gesenkt werden, da nur noch Zweifelsfälle manuell kontrolliert werden müssten. Wie auch aus den vorgenannten Metriken ließe sich so auch eine Qualitätsabschätzung erreichen.