Workshop „Verfahren zur Verbesserung von OCR-Ergebnissen“

Protokoll zu den Ergebnissen und Empfehlungen des Workshops

Ort: Geschäftsstelle der DFG, Kennedyallee 40, 53175 Bonn

Termin: Mittwoch, 12. März 2014, 13:00 Uhr – 18:30 Uhr

Donnerstag, 13. März 2014, 9:00 Uhr – 12:00 Uhr

Programm: siehe Anlage 1

Teilnehmer: siehe Anlage 2

Hintergrund

Das Positionspapier „Die digitale Transformation weiter gestalten – Der Beitrag der Deutschen Forschungsgemeinschaft zu einer innovativen Informationsinfrastruktur für die Forschung“ formuliert für das Förderprogramm „Erschließung und Digitalisierung“, das zum Förderbereich „Wissenschaftliche Literaturversorgungs- und Informationssysteme“ gehört, die Notwendigkeit, wo immer möglich und sinnvoll, neben der Bereitstellung von Image-Digitalisaten über eine DFG-Förderung auch Volltexte bereitzustellen. Für Drucke des 16., 17. und 18. Jahrhunderts lag der Schwerpunkt der DFG-Förderung bislang auf der Herstellung von Image-Digitalisaten. Auf der Grundlage der umfangreich vorliegenden und qualitativ hochwertigen Image-Digitalisate sollten zur Erhöhung der Anzahl der verfügbaren Volltexte entsprechend dem Positionspapier gezielt Förderimpulse gesetzt werden, um Verfahren zur Verbesserung der OCR-Ergebnisse von digitalisierten Drucken zu entwickeln und zu etablieren.

Ziel des Workshops war es, einen Überblick darüber zu gewinnen, welche OCR-Ergebnisse auf der Grundlage bestehender Werkzeuge und Verfahren bereits erzielt werden können und in welchen Bereichen Entwicklungsbedarf besteht. Auf der Grundlage von Impulsvorträgen zu verschiedenen Themenbereichen wurden mögliche Eckpunkte einer DFG-Förderung, z.B. im Format einer Ausschreibung, zu Verfahren zur Verbesserung von OCR-Ergebnissen von digitalisierten Drucken diskutiert.

Ergebnisse und Empfehlungen

Sowohl aus wissenschaftlicher als auch aus informationsinfrastruktureller Sicht wurde der **Auf- und Ausbau von historischen Textkorpora und lexikalischen Ressourcen** als zentrales Handlungsfeld angesprochen. Voraussetzung für ein sinnvolles Arbeiten mit digitalen Volltexten sei es aus wissenschaftlicher Sicht, dass einschlägige und ausreichend umfangreiche Korpora vorhanden seien. In diesem Zusammenhang wurde darauf hingewiesen, dass vor dem Hintergrund der sich voneinander unterscheidenden, sich zum Teil aber auch ergänzenden Anwendungsfelder und Fragestellungen der Forschung die Notwendigkeit bestehe, genre-, epochen- und sprachspezifische Korpora aus Nachschlagewerken, Wörterbüchern und Texten auf- und auszubauen. Aus informationsinfrastruktureller Sicht sei das Bestehen entsprechender Referenzkorpora die Voraussetzung für das Trainieren bereits vorhandener Softwares sowie die gezielte Weiterentwicklung von OCR-Verfahren. Nur über den Auf- und Ausbau von Korpora sei es möglich, die Bestimmung der Ground Truth von spezifischen Drucken zu erleichtern.

Da sowohl kommerzielle als auch Open-Source-OCR-Engines im Ergebnis von OCR-Durchläufen derzeit und auch perspektivisch nicht dazu in der Lage seien, eine Textgenauigkeit von 100% zu erzielen und die Ergebnisse schrift- und druckbezogen stark variierten, bestehe außerdem Entwicklungsbedarf im Feld der **Weiterentwicklung von Nachkorrekturanwendungen**. Zur Nutzung offener Korrektur und Annotationsanwendungen seien auch Methoden des Crowdsourcing zu berücksichtigen. Gegebenenfalls sei in diesem Zusammenhang auch die **Weiterentwicklung von Open-Source-OCR-Engines** zur Verbesserung aktuell möglicher OCR-Ergebnisse zu erwägen. Da auch bezogen auf bestehende Verfahren zur OCR-Bearbeitung von Drucken des 19. Jahrhunderts noch

Entwicklungsbedarf bestehe, sollte das 19. Jahrhundert unbedingt in entsprechende Entwicklungen einbezogen werden.

Ziel müsse es sein, bereits bestehende und noch aufzubauende Korpora umfassend nachnutzen zu können. Vor diesem Hintergrund sei bei eventuellen Förderaktivitäten im Feld der Weiterentwicklung von OCR-Verfahren unbedingt zu gewährleisten, dass die entstehenden Daten zur **offenen Nachnutzung** zur Verfügung gestellt würden. Die umfassende Nachnutzungsmöglichkeit müsse für Nutzerinnen und Nutzer klar erkennbar sein.

Bezogen auf akzeptable Fehlerquoten für Volltexte, die über OCR generiert werden, wurde festgestellt, dass sich keine einheitlichen Quoten definieren ließen. Vielmehr sei die Frage nach **akzeptablen Fehlerquoten** bzw. **erforderlichen Genauigkeitsquoten** abhängig von den Anwendungsfeldern bzw. der spezifischen Fragestellung, die an das Textkorpus gerichtet werde. Die Aussagen der DFG-Praxisregeln „Digitalisierung“ (Stand: 2/2013) zu akzeptablen Genauigkeiten seien vor diesem Hintergrund zu relativieren. Auch müsse bei der Auseinandersetzung mit der Frage nach akzeptablen OCR-Ergebnissen immer berücksichtigt werden, dass der Textbegriff differenziert zu betrachten und in Ergänzung des reinen Textes auch die Struktur einer gedruckten Seite von zentraler Bedeutung sei. In Ergänzung der Weiterentwicklung von OCR-Verfahren bezogen auf Texte seien auch Verfahren der Text-Bild-Erkennung in den Blick zu nehmen.

Bezogen auf die **Genauigkeitsberechnung** für die Ergebnisse von OCR-generiertem Text sei es von zentraler Bedeutung, dass die Berechnungsmethode **transparent** gemacht würde. Die DFG-Praxisregeln „Digitalisierung“ lieferten bereits konkrete Hinweise zu einer sinnvollen Berechnung – diese seien ggf. noch zu ergänzen und möglichst einheitlich anzuwenden.

Transparenz sei auch bezogen auf eventuelle **Nachbearbeitungen** bzw. **Versionierungen** von Volltexten für die wissenschaftliche Arbeit mit den Texten grundlegend. Nur wenn persistente Identifizierungen konkreter Versionen eines Volltextes möglich seien, könne wissenschaftlich präzise und eindeutig mit dynamisch sich veränderten Volltexten gearbeitet werden.

Auf organisatorisch-infrastruktureller Ebene sei es von besonderer Bedeutung, den **gesamten Workflow von Volltextgenerierung über OCR**, und insbesondere Lücken bzw. Probleme, wie z.B. die nur teilweise vorhandene Interoperabilität bestehender Datenformate (Import, Export und Speicherung) in den Blick zu nehmen. Es bestehe Bedarf, in dieser Hinsicht interoperable Standards zu entwickeln. Bezogen auf angereicherte bzw. korrigierte Volltexte sei zu fragen, wie die ergänzten Versionen standardisiert in das Ausgangsrepositorium zurückfließen können. Auch Fragen der Langzeitarchivierung von versionierten Volltexten sowie deren persistente Adressierung seien zu beantworten. Schließlich müssten in diesem Zusammenhang auch Visualisierungstools (weiter-)entwickelt werden, wobei unbedingt auf bestehenden Entwicklungen – auch außerhalb des Bibliothekssektors – aufgebaut werden solle (DFG-Viewer und Visualisierungsdienste außerhalb des Bibliothekssektors).

Bezogen auf bereits vorhandene Volltexte – z.B. auf der Grundlage der Public-Private-Partnership zwischen der BSB und Google erstelltem Volltext – wurde festgestellt, dass es

die Nachnutzbarkeit dieser Daten eine zentrale Voraussetzung für Weiterentwicklungen von OCR-Verfahren darstellten. Von zentraler Bedeutung sei es in diesem Zusammenhang, **Services und Workflows** zu etablieren, die es Nutzerinnen und Nutzern ermöglichen, weiter mit diesen Daten zu arbeiten und die Ergebnisse wiederum zur freien Nachnutzung zur Verfügung zu stellen. Für die denkbaren Nutzungsszenarien seien **Use Cases** und darauf basierend **Services** zu entwickeln, die die Felder Speicherung/Langzeitarchivierung, Werkzeuge zur Nachbearbeitung und Veredelung (z.B. über Crowdsourcing, Normdatenanreicherung etc.), Re-Integration der verbesserten bzw. angereicherten Texte in ein Repositorium bzw. Repositorien und Präsentation des Volltextes abdecken. Für sämtliche der genannten Felder seien **Standards und interoperable Datenformate** von grundlegender Bedeutung. Die Entwicklung von Use Cases könne eine gute Grundlage für eventuelle Eckpunkte bzw. modulare Bausteine einer größer angelegten Fördermaßnahme darstellen.

Im **Ergebnis der Diskussion** formulierten die Teilnehmerinnen und Teilnehmer, dass eine **koordinierte Fördermaßnahme** der DFG im Feld der Weiterentwicklung von OCR-Verfahren dringend erforderlich sei. Über eine entsprechende Ausschreibung im Bereich „Wissenschaftliche Literaturversorgungs- und Informationssysteme“ der DFG könnten Angebote für die Forschung geschaffen werden, die neue Arbeitsmöglichkeiten schaffen und ggf. auch neue Methoden und Fragestellungen stimulieren könnten. Ziel einer entsprechenden Fördermaßnahme solle es sein, nicht in die Masse der OCR-Bearbeitung von Texten einzusteigen, sondern vielmehr aufeinander abgestimmte Standardbildungen und Workflowverbesserungen herbeizuführen, die die Grundlage für eventuelle spätere OCR-Bearbeitungen in der Breite bilden. Bei sämtlichen Entwicklungen sei davon auszugehen, dass die Services in den meisten Fällen von Informationsinfrastruktureinrichtungen entwickelt und bereit gestellt würden, eine enge Abstimmung mit den Bedarfen und Anwendungsfeldern der Wissenschaft aber zentral sei.

Workshop „Verfahren zur Verbesserung von OCR-Ergebnissen“

Programm

Mittwoch, 12. März 2014

13:00 – 13:30 Uhr	**Begrüßung und Einführung** Dr. Franziska Regner, DFG-Geschäftsstelle
13:30 – 14:30 Uhr	**Möglichkeiten und Grenzen von OCR – Impulsvortrag mit anschließender Diskussion** Dr. Uwe Springmann, Centrum für Informations- und Sprachverarbeitung, LMU München
14:30 – 15:30 Uhr	**Forschungsfragen und Fehlertoleranz – vier Kurzstatements aus der Wissenschaft** Prof. Dr. Fotis Jannidis, Institut für Deutsche Philologie, Universität Würzburg Dr. Christoph Schöch, Institut für Deutsche Philologie, Universität Würzburg Prof. Dr. Charlotte Schubert, Historisches Seminar, Universität Leipzig Dr. Thomas Burch, Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften, Universität Trier
15:30 – 16:00 Uhr	Kaffeepause
16:00 – 16:30 Uhr	**Workflow Volltextgenerierung über OCR** Ralf Stockmann, Staatsbibliothek zu Berlin
16:30 – 17:00 Uhr	**Verfahren zur Feststellung von Genauigkeiten** Dr. Thomas Stäcker, Herzog August Bibliothek Wolfenbüttel
17:00 – 17:30 Uhr	**Bestehende Werkzeuge und Softwares** Dr. Günter Mühlberger, Institut für Germanistik, Universität Innsbruck
17:30 – 18:00 Uhr	**OCR-Kompetenzen im deutschsprachigen Raum** Christoph Stollwerk, Institut für Historisch-Kulturwissenschaftliche Informationsverarbeitung, Universität zu Köln

Donnerstag, 13. März 2014

9:00 – 10:00 Uhr	**Best Practices und Pilotprojekte** Dr. Markus Brantl und Karl Märker, Bayerische Staatsbibliothek Maria Federbusch, Staatsbibliothek zu Berlin Dr. Alexander Geyken, Arbeitsstelle Digitales Wörterbuch der Deutschen Sprache, Berlin-Brandenburgische Akademie der Wissenschaften
10:00 – 10:20	**Verarbeitung nicht OCR-geeigneter Dokumente** Prof. Dr. Arved C. Hübler, Institut für Print- und Medientechnik, TU Chemnitz
10:20 – 10:40 Uhr	Kaffeepause
10:40 – 11:00 Uhr	**Minimalstandards für die Bereitstellung von Volltexten über den DFG-Viewer** Sebastian Meyer, Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden
11:00 – 12:00 Uhr	**Abschlussdiskussion**

Workshop „Verfahren zur Verbesserung von OCR-Ergebnissen“

Teilnehmerinnen und Teilnehmer

Teilnehmer/in	Einrichtung
Dr. Sebastian Barteleit	Bundesarchiv
Dr. Markus Brantl	Bayerische Staatsbibliothek
Dr. Thomas Burch	Universität Trier, Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften
Professor Dr. Thomas Bürger	Sächsische Landesbibliothek - Staats- und Universitätsbibliothek
Dr. Holger Busse	Staatsbibliothek zu Berlin
Dr. Klaus Ceynowa	Bayerische Staatsbibliothek
Dr. Marianne Dörr	Universitätsbibliothek Tübingen
Maria Federbusch	Staatsbibliothek zu Berlin
Dr. Alexander Geyken	Berlin-Brandenburgische Akademie der Wissenschaften
Professor Dr.-Ing. Arved C. Hübler	Technische Universität Chemnitz, Institut für Print- und Medientechnik
Professor Dr. Fotis Jannidis	Julius-Maximilians-Universität Würzburg, Institut für Deutsche Philologie
Karl Märker	Bayerische Staatsbibliothek
Sebastian Meyer	Sächsische Landesbibliothek - Staats- und Universitätsbibliothek
Dr. Günter Mühlberger	Universität Innsbruck, Institut für Germanistik
Professor Dr. Stephan Walter Müller	Universität Wien, Institut für Germanistik
Dr. Sven Schlarb	Österreichische Nationalbibliothek
Dr. Christof Schöch	Julius-Maximilians-Universität Würzburg, Institut für deutsche Philologie
Professor Dr. Charlotte Schubert	Universität Leipzig, Historisches Seminar
Dr. Uwe Springmann	Ludwig-Maximilians-Universität München, Centrum für Informations- und Sprachverarbeitung (CIS)
Dr. Thomas Stäcker	Herzog August Bibliothek Wolfenbüttel
Ralf Stockmann	Staatsbibliothek zu Berlin
Christoph Stollwerk	Universität zu Köln, Institut für Historisch-Kulturwissenschaftliche Informationsverarbeitung

Teilnehmer/in	Einrichtung
Ulrike Hintze	Geschäftsstelle der DFG
Dr. Angela Holzer	Geschäftsstelle der DFG
Dr. Anne Lipp	Geschäftsstelle der DFG
Dr. Franziska Regner	Geschäftsstelle der DFG