Ausschreibung

Implementierung der OCR-D-Software zur Volltextdigitalisierung historischer Drucke

Eine Ausschreibung im Rahmen des Förderprogramms

„e-Research-Technologien“

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

Hintergrund

Die Bilddigitalisierung der im deutschen Sprachraum erschienenen Drucke des 16. bis 19. Jahrhunderts wurde vielfach vollzogen, sodass viele rare und unikale Drucke als Bilddigitalisate für die Forschung sichtbar und verfügbar sind. Um den Ansprüchen der Wissenschaft an die Nutzbarkeit der digitalisierten Drucke zu genügen, werden vor allem (maschinenlesbare) Volltexte der Bilddigitalisate benötigt. Volltexte von Bilddigitalisaten können durch Verfahren der automatisierten Layout-, Struktur- und Zeichenerkennung erzeugt werden. Die Entwicklung eines umfassenden Verfahrens zur Massenvolltextdigitalisierung der Drucke hat die DFG mit OCR-D (www.ocr-d.de) initiiert.

Im Rahmen einer ersten Ausschreibung wird seit 2015 ein Koordinierungsprojekt zur Weiterentwicklung von Verfahren der *Optical-Character-Recognition* gefördert. Basierend auf den Vorarbeiten des OCR-D-Koordinierungsprojekts werden seit 2018 im Rahmen einer zweiten Ausschreibung insgesamt acht Modulprojekte gefördert (https://ocr-d.github.io/de/module-projects), die die einzelnen Teilschritte des OCR-D-Workflows für die Text- und Strukturerkennung zur Volltextdigitalisierung historischer Drucke entwickeln. Mit Abschluss dieser Arbeiten wird im Jahr 2020 ein Prototyp der OCR-D-Software zur Verfügung stehen.

Die bisherigen Arbeiten in der OCR-D-Initiative haben zu wesentlichen Verbesserungen der Verfahren zur automatischen Text- und Strukturerfassung historischer Drucke geführt. Der Software-Prototyp verspricht durch seinen modularen Aufbau eine flexible Integration in bestehende (bibliothekarische) Workflow- und Digitalisierungssysteme. Die Implementierung der OCR-D-Software in Bibliotheken, Archiven sowie anderen bestandshaltenden bzw. bestandsverarbeitenden Einrichtungen ist folglich der nächste notwendige Schritt, damit die Erzeugung hochqualitativer Volltexte stattfinden kann.

Ziel der Ausschreibung

Zentrales Ziel der Ausschreibung ist die **Entwicklung (generischer) Implementierungspakete** mit akzeptabler Performanz für unterschiedliche Anforderungen. Unterschiedliche Anforderungen ergeben sich aus verschiedenen Digitalisierungsstrategien der Einrichtungen, wenn z. B. Dienstleister oder eigene Workflowsysteme genutzt werden. Je nach Bedarf der Einrichtungen sollen die Implementierungspakete möglichst viele unterschiedliche Anwendungsszenarien abdecken. Mögliche Anwendungsszenarien sind beispielsweise die Implementierung der OCR-D-Software auf üblichen Arbeitsplatzrechnern, die parallelisierte Implementierung auf Standardrechnern, die Implementierung als Webdienst, die Implementierung auf einem Hochleistungsrechner oder jede andere Implementierung, die nötig sein kann, um die massenhafte Erzeugung der Volltexte aus den Bilddigitalisaten zu ermöglichen. Ein Implementierungspaket soll das komplette Verfahren umfassen, das eine am OCR-D-Workflow interessierte Einrichtung benötigt, um die Erzeugung von Volltexten aus den Bilddigitalisaten im Rahmen der jeweiligen Workflowsysteme selbstständig betreiben zu können. Je nach Anwendungsszenario sollen die unterschiedlichen Ansprüche an Benutzbarkeit und Nutzerfreundlichkeit berücksichtigt werden.

Um das Ziel der Ausschreibung zu erreichen, ist zudem die Förderung der **Koordinierung bei der Entwicklung der Implementierungspakete** möglich. Dies umfasst u. a. die Abstimmung von Maßnahmen zur Qualitätsverbesserung der OCR-D-Software, die Standardisierung von Implementierungspaketen (z. B. Ein- und Ausgabeformate, Schnittstellen, Metadaten u. v. m.) und die Unterstützung für eine sinnvolle Abgrenzung der verschiedenen Implementierungspakete. Daher ist im Rahmen der Ausschreibung die Förderung eines Koordinierungsprojekts vorgesehen.

Voraussetzungen für die Förderung

Voraussetzung für eine Antragstellung ist eine bis zu sechsmonatige Pilotierung der OCR-D-Software für das jeweils angestrebte Anwendungsszenario; diese Arbeiten sind in Eigenleistung zu erbringen. Die Pilotierung dient der Qualitätssicherung der OCR-D-Software für den Praxiseinsatz. In dieser Zeit soll eine intensive Auseinandersetzung mit dem Prototyp der OCR-D-Software erfolgen, sodass auf dieser Basis die prinzipielle Durchführbarkeit eines Implementierungsvorhabens belegt wird (*proof of concept*) und so die konkrete Ausgestaltung eines Antrags für die Implementierung erfolgen kann. Während der Pilotierungszeit wird die Fortführung des Engagements des bestehenden OCR-D-Koordinierungsprojekts erwartet.

Die Pilotierung soll für unterschiedliche Anwendungsszenarien erfolgen, um so den späteren Einsatz der Implementierungspakete an möglichst vielen bestandshaltenden bzw. bestandsverarbeitenden Einrichtungen zu ermöglichen. Während der Pilotierungsphase sollen an den interessierten Einrichtungen die einzelnen OCR-D-Module sowie der gesamte OCR-D-Workflow getestet werden. Die Korrektheit der erzeugten Volltexte soll mit Blick auf die wissenschaftlichen Bedarfe und Ansprüche ebenfalls geprüft werden. Mit systematischen Softwaretests oder mit anwendungsbezogenen Funktions- und Implementierungstests sollen mögliche Fehlerquellen der OCR-D-Software aufgezeigt und dokumentiert werden. Diesen soll mit konkreten Verbesserungs- und Korrekturvorschlägen des Quellcodes der Software begegnet werden. Außerdem sollen Kennzahlen zu Funktionen, zur Stabilität, Laufzeit und Performanz erhoben und dokumentiert werden. Je nach Implementierungsszenario sollen verschiedene Schnittstellen zu anderen (bibliothekarischen) Workflow- und Digitalisierungssystemen erprobt werden. Die Erfahrungen aus dem Einsatz der OCR-D-Software in den bereits testenden Pilotbibliotheken (UB Darmstadt, SLUB Dresden, UB Heidelberg, UB Halle, UB Mannheim, SUB Göttingen, SBB Berlin, BBAW Berlin, HAB Wolfenbüttel) sollen berücksichtigt werden. Die Ergebnisse der Pilotierung sollen im Antrag unter dem Abschnitt „Vorarbeiten“ dargestellt werden, eine maximal fünfseitige Anlage ist ebenfalls möglich.

Voraussetzung für die Förderung eines OCR-D-Koordinierungsprojekts ist die Analyse der Koordinierungsbedarfe für die Zeit der Implementierungsprojekte. Diese Analyse kann technische und organisatorische Aspekte beinhalten und soll im Austausch mit den Einrichtungen, die Implementierungsanträge stellen wollen, erarbeitet werden. Die Zusammensetzung des Koordinierungsprojekts soll die Kompetenzen umfassen, die nötig sind, um den Bedarfen, die während der Pilotierungsphase ermittelt werden, gerecht werden zu können. Sollte sich herausstellen, dass für die Erfüllung dieser Bedarfe eine andere Zusammensetzung als diejenige

des derzeitigen OCR-D-Koordinierungsprojekts notwendig ist, kann dies entsprechend umgesetzt werden. Im Rahmen der Ausschreibung kann nur ein Antrag zur weiteren Koordinierung der OCR-D-Initiative gefördert werden, sodass entsprechende Absprachen und Abstimmungen im Vorfeld getroffen (Selbstorganisation) und im Antrag dargestellt werden müssen. Für einen solchen Antrag gelten die allgemeinen Anforderungen und Voraussetzungen des Förderprogramms e-Research-Technologien, jedoch die unten genannten Fristen.

Weitere Voraussetzung zur Förderung von Anträgen zur Entwicklung von Implementierungspaketen ist die Bereitschaft und im Antrag anzugebende Zusage der entsprechenden Einrichtungen zur abgestimmten Zusammenarbeit mit dem OCR-D-Koordinierungsprojekt während der Pilotierungs- und der zukünftigen Implementierungsprojektlaufzeit.

Im Bedarfsfall: Optimierung einzelner OCR-D-Module

In der den eigentlichen Implementierungsprojekten vorausgehenden Pilotierungsphase soll auch geklärt werden, ob begründeter Bedarf für die Weiterentwicklung oder Verbesserung der einzelnen OCR-D-Softwaremodule besteht. Sollte dies der Fall sein, können diese Bedarfe – in Absprache mit anderen Einrichtungen und unterstützt durch das OCR-D-Koordinierungsprojekt – gebündelt an die primären Modulentwickler herangetragen werden, die zur Umsetzung dieser Bedarfe einen Antrag im Rahmen der Ausschreibung stellen können. Für diese Anträge gelten die allgemeinen Anforderungen und Voraussetzungen des Förderprogramms e-Research-Technologien, jedoch die unten genannten Fristen.

Anforderungen an das Arbeitsprogramm

Förderanträge, sowohl für die Implementierung als auch für die Koordination sollen eine detaillierte Projekt- und Zeitplanung sowie ein klares Arbeitsprogramm mit Arbeitspaketen, Meilensteinen, selbst definierte Erfolgskriterien, Mengenangaben und Personalressourcen für das Vorhaben enthalten. Potentielle Risiken und Planungen zum Umgang mit diesen Risiken sollen ebenfalls skizziert werden. Um Synergieeffekte zwischen den Implementierungsprojekten zu nutzen, sind zeitnahe und aufeinander abgestimmte Starts aller Projekte wünschenswert.

Die Implementierungspakete sind für die spezifischen Anforderungen einer Einrichtung zu beschreiben und müssen zugleich so generalisierbar sein, dass ein entsprechendes Implementierungspaket auch an anderen Einrichtungen mit vergleichbaren Anforderungen nachgenutzt werden kann. Es bietet sich an, für die einzelnen Anwendungsszenarien jeweils Absprachen zwischen Einrichtungen mit vergleichbaren Anforderungen zu treffen (Selbstorganisation). Mit der Förderung sollen die jeweiligen Implementierungspakete möglichst weitgehend für die weitere Nutzung vorbereitet werden (*close to ready-to-use*), damit künftig auf dieser Basis die Massenvolltextdigitalisierung von Drucken aus dem 16. bis 19. Jahrhundert entsprechend wissenschaftlicher Anforderungen begonnen und durchgeführt werden kann.

Für die Anordnung flexibler, modularer Verarbeitungsketten sind die zur Interaktion und Integration vom (künftigen) OCR-D-Koordinierungsprojekt vorgegebenen Schnittstellen zu be-

dienen. Diese Interaktion soll durch das Koordinierungsprojekt unterstützt werden. Jedes geförderte Implementierungsvorhaben ist zudem verpflichtet, für die entwickelten Implementierungsszenarien nach den Vorgaben des Koordinierungsprojekts eine Software- und Benutzerdokumentation anzufertigen.

Das künftige OCR-D-Koordinierungsprojekt soll die Funktionsweise des OCR-D-Gesamtworkflows weiterhin sicherstellen und eine Lösung für die nachhaltige und dauerhafte Koordinierung und Betreuung der auch in Zukunft notwendig werdenden Support-Struktur für die OCR-D-Software erarbeiten und aufsetzen. In der Zeit nach der Projektförderung können zur Identifikation und Moderation von Maßnahmen zum Community-Building auch Mittel für Rundgespräche beantragt werden.

Allgemeine Anforderungen

Alle durch die geförderten Projekte zustande gekommenen Ergebnisse sind in der Fachöffentlichkeit bekannt zu machen und zur kostenlosen Nachnutzung für die Wissenschaft zur Verfügung zu stellen. Die Offenlegung der ggf. produzierten Quellcodes (*Open Source*) ist verpflichtend, die Bereitstellung der Projektergebnisse mittels eindeutiger Lizenzen an geeigneter Stelle (z. B. GitHub) wird vorausgesetzt. Das schließt die umfassende Dokumentation mit ein.

Sämtliche mit DFG-Förderung erstellte, über das Internet verfügbare Inhalte – auch Softwareentwicklungen – sind so aufzubereiten, zu indexieren und zu bewerben, dass eine maximale Auffindbarkeit gewährleistet wird. Entsprechende Metadaten müssen informationsfachliche Standards erfüllen und sich dazu eignen, auch in internationale, fachspezifische und informationsfachliche Nachweissysteme integriert zu werden. Für die im jeweiligen Implementierungsprojekt erzeugten Volltexte wird die Integration in die entsprechenden Nachweissysteme erwartet. Die Nutzung des DFG-Viewers zur Darstellung der Volltexte wird empfohlen.

Zur Gewährleistung der abgestimmten Zusammenarbeit verpflichten sich die Implementierungsprojekte gegenüber der DFG und gegenüber dem derzeitigen bzw. zukünftigen OCR-D-Koordinierungsprojekt zur Zusammenarbeit und formulieren dies entsprechend im Antrag als Erklärung einer Selbstverpflichtung. Im Gegenzug wird erwartet, dass sich das OCR-D-Koordinierungsprojekt zur Zusammenarbeit mit den Implementierungsprojekten bei deren Antragstellung durch einen *Letter of Support* verpflichtet.

Aufgrund der erwarteten Qualitätssicherung der OCR-D-Software während der Pilotierung müssen für die Durchführung der Projekte dieser Ausschreibung, d. h. Implementierungsprojekte und Koordinierungsprojekt, keine weiteren Eigenleistungen im Antrag aufgezeigt und während der Projektlaufzeit erbracht werden.

Art und Dauer der Förderung

Antragsberechtigt sind Hochschulen, Bibliotheken, Archive und andere Infrastruktureinrichtungen sowie außeruniversitäre Forschungseinrichtungen. Eine Antragstellung durch Konsortien, die mehrere Einrichtungen umfassen können, wird begrüßt.

Im Rahmen des Förderangebots können sämtliche im Programm e-Research-Technologien mögliche Module beantragt werden. Die Mittel müssen projektspezifisch begründet sein.

Die Dauer der Implementierungsprojekte soll in der Regel auf 24 Monate begrenzt sein. Die Dauer des Koordinierungsprojekts soll auf 36 Monate begrenzt sein.

Termine und Antragstellung

Einrichtungen, die planen für ein Anwendungsszenario einen Antrag zur Entwicklung eines Implementierungspakets zu stellen, werden um eine verbindliche, maximal dreiseitige Absichtserklärung zur Kurzdarstellung der Projektziele, des Anwendungsszenarios und der Vorarbeiten gebeten. Darin sollen vor allem auch die verantwortlichen Antragstellenden genannt werden. Alle Absichtserklärungen werden auf der Webseite der DFG veröffentlicht, damit eine Abstimmung zwischen den Einrichtungen, die ähnliche Implementierungspakete anstreben, erfolgen kann und damit das künftige OCR-D-Koordinierungsprojekt die Bedarfsanalyse vornehmen kann. Für ein künftiges OCR-D-Koordinierungsprojekt wird ebenfalls eine Absichtserklärung erwartet. Die vollständigen Absichtserklärungen sind bis 5. Mai 2020 an LIS@dfg.de abzugeben.

Förderanträge sind in deutscher Sprache zu verfassen und können bis 7. Oktober 2020 über das elan-Portal (https://elan.dfg.de) eingereicht werden. Die Antragstellenden werden gebeten, im Antrag auch dazu Stellung zu nehmen, wann sie mit der Arbeit beginnen können. Die Gesamtzahl der Seiten eines Antrags soll den Umfang von 20 Seiten nicht überschreiten. Für die Ergebnisse der Pilotierung kann eine Anlage (max. 5 Seiten) hinzugefügt werden.

Beachten Sie auch den Leitfaden für die Antragstellung - Projektanträge im Bereich Wissenschaftliche Literaturversorgungs- und Informationssysteme (DFG-Vordruck 12.01) sowie das Merkplatt zum Förderprogramm e-Research-Technologien (DFG-Merkblatt 12.19).

Handelt es sich bei dem Antrag um Ihren ersten Antrag bei der DFG, berücksichtigen Sie bitte, dass Sie sich vor der Antragstellung im elan-Portal registrieren müssen. Ohne Registrierung ist eine Antragstellung nicht möglich. Für die Umsetzung der Registrierung sollten mindestens 48 Stunden eingeplant werden.

Ansprechpersonen

Bei Rückfragen und zur Beratung wenden Sie sich bitte an:

- Fachliche Fragen und Förderbedingungen:
  Dr. Matthias Katerbow, Tel. +49 228 885-2358, Matthias.Katerbow@dfg.de
  Dr. Florian Werner, Tel. +49 228 885-2212, Florian.Werner@dfg.de
- Formale und organisatorische Fragen:
  Petra Stötzel, Tel. +49 228 885-2235, Petra.Stoetzel@dfg.de

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