Konzept Use Cases
(R 5.3.1)

Version 2015-03-27
Cluster 5
Verantwortlicher Partner Minf-BA, TUD, INFAI

DARIAH-DE
Aufbau von Forschungsinfrastrukturen für die e-Humanities

Dieses Forschungs- und Entwicklungsprojekt wird / wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF), Förderkennzeichen 01UG1110A bis N, gefördert und vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR) betreut.

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Projekt: DARIAH-DE: Aufbau von Forschungsinfrastrukturen für die e-Humanities
BMBF Förderkennzeichen: 01UG1110A bis N
Laufzeit: März 2011 bis Februar 2016

Dokumentstatus: Final
Verfügbarkeit: öffentlich
Autoren:

Frederik Baumgardt, INFAI
Tobias Gradl, MInf-BA
Nils Reimers, TUD

Revisionsverlauf:

Datum	Autor	Kommentare
27.10.14	Frederik Baumgardt	Arbeitsplan zu Use-Case 3
28.10.14	Nils Reimers	Use-Case 1: Narrative Techniken und Untergattungen im deutschen Roman
09.12.14	Tobias Gradl	Use-Case 2: Biographien
09.12.14	Frederik Baumgardt	Ergänzungen zum Arbeitsplan in Use-Case 3
04.03.15	Nils Reimers Frederik Baumgart Tobias Gradl	Einarbeitung der Kommentare von Nadja Grupe (SUB)
27.03.15	Tobias Gradl	Einarbeitung der Anmerkungen von Klaus Thoden (MPI WG)

Inhaltsverzeichnis:

1. Überblick über die Use-Cases

Cluster 5: Big Data beschäftigt sich[1] mit der Identifikation und Bewertung quantitativer Methoden (AP 5.2) mit Blick auf deren Verwendbarkeit im Rahmen geisteswissenschaftlicher Forschung, der Adaption und Umsetzung geeigneter Methoden im Rahmen fachwissenschaftlicher Dienste (AP 5.3), sowie der Vermittlung von Kompetenzen für den Umgang mit quantitativen Methoden (AP 5.4).

Um die zur Verfügung stehenden Ressourcen und Kompetenzen des Clusters möglichst optimal einzusetzen, werden alle Aufgaben der Arbeitspakete auf drei behandelte Use-Cases verteilt und zeitlich parallel behandelt:

- In *Use Case 1: Narrative Techniken und Untergattungen im deutschen Roman* wird mittels quantitativer Verfahren die Entwicklung narrativer Techniken – und in weiterer Folge auch die Entwicklung darauf aufbauender literarischer Kategorien – analysiert.

- *Use-Case 2: Biografien* dient der Entdeckung von Korrelationen zwischen Personen, Orten und Ereignissen in biographischen Texten und Daten. Geisteswissenschaftlich ist der Anwendungsfall motiviert aus dem am IEG angesiedelten Forschungsprojekt *Cosmobilities – Grenzüberschreitende Lebensläufe in den europäischen Nationalbiografien des 19. Jahrhunderts*[2].

- Im Rahmen von *Use-Case 3: Identifikation von griechischem und lateinischem Text in einer Sammlung von 2 Millionen Texten* basiert auf Digitalisaten des OpenMigne Projekts und hat die Erkennung intertextueller Korrelationen zum Ziel.

Die Use-Cases 2 und 3 werden aus organisatorischen Gründen zeitlich nicht vollständig parallel, sondern zeitversetzt betrachtet. Die Arbeiten am dritten Use-Case haben demnach später begonnen, so dass die konzeptuellen Überlegungen zu diesem Use-Case derzeit noch in einer Vorphase sind. Da der zweite Use-Case aber zwischenzeitlich mit größerer Intensität bearbeitet werden konnte, als in der ursprünglichen Planung vorgesehen, ergeben sich durch den versetzten Ablauf keine zusätzlichen Probleme und Risiken im Hinblick auf die Gesamtplanung.

Im Rahmen des vorliegenden Berichts erfolgt eine Beschreibung des konzeptionellen Fortschritts in den einzelnen, jeweils weitgehend autonom durchgeführten Use-Cases und weiterer Schritte. Eine weiterführende Detaillierung der Ziele und Motivation findet sich in *Report 5.2.1: Beschreibung der Use-Cases*.[3]

---

1 neben der Koordination des Clusters in AP 5.1

2 Vgl. http://www.ieg-mainz.de/Forschungsprojekte------_site.site..ls_dir._nav.17_f.69_likecms.html

3 https://dev2.dariah.eu/wiki/download/attachments/14651583/R%205.2.1%20%E2%80%93%20Beschreibung%20der%20Use%20Cases.pdf

2. Use-Case 1: Narrative Techniken und Untergattungen im deutschen Roman

Der Ausgangspunkt des ersten Use-Case ist eine Textsammlung von 2.000 deutschsprachigen Romanen aus dem Zeitraum 1500 bis 1930. Von diesen 2.000 Romanen entstammen 400 der digitalen Bibliothek von *TextGrid* und sind nach *TEI* kodiert. Die übrigen entstammen der *Gutenberg* Sammlung.

Die dabei entwickelten Verfahren sollen sprachunabhängig funktionieren. Um dies zu überprüfen werden die Verfahren ebenfalls auf eine *TEI*-kodierte Sammlung von 200 französischen Kriminalromanen des 19. und 20. Jahrhunderts angewandt.

2.1. Technische Grundlage

Der Austausch ist ein zentrales Konzept in der Wissenschaft. Der Austausch über Versuchsaufbaue und gewonnene Resultate verbreitet nicht nur gewonnenes Wissen, sondern erlaubt es anderen Wissenschaftlern, diese zur validieren oder zu verbessern. Dieses bewährte Grundprinzip möchte man ebenfalls nutzen für die Analyse von Textdaten mittels quantitativer Verfahren. Die genutzt und / oder entwickelten Analyseschritte sollten hierbei leicht zwischen Forschern ausgetauscht werden können. Ebenfalls sollten Ergebnisse reproduzierbar sein. Zum einen zur späteren Verifizierung dieser, zum anderen aber um darauf aufzubauen und verbesserte Analysemethoden entwickeln zu können. Die quantitative Analyse der narrativen Techniken wird im beschrieben Use Case computergestützt durchgeführt, d.h. mittels entsprechender Algorithmen werden die Textsammlungen verarbeitet, Informationen werden extrahiert und in Beziehung zueinander gesetzt. Einmal entwickelt lassen sich diese automatisierten Methoden leicht verteilen und die Ergebnisse können beliebig reproduziert werden. Neben der Reproduzierbarkeit sind aber vor allem die Anpassung der Analysemethode auf eigene Forschungsfragen und die Erweiterung der Methoden von zentralem Interesse. Anstatt eine fertige Software zu entwickeln, die nur eine fest definierte Analysemethode verwendet, haben wir auf die Bereitstellung einzelner Komponenten zum Ziel, die wie eine Art Baukasten verwendet werden können. Hierbei bildet das *Apache UIMA (Unstructured Information Management Architecture)* Framework den Rahmen für den Baukasten. Auf UIMA aufbauend stellt das von der TU Darmstadt entwickelte *DKPro (Darmstadt Knowledge Processing Software Repository)* viele essentielle Grundkomponenten der Textverarbeitung zur Verfügung. Dadurch können viele Analysemethoden Plug & Play-artig zusammengestellt und ausgeführt werden.

Das *Apache UIMA Project* ist ein Framework speziell zugeschnitten auf die Extraktion von Wissen aus unstrukturierten Daten. Dieses Framework ist Open Source und lizensiert unter der freien *Apache-Lizenz*. Diese erlaubt es, dass UIMA frei in jedem Umfeld frei verwendet, modifiziert und verteilt werden darf.

UIMA kann verwendet werden für Informationen die in beliebigen Formaten vorliegen, beispielsweise für Bilder, für Videosequenzen oder für Audio. Das meist verwendete Anwendungsszenario ist allerdings die Extraktion von Wissen aus natürlich-

sprachigen Texten. Dafür stellt das Framework eine große Anzahl an Schnittstellen zur Verfügung, die die einheitliche Erstellung von Analysekomponenten erlaubt.

UIMA verwendet dabei das Konzept von Pipelines. Daten beliebiger Formate werden zunächst eingelesen und in mehreren sukzessiven Schritten analysiert und weiterverarbeitet. Abschließend werden die Ergebnisse in einem gewünschten Format ausgegeben beziehungsweise abgespeichert. Abbildung 1 illustriert eine solche Pipeline zum Verarbeiten von Textdokumenten. Dieses Konzept der Modularisierung bildet die Grundlage für die Umsetzung des beschriebenen Use Case.

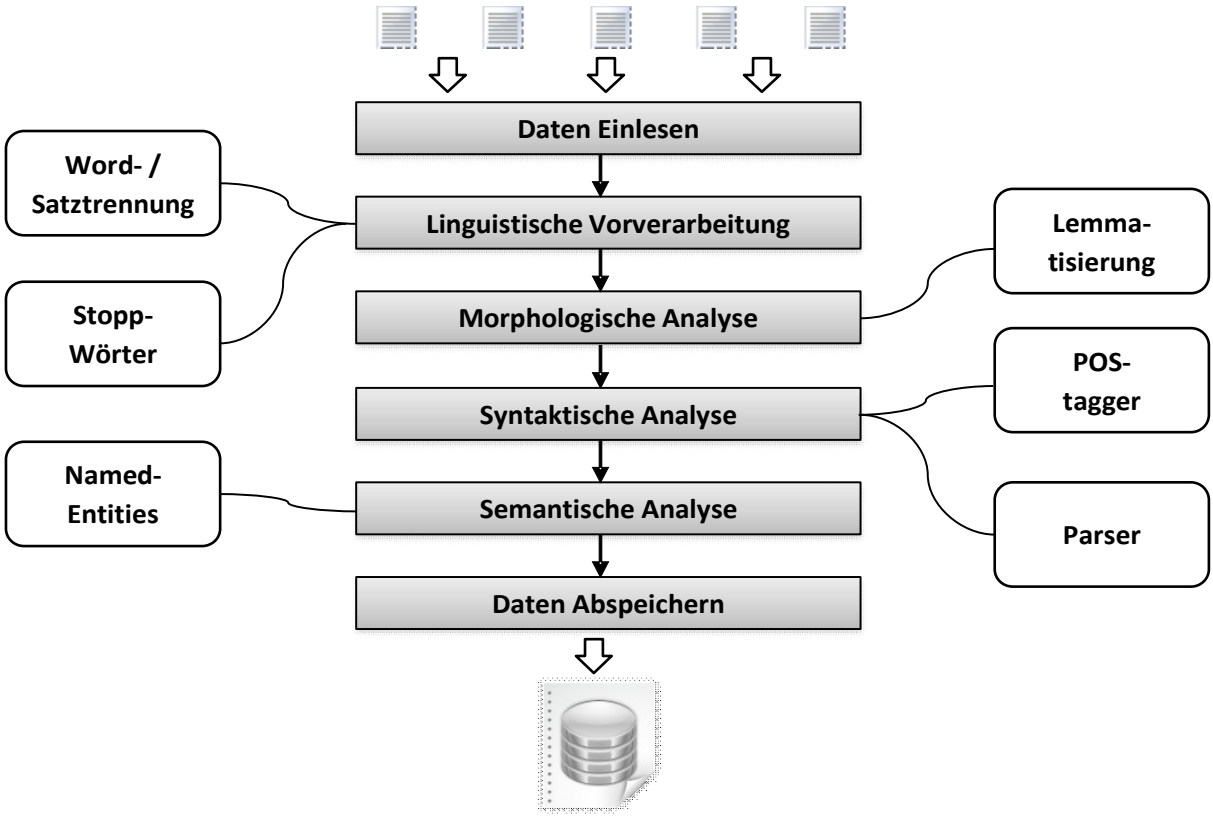

Abbildung 1: Illustrative Pipeline zum Verarbeiten von Textdokumenten

Die Modularisierung der Analyseschritte und das Zusammenschalten zu einer Pipeline ermöglicht es, die Methoden mit nur geringem Anpassungsaufwand für unterschiedlichste Szenarien zu verwenden. Möchte man anstelle von Textdateien XML-Dateien einlesen und deren Inhalt verarbeiten, so muss nur die Komponente zum Einlesen der Daten angepasst werden. Die weiteren Komponenten zur Durchführung der Analysen bleiben identisch. Dies erhöht ebenfalls die Kollaboration zwischen den Anwendern, da einzelne Komponenten oder ganze Pipelines zwischen Nutzern ausgetauscht werden können. Mit nur geringem Aufwand können diese ausgetauschten Komponenten oder Pipelines für die eigene Forschungsfrage genutzt werden. Dies ist insbesondere auch dann nützlich, wenn neuere, bessere Komponenten verfügbar gemacht werden, beispielsweise ein genauerer Named Entity Recognizer. Mit nur wenigen Zeilen Programmcode kann dieser in alle bestehenden Pipelines integriert werden.

Die Vorverarbeitungsschritte für viele Anwendungsszenarien sind oft sehr ähnlich, beispielsweise das Auffinden von Satzgrenzen, die Bestimmung von Wortarten oder das Identifizieren von Entitäten im Text. Für manche dieser Vorverarbeitungsschritte existieren spezialisierte Komponenten, wie Beispielsweise MaltParser[4]. Ebenfalls existieren entsprechende Tool Suites, wie Beispielsweise ClearNLP[5] oder OpenNLP[6], die eine große Anzahl an Verarbeitungsschritte anbieten. Jede dieser Komponenten ist mit einer eigenen Programmierschnittstelle (*Application Programming Interface, API*) versehen und einem eigenen Typensystem versehen. Dies macht es oftmals aufwendig, die verschiedenen Komponenten miteinander zu kombinieren oder gewisse Komponenten in einer Pipeline durch die einer anderen Tool Suite zu ersetzen. Keine der Tool Suiten kann alle gewünschten Funktionen abdecken und oftmals sind diese auch nur für bestimmte Sprachen verfügbar, beispielsweise nur für Englisch. Dies verhindert oftmals die Wiederverwendung einer geschriebenen Pipeline für neue Daten beziehungsweise für neue Versuche.

Ein weiteres Hindernis beim Austausch von entwickelten Pipelines ist zum einen die häufig fehlende eindeutige Identifizierung von Komponenten und zum anderen die notwendige manuelle Installation dieser in der korrekten Version. Dies macht den Austausch von Pipelines aufwendig. Ebenfalls sind Ergebnisse teilweise nicht mehr reproduzierbar, da die Information zu den ursprünglich genutzten Versionen fehlt.

Verarbeitungsschritt	**# Komponenten**	**Sprachen**
Sprachen-Identifizierung	2	de, en, es, fr, +65
Wort- und Satzgrenzbestimmung	5	de, en, es, fr, +25
Lemmatisierung	7	de, en
Wortstamm-Bestimmung	1	de, en, es, fr, +11
Wortart-Bestimmung	9	de, en, es, fr, +14
Morphologische Analyse	2	de, en, fr, it, +14
Named Entity Recognition	2	de, en, es, nl
Chunking	2	de, en
Constituency Parsing	3	de, en, fr, zh, +1
Dependency Parsing	5	de, en, fr, it, +7
Koreferenz-Analyse	1	en
Semantic Role Labeling	1	en
Rechtschreibüberprüfung	3	de, en, es, fr, +25

Abbildung 2: Übersicht über Analyse-Komponenten in DKPro Core (Eckart de Castilho und Gurevych, 2014)

Um diese Probleme zu beheben wird im Use Case das von der TU Darmstadt entwickelte *DKPro (Darmstadt Knowledge Processing Software Repository)* verwendet. DKPro integriert wie in Abbildung 2 dargestellt eine große Anzahl von Komponenten für grundlegende Verarbeitungsschritte im Bereich der Textverarbeitung. Über ein einheitliches Typsystem können ebenfalls Komponenten unterschiedlicher Tool Suiten, beispielsweise von ClearNLP und OpenNLP, ohne zusätzlichen Programmieraufwand gemeinsam verwendet werden. Dies ist insbesondere relevant für weniger unterstützte Sprachen. Die meisten Tool Suiten bieten eine hohe Abdeckung an

[4] www.maltparser.org
[5] www.clearnlp.com
[6] opennlp.apache.org

Komponenten für Englisch. Für Deutsch, Französisch oder gar Niederländisch stehen deutlich weniger Komponenten zur Verfügung. Um gewünschte Vorverarbeitungen zu erzielen ist man so dann oftmals gezwungen, Komponenten aus unterschiedlichen Suiten zu verwenden.

Ein weiterer Grund für den Einsatz von DKPro ist das zentrale Repository zur Speicherung der Komponenten. On-demand können entwickelte Pipelines alle notwendigen Ressourcen herunterladen und verwenden. Hiermit entfällt das aufwendige Installieren und Konfigurieren von Komponenten. Gerade der Wechsel zwischen verschiedenen Sprachen, beispielsweise von Englisch nach Deutsch, wird so vereinfacht. Sofern vorhanden, werden die notwendigen Ressourcen für die neu gewählte Sprache geladen und in der Pipeline verwendet. Um Ergebnisse selbst über Jahre reproduzierbar zu machen, verwendet DKPro eine strikte Versionierung aller Komponenten. Insgesamt ermöglicht dies, dass Pipelines ohne weiteren Konfigurationsaufwand ausgetauscht werden können und so einfach einem breiten Publikum zugänglich gemacht werden können. Ebenfalls lassen sich bestehende Pipelines einfach um weitere Komponenten erweitern beziehungsweise durch andere Komponenten austauschen, jeweils passend zum jeweiligen Versuchsaufbau.

DKPro ist in Java entwickelt und lässt sich sowohl in Java, in Groovy als auch in Jython einbinden. Aufgrund der einfachen und gut lesbaren Syntax fiel die Entscheidung für den Use Case auf Jython. Jython ist eine Java-Implementierung der Skriptsprache *Python* und lässt sich auf jeder Java-Plattform ausführen. Die Syntax von Jython ist identisch zur Syntax von Python, allerdings ermöglicht Jython es zusätzlich auf Java-Funktionalitäten zurückzugreifen. Die Python/Jython-Syntax betont Programm-lesbarkeit und viele Konzepte können in wenigen Zeilen Programmcode umgesetzt werden. Als Skriptsprache ist Jython insbesondere Vorteilhaft für die Entwicklung kleinerer Programme zur Generierung erster Ergebnisse (sogenanntes Prototyping). Diese Merkmale der Sprache führen zu einer flachen Lernkurve, welches insbesondere im Rahmen der *Digital Humanities* von großem Vorteil ist. Erste Ergebnisse können dadurch auch von Personen mit bisher nur wenig Hintergrundwissen in der Programmierung schnell erzielt werden. In Abbildung 3 ist der notwendige Code aufgeführt zum Einlesen und Segmentieren einer Textdatei und zur Bestimmung der Wortarten (Part-of-Speech).

```
# Assemble pipeline
pipeline = iteratePipeline(
  createReaderDescription(TextReader,
    TextReader.PARAM_PATH, "example.txt",
    TextReader.PARAM_LANGUAGE, "en",
     ),
  createEngineDescription(OpenNlpSegmenter),
  createEngineDescription(OpenNlpPosTagger));

# Run pipeline
for jcas in pipeline:
  for token in select(jcas, Token):
    print token.coveredText + " " + token.pos.posValue
```

Abbildung 3: Pipeline die zu jedem Wort in der spezifizierten Textdatei die Wortart (Part-of-Speech) ausgibt.

2.2. Technische Umsetzung: Interaktionen zwischen Charakteren

Wie in Report 5.2.1 schon näher beschrieben, sollen in diesem Use Case narrative Techniken in Romanen analysiert werden. Im Speziellen liegen der Fokus auf der Erkennung und Auswertung von deskriptiven und narrativen Passagen, der Erkennung möglicher Plotähnlichkeiten und der Analyse der Gattungen der Texte.

Als erste konkrete Forschungsfrage soll im Use Case die Interaktionen und die Beziehungen zwischen den Charakteren ausgewertet und entsprechend aufbereitet werden. Beispielsweise soll analysiert werden, ob ein Hauptcharakter existiert und wie die weiteren Charaktere zu diesem eingestellt sind. Eine grundlegende Technik ist hierfür das Erkennen von Charakteren in Romanen. Hierzu wird im ersten Schritt die Qualität von existenten Named Entity Recognizer evaluiert. Die meisten dieser Systeme wurden auf Nachrichtenartikel trainiert und liefern nur bedingt gute Ergebnisse zu Romanen. Um dieses objektiver beurteilen zu können, werden Namen in Romanen annotiert und die Performance der Systeme wird evaluiert. Darauf aufbauend wird entschieden, ob und wie bessere Modelle speziell für diese Art von Texten erzeugt werden können.

Neben der expliziten Nennung von Charakternamen existieren auch entsprechende, implizite Bezüge, beispielsweise mittels Pronomen (*„Er sagte…“*). Gerade in Dialogen finden sich eine hohe Anzahl dieser impliziten Bezüge. Um die Interaktionen zwischen den Charakteren analysieren zu können müssen deswegen diese impliziten Bezüge aufgelöst werden. Diese sogenannte *Coreference Resolution* soll im Use Case mittels *BART*[7] realisiert werden. Dieses System ist noch nicht in DKPro vorhanden und wird, sofern technisch möglich, entsprechend integriert.

Neben der Erkennung der Charaktere ist auch die Auswertung der Interaktionen zwischen diesen von Interesse. Mittels *Semantic Role Labeling* können die semantischen Argumente für die Prädikate eines Textes bestimmt werden. Beispielsweise für den Satz *„Max verkaufte das Buch an Tom“*, soll für das Prädikat *verkaufen* die involvierten Personen *Max,* der Verkäufer, und *Tom,* der Käufer, identifiziert werden sowie der verkaufte Gegenstand, ein *Buch*. Als Komponente zum *Semantic Role Labeling* wurde *mate-tools srl*[8] ausgewählt und wurde in der Version 1.7.0 in DKPro integriert.

Sobald *BART* und *mate-tools* in DKPro integriert sind, können die Charaktere und die Interaktionen zwischen diesen aus den Romanen extrahiert werden. Daraus lassen sich gewisse Informationen ableiten, zum Beispiel können so die Hauptcharaktere identifiziert werden.

Da die eingesetzten Systeme nicht fehlerfrei funktionieren und darüber hinaus nicht direkt für Romane konzipiert wurden, steht die Evaluation, ob und wie weit diese für diese Aufgabe verwendet werden können, noch aus.

---

7 www.bart-anaphora.org

8 https://code.google.com/p/mate-tools/

2.3. Technische Umsetzung: Gattungen der Texte

Der Gattungsbegriff steht in diesem Use Case stellvertretend für theoriegeleitete literarische Kategorien, die anhand quantitativer Methoden nachvollzogen werden können und dadurch möglicherweise in einem neuen, empirischen, Licht erscheinen. Für die Modellierung von Untergattungen werden im Rückgriff auf die vorangegangenen Arbeitsschritte in einem iterativen Prozess Feature-Vektoren aufgebaut, die es erlauben, Rückschlüsse darüber zu ziehen und zu beobachten, in welcher Form sich derartige Merkmale hin zu Strukturen auf einer gattungsstilistischen Meso-Ebene aggregieren.

Um Romane nach Gattungen gruppieren zu können, werden diese Feature-Vektoren in ein *Topic Model* eingebracht. Ein häufig verwendetes Topic Model ist dabei die *Latent Dirichlet Allocation (LDA)* von Blei et al. (2002). Beim LDA-Verfahren wird für jedes Dokument die *latent topics* identifiziert, d.h. es werden zugrundliegende Themen identifiziert, die das Dokument zu erklären versuchen. Zur Erzeugung des LDA-Models wird MALLET[9] verwendet, welches seit dem Versionsupdate 1.7.0 zur Verfügung steht.

2.4. Dissemination

Im Rahmen des Use Cases wird aufbauend auf UIMA, DKPro und Jython eine Sammlung von Best-Practices und Workflows für Verfahren der quantitativen Textanalyse entwickelt. In Form von didaktisch aufbereitetem Lehr- und Lernmaterial werden zuerst die grundlegenden Komponenten zur Textvorverarbeitung präsentiert. Dies sind beispielsweise das Part-of-Speech-Tagging, die Named Entity Recognition, das Dependency Parsing und das Semantic Role Labeling. Diese und weitere Komponenten werden in Form einer SVN-integrierten Wiki-Dokumentation[10] präsentiert und entsprechende Code-Beispiele sind aufgeführt. Über das angebotene Projektarchiv können die stets aktuellsten Code-Beispiele als auch Beispieltexte heruntergeladen und lokal getestet werden.

Durch das modulare Konzept von UIMA und DKPro können diese grundlegenden Komponenten einfach zusammengeschaltet werden, um entsprechende Pipelines für die jeweils zu untersuchende Forschungsfrage zu erstellen. Ebenso erlaubt dieses Konzept eine flache Lernkurve und somit sind vorab keine umfangreichen Programmierkenntnisse notwendig.

[9] http://mallet.cs.umass.edu/

[10] https://code.google.com/p/dkpro-tutorials/wiki/Jython_Tutorial

3. Use-Case 2: Biographien

Der Fokus der Konzeption und prototypischen Implementierung liegt im zweiten Use-Case primär auf Anforderungen, die aus der qualitativen, historischen Forschung um das *Cosmobilities* Projekt abgeleitet werden können. So soll mit Hilfe automatischer Methoden zur Analyse und Visualisierung von Daten die qualitative Forschung wie folgt unterstützt werden:

- Biographische Informationen aus unterschiedlichen Quellen sollen zu (potenziell) transnationalen Lebens- und Bewegungsprofilen historischer Personen zusammengeführt werden.

- Auf Basis dieser Profile sollen Eigenschaften und Regeln identifiziert werden können, welche als so genannte *Internationalitätskriterien* Rückschlüsse über die Wahrscheinlichkeit einer Mobilität korrelierter Personen erlauben.

Anwendungsfälle für die Analyse und Verarbeitung biographischer Daten mit Hilfe informatischer Methoden sind dabei nicht auf das *Cosmobilities* Projekt beschränkt. Durch erste weitere Anwendungsfälle kann bereits in diesem frühen Entwicklungsstadium a) die Menge der insgesamt verfügbaren biographischen Daten erhöht werden (ggf. multilingual) und b) eine Untersuchung der Übertragbarkeit des Konzepts auf artverwandte Fälle erfolgen. Neben der Umsetzung der Funktionalität zur Unterstützung des *Cosmobilities* Projekt (Fokus: Internationalität in Lebensläufen historischer Personen des 19. Jahrhunderts) besteht eine weitere Zielsetzung insbesondere in der Wiederverwendbarkeit der erzeugten Datenbasis für weitere Anwendungsfälle und Forschungsfragen, z. B. nach dem Migrationshintergrund von Personen (Internationalität der Vorfahren).

3.1. Konzeptionelle Grundüberlegungen

Vor dem Hintergrund der primären, aus *Cosmobilities* motivierten Anforderungen und der Abgrenzung von weiteren artverwandten Anwendungsfällen können zwei wesentliche Anforderungskategorien unterschieden werden:

- Die *Analyse biographischer Daten* umfasst die semi-automatische[11] Erkennung von biographischen Ereignissen (Korrelationen zwischen Personen, Orten und Ereignissen) zur Erstellung personenbezogener Bewegungsprofile.

- Die *Aggregation generierter Profile* zur Unterstützung spezifischer Fragestellungen—im konkreten Fall von Cosmobilities zur Ableitung von Internationalitätskriterien.

Die Analyse biographischer Daten kann dabei auf unterschiedliche Konzepte und technische Implementierungen z. B. im Kontext von *Named Entity Recognition* zurückgreifen und zudem als wiederverwendbares Paket auch für weitere Anwendungsfälle eingesetzt werden. Die Aggregation und Visualisierung der Profile wird

[11] *Semiautomatisch* insbesondere weil die manuelle Korrektur automatisch generierter Inhalte und eine Berücksichtigung in ggf. anschließenden automatischen Phasen möglich sein muss.

dagegen spezifisch für *Cosmobilities* umgesetzt und muss ggf. für ähnliche Anwendungsfälle individuell angepasst werden.

Person als zentrales Zugriffsobjekt

Die Analyse biographischer Daten führt zu Datensätzen, die die Korrelation zwischen Person, Ort, Zeit und Ereignis widerspiegeln. Auf einer logischen Ebene entsteht hierdurch eine graphartige Struktur, in der beliebig zwischen unterschiedlichen Objekttypen navigiert werden kann.

Obwohl diese Struktur problemlos als solche in einer Datenhaltung abgelegt und verwaltet werden könnte, stehen im konkreten Fall von *Cosmobilities* Personen und deren Attribute im Vordergrund der Analysen.

Biographische Daten werden aus diesem Grund im Rahmen personenbezogener, biographischer Profile zusammengeführt. Diese Profile können durch die Analyse weiterer Quellen oder manuelle Intervention erweitert und korrigiert werden. Technische Details wie Verarbeitungs- und Indexierungsstrategien können so auf die robuste und leistungsfähige Durchführung personenbezogener Analysen und die Bearbeitung der Profile hin optimiert werden. Weitere Objekttypen (Ort, Zeit, Ereignis) bleiben als solche erhalten, wodurch prinzipiell auch andere Zugriffe z. B. aus dem Benutzerinterface heraus möglich sind. Da der Index jedoch für die Auswertung von Biographien ausgehend von personenbezogenen Fragen optimiert ist, könnten derartige Zugriffe mit bedeutend längeren Wartezeiten einhergehen. Mit dem Aufkommen entsprechender Forschungsfragen, die z. B. eine ortsabhängige Abstraktion und Verdichtung von Daten erfordern, könnten jederzeit entsprechende Indexstrukturen eingerichtet werden.

Quantitative Analysen

Die Leistungsfähigkeit der Analysen und insbesondere der anfragespezifischen Auswertung von biographischen Profilen ist insbesondere dann relevant, wenn eine Abstraktion von der Ebene individueller biographischer Profile im Hinblick auf spezifische Fragestellungen erfolgen soll.

Transnationale Biographien historischer Personen werden im Rahmen von *Cosmobilities* aus zwei Zielsetzungen heraus erstellt:

- *Wirtschaft*, *Politik*, *Wissenschaft* und *Kunst* wurden als übergeordnete Gruppen mit einem hohen angenommenen Internationalitätsgrad identifiziert. Diese vier Gruppen sollen miteinander, sowie mit einer Kontrollgruppe verglichen werden können.

- Neben Berufen/Professionen wird die Existenz von Attributen (z. B. Geschlecht, Studium, Mitgliedschaften in Organisationen) angenommen, die in bestimmter Kombination als Indiz für die Internationalität von Lebensläufen dienen könnten. Mit Hilfe quantitativer Analysen sollen Vorschläge für derartige Attribute und Verbindungen erarbeitet und für eine qualitative Überprüfung bereitgestellt werden.

Neben der Gruppierung von Personen anhand ihrer Profession oder anderer Attribute und der Auswertung ihrer Biographien mit Blick auf Internationalität, können weitere Forschungsfragen identifiziert werden, die durch die erstellten Profile unterstützt werden können. Durch die Erfassung und Analyse von Verwandtschaftsverhältnissen kann eine Analyse z. B. auf die Vorfahren untersuchter Personen ausgeweitet und die erwähnte Migrationsanalyse durchgeführt werden. So wird derzeit in einer Bamberger Kollaboration mit der Professur für vergleichende Politikwissenschaft an der Analyse politischer Reden in Abhängigkeit des Migrationshintergrunds des jeweiligen Redners durchgeführt.

3.2. Lösungsansatz

Zwei wesentliche Grundbedingungen für die Unterstützung des *Cosmobilities* Projekts können im Hinblick auf die notwendige Datenbasis identifiziert werden:

- Der Einsatz quantitativer Methoden bedingt eine Datenbasis, die a) eine ausreichende Zahl von Datensätzen beinhaltet, um die Methoden anwenden und Ergebnisse überprüfen zu können und b) nicht allein durch den Einsatz qualitativer Arbeit vollständig erschlossen und ausgewertet werden kann.

- Da für den konkreten Einsatz im *Cosmobilities* Projekt biographische Daten insbesondere auch zu international tätigen Personen zusammengestellt werden, sollten Informationen aus unterschiedlichen Quellen kombiniert werden, um nationale Sichtweisen auf diese Personen auszugleichen.

Als Konsequenz wurde die Datenbank von Wikidata als erste Basis für die Analyse und Korrelation biographischer Daten identifiziert. Neben biographischen Daten, wie Geburts- und Sterbedaten, Berufsbezeichnungen oder nächsten Verwandten zeichnet sich Wikidata insbesondere auch durch beinhaltete Referenzen zu weiteren Datenbanken[12] aus: typische Verweise umfassen beispielsweise die Bezeichner von *Gemeinsamer Normdatei (GND)*, Virtual International Authority File (VIAF) oder des *Library of Congress Name Authority File (LCNAF)*.

Abbildung 4 zeigt die Mengenstruktur der im derzeitigen Prototypen verarbeiteten Wikidata Einträge. Etwa 1,8 Mio. Personeneinträge wurden hierbei verarbeitet, analysiert, mit Orten (ca. 2,7 Mio.) und weiteren Einträgen (z. B. Berufsbezeichnungen) korreliert und indexiert.

[12] Eine umfassende Untersuchung von *Normdaten in Wikidata* bietet das gleichnamige Handbuch, verfügbar unter http://hshdb.github.io/normdaten-in-wikidata/, zuletzt aufgerufen am 01.12.2014

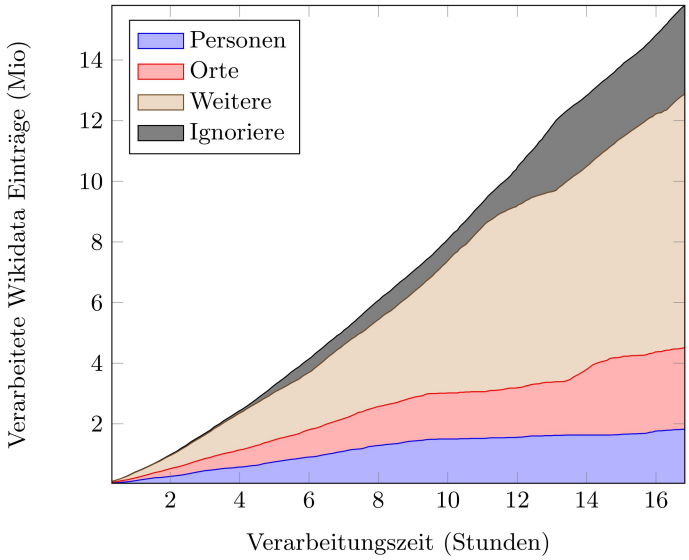

Abbildung 4: Parsing und Analyse der Wikidata im Rahmen des CosmoTool Prototypen

Wikidata als strukturierte Basis

Erste biographische Daten können somit direkt aus Wikidata abgeleitet werden: Neben Daten zu Geburt, Studium oder Tod einer analysierten Person werden dabei auch die biographischen Daten der direkten Verwandten als Indizien für den Zeitpunkt und Ort von Aufenthalten verwertet. So können Geburtsort und -datum eines leiblichen Kindes als sichere biographische Korrelation im Profil der Mutter, als wahrscheinliches Element im Profil des Vaters aufgenommen werden. Ebenso dient z. B. auch der Tod von Elternteilen als Indiz für den Aufenthaltsort des Kindes – in diesem Fall besteht jedoch eine direkte Abhängigkeit a) zur Lebensstufe des Kindes und b) der Epoche. So kann der Sterbeort und -zeitpunkt des Elternteils z. B. mit hoher Wahrscheinlichkeit dem Aufenthaltsorts eines Kindes mit einem Lebensalter von 2 Jahren zugeordnet werden. Mit zunehmendem Alter des Kindes nimmt diese Wahrscheinlichkeit ab, bis das Indiz schließlich als *zu unsicher* eingestuft werden muss und keine weitere Relevanz für das Profil des Kindes erhält.

Erste Heuristiken und Regeln wurden im aktuellen Prototypen bereits implementiert, die wissenschaftlich fundierte Analyse wird derzeit von Historikern am IEG vorgenommen – parallel zur weiteren Konzeption und Implementierung weiterer Analysewerkzeugen bzw. der Anbindung weiterer Quellen, darunter vor allem den europäischen Nationalbiografien.

Unstrukturierte Daten

Biographische Beschreibungen in Form von Volltexten werden im Rahmen des Use-Cases mit Hilfe existierender Werkzeuge aus dem Bereich des Natural Language Processing (NLP) analysiert. Die primäre Zielsetzung der Volltextanalysen besteht in der Ergänzung der aus strukturierten Daten gewonnenen biographischen Profile durch die Erkennung und Korrelation von Personen, Orten, Zeit und Ereignissen. Anhand a) bereits erkannter und ggf. durch qualitative Analysen bestätigter Daten und

b) formalisiertem Hintergrundwissen (z. B. durch Heuristiken oben) können die Erkenntnisse aus der Analyse von Volltexten auf ihre Plausibilität hin überprüft und bezüglich ihrer Sicherheit eingeordnet werden. Ein übergeordneter Rahmen wird so beispielsweise durch die Geburts- und Sterbedaten einer Person gebildet – biographische Einträge außerhalb dieses Zeitraums sind nicht möglich. Die so genannten Lebensstufen[13] teilen das menschliche Leben historischer Personen epochenabhängig ein. In der mittelalterlichen Vorstellung wurden so z. B. die sechs Stufen *infantia* (Säuglingszeitalter), *pueritia* (Kindheit), *adolescentia* (frühe Jugend), *iuventus* (spätere Jugend), *gravitas* (Höhe des Lebens) und *senectus* (Greisenzeitalter) unterschieden. Ein Mapping derartiger Stufen mit Ereigniskategorien, wie z. B. der Zuordnung des Ereignisses *Studium* mit der (späteren) Jugend eines Menschen erlaubt weiterführende Plausibilitäts- bzw. Wahrscheinlichkeitsbeurteilungen.

Zusammenfassend basiert der konzeptionelle Ansatz im Use-Case Biographien auf der iterativen Kombination von biographischen Erkenntnissen aus unterschiedlich strukturierten Quellen, um hieraus möglichst umfassende Lebens- und Bewegungsprofile für (historische) Personen ableiten zu können. Aufgrund der Fokussierung auf Internationalitätsaspekte im Projekt Cosmobilites kann die Zusammenführung verschiedener national geprägter Biographien zu einem integrierten, transnationalen Lebensprofil zu weiteren Erkenntnissen führen, die ansonsten nur im Rahmen manueller, qualitativer Betrachtungen möglich wären.

3.3. Technische Grundlagen

Verschiedene Eigenschaften und Rahmenbedingungen des Use-Cases führen zu der Nachnutzbarkeit und Weiterentwicklung bestehender DARIAH-DE Komponenten. Insbesondere die Heterogenität und Verteilung der potenziell relevanten Quellen biographischer Daten und die Notwendigkeit einer logischen Zusammenführung in einer integrierten Sicht auf Lebensprofile führen zu der Interpretation als Use-Case für die *Schema* und *Crosswalk Registry*[14] und das Konzept der forschungsorientierten Föderation von Daten.[15]

Mit Hilfe des im Rahmen der generischen Suche[16] von DARIAH-DE konzipierten und derzeit entwickelten *Transformationsframeworks* und der *Schema* und *Crosswalk Registry* kann die Föderation semi-strukturierter und strukturierter Daten bereits umgesetzt werden – Parsing, Transformation und Indexierung der Wikidata Daten (vgl. Abbildung 4) basieren bereits auf diesem Framework. Die Vorbereitung von unstrukturiertem Text (z. B. die generische Extraktion von Text aus MediaWiki oder HTML Markup), sowie die Ansteuerung analytischer und verarbeitender Tools (wie z. B. Komponenten aus dem Stanford NLP Portfolio[17]) aus dem Framework heraus, werden derzeit umgesetzt und evaluiert.

---

13 Vgl. https://dev2.dariah.eu/wiki/x/lQQMAg, zuletzt aufgerufen am 02.12.2014

14 Vgl. http://dev3.dariah.eu/schereg/, zuletzt aufgerufen am 02.12.2014

15 Vgl. http://dharchive.org/paper/DH2014/Paper-779.xml und http://goo.gl/i7i0pa, beide zuletzt aufgerufen am 02.12.2014

16 Vgl. http://search.de.dariah.eu, zuletzt aufgerufen am 02.12.2014

17 http://nlp.stanford.edu/software/index.shtml, zuletzt aufgerufen am 02.12.2014

3.4. Prototypische Umsetzung

Erste konzeptuelle Ideen konnten im Rahmen eines Prototypen[18] bereits umgesetzt werden. Der primäre Fokus bestand (und besteht auch kurz- bis mittelfristig weiterhin) in der Erkennung individueller persönlicher Profile, der Suche und Visualisierung von Profilen in Einzelansicht und der manuellen Korrigierbarkeit der Einträge. Durch den Verzicht auf Aggregations- und Gruppierungsfunktionalität wird sichergestellt, dass zunächst eine möglichst hohe Qualität bei der Erstellung und Zusammenfassung der einzelnen Profile erreicht werden kann. Basierend auf einer möglichst umfangreichen und qualitativ hochwertigen Basis biographischer Einzelprofile werden in folgenden Schritten weitere Aspekte im Rahmen des Prototypen umgesetzt, um schließlich den Anforderungen aus *Cosmobilities* und weiteren Use-Cases entsprechen zu können.

Der Bildschirmausschnitt in Abbildung 5 zeigt die einfache Suchoberfläche des Prototypen mit Eingabefeldern zur Spezifikation von Zeiträumen für Geburts- und Sterbedatum, Wikidata Occupations und einem freien Suchausdruck, welcher derzeit über der Menge der Personennamen ausgewertet wird. Im unteren Bereich des Bildschirmausschnitts werden passende Einträge in Form einer sortierten Ergebnisliste zurückgegeben.

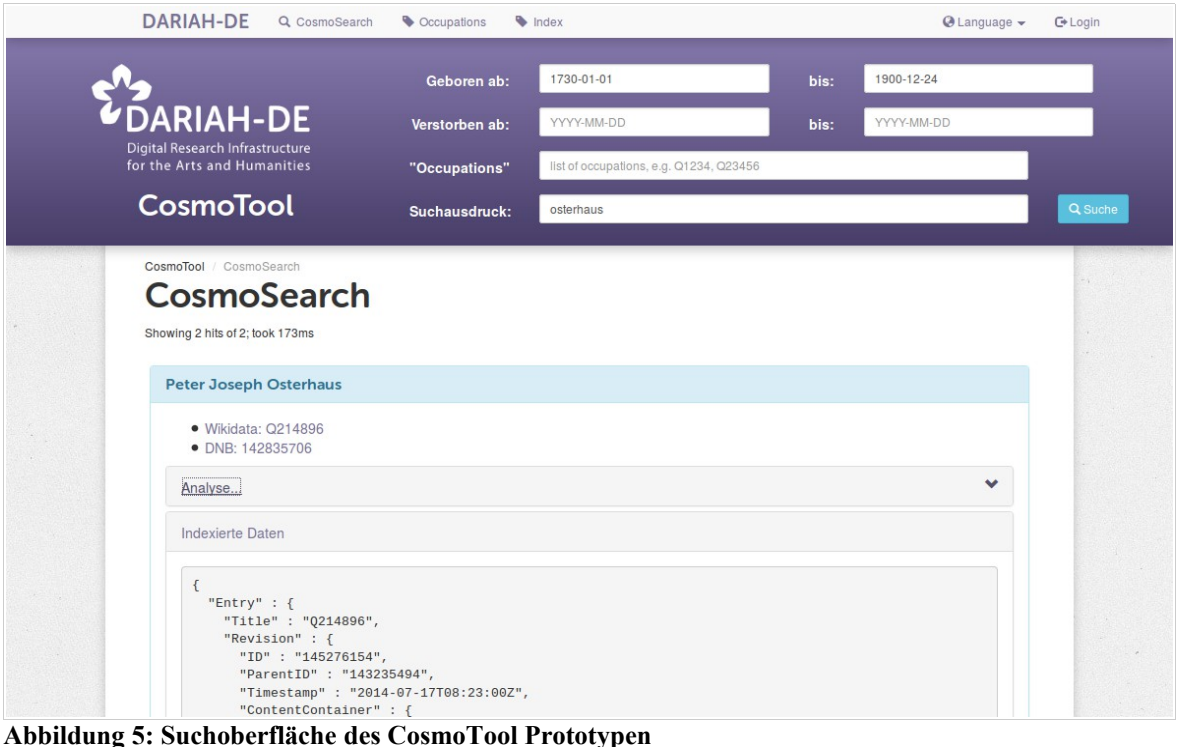

Abbildung 5: Suchoberfläche des CosmoTool Prototypen

Abbildung 6 zeigt ein Ergebnis im Detail und verdeutlicht die Generierung der Lebensprofile. Am Beispiel von Johann Wolfgang von Goethe können neben den direkten Attributen, wie Geburt, Studium und Tod auch Ereignisse direkter Verwandter in die Bildung des Profils einfließen.

[18] http://search.de.dariah.eu/cosmotool

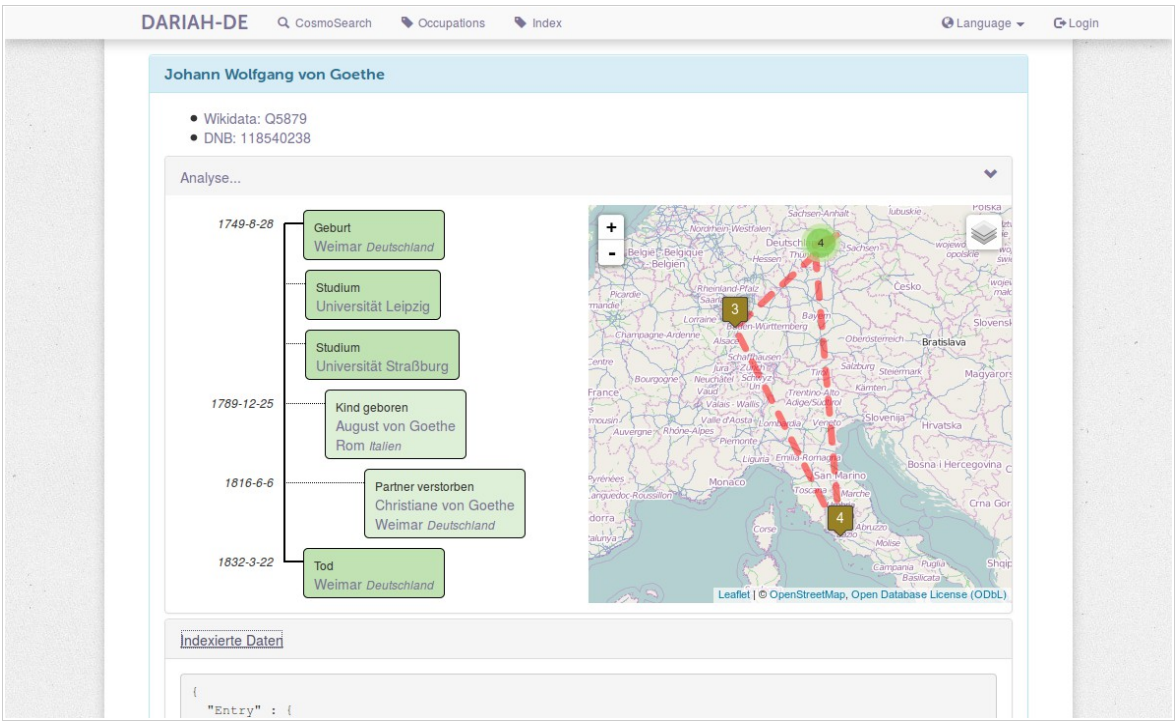

Abbildung 6: Profil von J. W. v. Goethe—gebildet aus Wikidata Einträgen

Abbildung 7 zeigt schließlich einen Ausschnitt des „Occupation-Baums“ - einer hierarchischen Anordnung von Berufen und Tätigkeiten zur weiteren Verwendung in Suchanfragen bzw. für erste Berufsgruppen-bezogene Analysen.

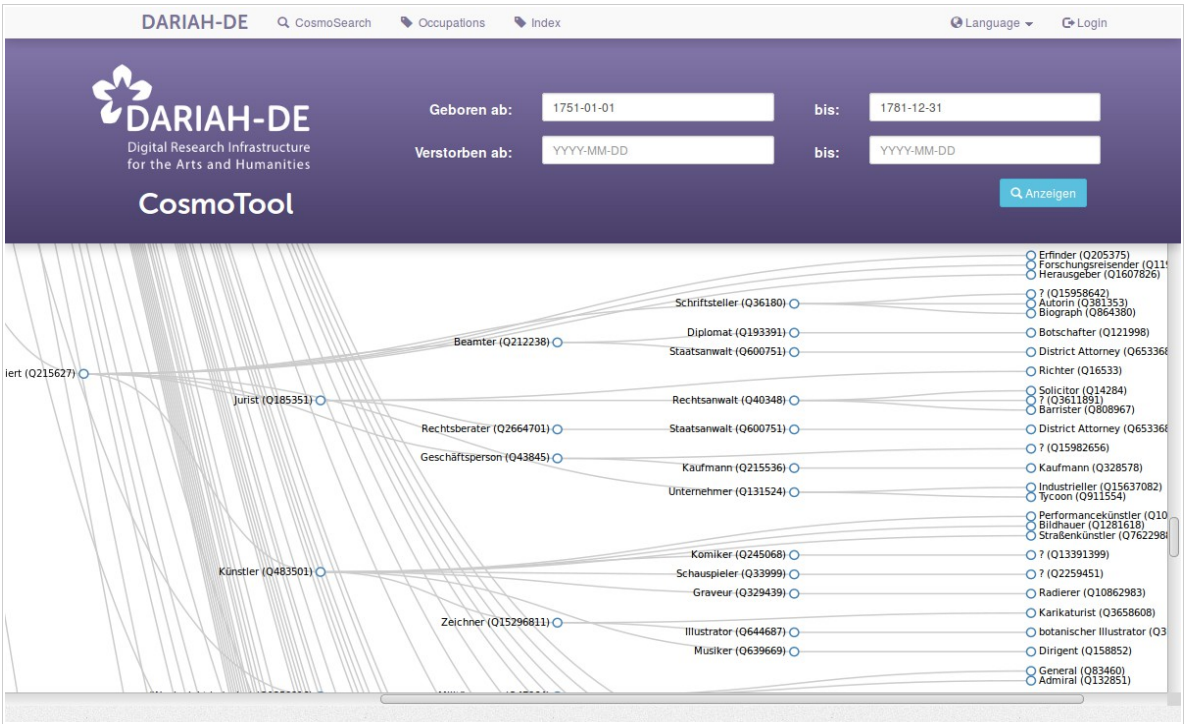

Abbildung 7: Übersicht der in Wikidata verwendeten Professionen (Occupations) als erste Basis für die Aggregation von Personengruppen

4. Use-Case 3: Identifikation von griechischem und lateinischem Text in einer Sammlung von 2 Millionen Texten

Der dritte Use-Case basiert auf den Texten der Patrologia Latina und Patrologia Graecae, die im Rahmen des OpenMigne Projekts am Humboldt-Lehrstuhl für Digital Humanities der Universität Leipzig digitalisiert werden. Die *Identifikation von griechischem und lateinischem Text* soll hierbei eine erste Erschließung der intertextuellen Struktur des Korpus leisten und mit Hilfe quantitativer Verfahren Zitationsgraphen der Dokumentensammlung errechnen, evaluieren und visualisieren.

4.1. Technische Umsetzung

1. Bereitstellung eines partiellen Korpus auf EpiDoc[19] / TEI P5[20] Grundlage mit Addressierbarkeit nach CITE-Architektur Die zugrundeliegenden Daten des OpenMigne Projekts sind derzeit noch nicht vollständig Standard-konform (es gibt ein Problem auf Seiten der Data Entry-Partner). D.h. die u.s. Methoden werden mit einer Teilmenge des Korpus entwickelt und anschliessend auf einen kompletten Korpus angewandt.

2. Konvertierung des Korpus in ein MongoDB-kompatibles und Roundtrip-fähiges Schema nach BSON-Spezifikation[21]. Strukturell entspricht das Korpus damit einem Directed Rooted Tree mit individuell addressierbaren Knoten, die Ebenen der CTS-Hierarchie[22] darstellen und Dokumentensammlungen, einzelnen Dokumenten oder TEI-Elementen entsprechen. Eine CTS-Query resultiert in einem Teilbaum.

3. Skizzierung eines vorläufigen Evaluierungsverfahrens unter Einsatz von manuell erzeugten und PAN[23] Ressourcen und Testen der Praxistauglichkeit von Signal/Noise-Ratio als Evaluationsmetrik. Durch differenzielles Testen von Precision und Recall soll die Kombination der 3 o.g. Elemente einen praktikablen Ersatz für umfangreiche und kostenintensive Goldstandards bieten.

4. Implementierung einer RESTful API, basierend auf CTS-API v5[24] in Scala / Play Framework 2 / Akka / ReactiveMongo (50%). Insbesondere CTS-API und Datenbankanbindung in Form von JSON/BSON/XML-Konvertierungen sind spezifische Entwicklungen für den Use Case.

5. Integration zweier Alignment-Verfahren mittels in Scala/Akka implementierten Wrappern und entsprechenden Adaptern für Datenkonvertierung in den Input-/Output-Layern.

---

19 http://www.stoa.org/epidoc/gl/latest/
20 http://www.tei-c.org/release/doc/tei-p5-doc/en/html/
21 http://bsonspec.org/
22 http://www.digitalhumanities.org/dhq/vol/3/1/000028/000028.html
23 http://pan.webis.de/
24 https://github.com/cite-architecture/cts_spec/blob/master/md/specification.md

6. Berechnung eines ersten Datensatzes von ca. 40.000 Alignments zwischen Texten der Patrologia Latina, des Corpus Scriptorum Ecclesiasticorum Latinorum und weiteren Texten des Open Greek And Latin Projekts der Universität Leipzig.

7. Entwicklung einer Web-Anwendung zum Alignment zweier eng verwandter Text, bspw. Editionen oder Ergebnisse verschiedener OCR-Verfahren. Durch die Anwendung können Korrekturen eines Textes mit den parallelen Fragmenten des anderen Textes durchgeführt werden.

4.2. Weitere Planung

Mit der bisherigen Arbeit wurde eine Grundlage zur effizienten Entwicklung und Anwendung von Text-Reuse-Verfahren geschaffen, sowie eine prototypische Anwendung auf den Ergebnissen. Bis Ende März wird die Arbeit in einer gemeinsamen Plattform für die Entwicklung und Evaluation unterschiedlicher Alignment-Algorithmen und Verfahren zur Erkennung von Text-Reuse gebündelt.

Hier aufbauend werden bis Jahresende Anwendungen inkl. ihrer zugrundeliegenden Methoden zur Text-Reuse-Erkennung entwickelt, die helfen sollen, Texte, Korpora und deren Autoren aus hermeneutischen und textkritischen Fragestellungen heraus zu beleuchten.