Absichtserklärung
zur Ausschreibung der DFG - Implementierung der
OCR-D-Software zur Volltextdigitalisierung historischer Drucke

Die Bibliotheken

Staats- und Universitätsbibliothek Bremen	Universitätsbibliothek Johann Christian Senckenberg, Frankfurt	Universitäts- und Landesbibliothek Bonn

erklären die Absicht, sich mit dem Projekt *"OCR-D für Visual Library"* an der Ausschreibung der DFG "*Implementierung der OCR-D-Software zur Volltextdigitalisierung historischer Drucke"* zu beteiligen. Dazu ist ein Projekt mit einer Laufzeit von 24 Monaten ab dem zweiten Quartal 2021 vorgesehen. Eine Zusammenarbeit mit dem OCR-D-Koordinationsprojekt wird hier und auch zur Antragstellung bestätigt. Das hier beabsichtigte Projekt ist zwar auf die Visual Library zugeschnitten, jedoch sollen die Projektergebnisse grundsätzlich auch für andere national wie international eingesetzte Softwaresysteme nachgenutzt werden können, mit dem Ziel, die OCR-D Community weiter zu vergrößern.

Das Konsortium der antragstellenden Bibliotheken verfügt über langjährige Erfahrungen bei der Bearbeitung großer Digitalisierungsprojekte. OCR wird intensiv eingesetzt, teilweise wurden Projekte zur OCR-Nachkorrektur sowie Training und Evaluation mit Ground Truth-Texten durchgeführt. Digitalisiertes Quellenmaterial unterschiedlichster Ausprägungen (Textsammlungen von Büchern, Zeitschriften und historische Zeitungen) verschiedener Schwierigkeitsstufen (Antiqua, Fraktur, Schriftqualität, etc.) sind vorhanden.

Das Softwaresystem Visual Library (VL) wurde von der Firma *semantics Kommunikations-management* GmbH entwickelt. Die Digitalisierungsinfrastruktur Visual Library bildet den gesamten Workflow der Digitalisierung ab, inkl. OCR-Verarbeitung und wird kontinuierlich durch *semantics* mit Support und Weiterentwicklung gepflegt. Mit dem Fokus auf Bibliotheken mit VL können direkt oder indirekt (siehe unten die Konzeptstudie) allein in Deutschland 16 Bibliotheken erreicht werden (potenziell weitere Bibliotheken in Österreich und der Schweiz), die OCR einsetzen bzw. daran interessiert sind. Über die VL Anwendercommunity ist diese Gruppe von Bibliotheken vernetzt. Seit 2013 finden fast jährlich Anwendertreffen statt. Längerfristig können über einen intensiven Know-How Transfer (oder Services) auch kleinere Bibliotheken von der OCR-D-Software profitieren. Bei einer Implementierung von OCR-D in Mandantensysteme der Visual Library ist dieses Potenzial noch größer.

Ziel des Projektes ist die Entwicklung einer offenen OCR-Infrastruktur für Bibliotheken, die ihre Digitalisierungsaktivitäten auf der Software Visual Library der Firma *semantics* aufbauen. Aktuell kann innerhalb der VL nur Abbyy Finereader verwendet werden. Zur Flexibilisierung und Verbesserung der OCR-Ergebnisse soll im Rahmen des Projektes die VL in Form einer losen Kopplung an eine lokale OCR-D Instanz angebunden werden. Die lose Kopplung ermöglicht, die offene, flexible und skalierbare Systemarchitektur von OCR-D mit den bewährten Eigenschaften der VL zu verbinden. Im Rahmen des Projektes erfolgt eine Implementierung und Evaluierung bei den Partnerbibliotheken. Weitere Bibliotheken

sind eingeladen, sich als “assoziierte Partner” zu beteiligen. Für “interessierte Anwender” wird es Gelegenheiten für einen Know-How Transfer geben.

Vorarbeiten im Rahmen einer Pilotierung werden die folgenden Aktivitäten sein:

- Einarbeitung in die OCR-D-Software sowie Installation auf Compute Servern
- Einsatz der OCR-D-Software und Evaluationen im kleineren Rahmen
- Abschließende Klärung der technischen Anbindung von Prozessen und Dateiformaten an die Visual Library mit der Firma semantics
- Ausbau des Konsortiums durch Anbindung assoziierter Bibliotheken an das Projekt

Eine performante IT-Infrastruktur ist bei dem Einsatz der OCR-D-Software geboten. Ein Compute Server mit einer ausreichenden Anzahl an CPU Kernen, genügend Hauptspeicher und Festplattenkapazität wäre die minimale Anforderung. Ein weiterer Ausbau der IT-Infrastruktur hängt von den Erfahrungen zu Laufzeiten und Performance der einzelnen getesteten OCR-D- Workflows und von dem Ausmaß weiterer Digitalisierungsprojekte ab.

Die vom Konsortium betrachteten Anwendungsszenarien für die technische Umsetzung sind der Einsatz von Compute Servern, gegebenenfalls vorbereitet für eine Erweiterung der Leistungsfähigkeit durch Cloud-Computing. Inwiefern Webdienste die Systemarchitektur bzw. die Nutzerfreundlichkeit der Dienste gestalten helfen, ist abhängig von den einzelnen OCR-D-Workflows (Training, OCR, Evaluation) und damit Gegenstand des Projektes.

Datum 19. Mai 2020

Maria Elisabeth Müller	Daniela Poth	Dr. Ulrich Meyer-Doerpinghaus
SuUB Bremen	UB Frankfurt	ULB Bonn

Projektskizze

1. Projektziele
    a. Implementierung der OCR-D Software als offener Dienst in einem Konsortium von Bibliotheken, die die Digitalisierungssoftware Visual Library verwenden
        i. IT-technische Implementierung – siehe Arbeitspaket 1
        ii. Softwaretechnische Implementation – siehe Arbeitspaket 2
        iii. Workflow implementieren – siehe Arbeitspaket 3
    b. Evaluation und Dissemination – siehe Arbeitspakete 4 und 5
    c. Zusammenarbeit mit dem OCR-D-Koordinationsprojekt
2. Projektpartner
    a. SuUB Bremen, UB Frankfurt, ULB Bonn
    b. Weitere “assoziierte Bibliotheken” (mit oder ohne VL; siehe oben die Einladung)
3. Arbeitspakete
    a. Implementation
        i. AP1: Anwendungsszenarien für die technische Umsetzung
            1. Compute Server, SuUB
            2. Webdienste

3. Vorbereiten auf einen Cloud-basierten Ansatz

ii. AP2: Softwaretechnische Implementation
   1. Offene Schnittstellen für Dienste
   2. Anbindung an die Visual Library
   3. Intuitive Benutzerschnittstellen

iii. AP3: Umsetzung des Workflows (ggfs. iterativ)
   1. Ground Truth Text erstellen bzw. OCR-Volltexte nachkorrigieren; dazu Dienstleister-Angebote prüfen
   2. Modell [nach-]trainieren oder ein vorbereitetes Modell auswählen
   3. Bildvorverarbeitung
   4. Layouterkennung
   5. OCR Volltexte erstellen
   6. Import von Volltexten in die Visual Library

b. AP4: Evaluation
   i. Einsatz der OCR-D Software in größerem Rahmen
   ii. Erhebung und Dokumentation von Kennzahlen zu Funktionen, Stabilität, Laufzeit und Performance
   iii. Evaluation der Prozesse mit und ohne Anbindung von Dienstleistern

c. AP5: Dissemination
   i. Offenlegung der produzierten Quellcodes (Open Source) inklusive Dokumentation auf GitHub
   ii. Hausinterne Schulungen „Nutzung der OCR-D-Software“
   iii. Workshop I – Erfahrungsaustausch im Projektkonsortium
   iv. Workshop II – Erfahrungsaustausch mit weiteren interessierten Bibliotheken

4. Zeitplan / Meilensteinplan
   a. Laufzeit 24 Monate
   b. Konsortium Kick Off Meeting
   c. Meilenstein M1 (Monat 9): Die Implementation ist umgesetzt (APs 1, 2 und 3)
   d. Meilenstein M2 (Monat 21): Evaluation beendet (AP 4)
   e. Meilenstein M3 (Monat 24): Dissemination durchgeführt (AP 5)
   f. In der Zeit nach der Projektförderung Mittel für Rundgespräche beantragen; zur Identifikation und Moderation von Maßnahmen zum Community-Building
      i. Konzeptstudie (während oder nach der Projektlaufzeit): Erstellung eines Konzeptes zur Implementation der OCR-D Software in Visual Library-Mandantensystemen
         1. Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz)
         2. Gemeinsamer Bibliotheksverbund (GBV)
         3. Potenzial für eine Zusammenarbeit mit VL-Mandantensystemen in der Schweiz und Österreich
      ii. „lokale OCR-Kompetenzzentren bzw. OCR-Services für kleinere anfragende Bibliotheken“
5. Anhang *<zur Antragstellung im Oktober>*
   a. LOI – Firma Semantics
   b. LOIs von „assoziierten Projektpartnern“