M 6.3.1 – Integration

Weiterentwickelte und integrierte Applikationen und Dienste

Version 5.4.2016
Cluster 6
Partner TUD, KIT, DT/PB, HKI, STI

DARIAH-DE
Construction of Research Infrastructures for the e-Humanities

This research and development project is / was funded by the German Federal Ministry of Education and Research (BMBF), fund number 01UG1110A to N, and managed by the Project Management Agency of the German Aerospace Center (Deutsches Zentrum für Luft- und Raumfahrt, PT-DLR).

SPONSORED BY THE

Federal Ministry of Education and Research

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

Dokumentstatus: Final
Verfügbarkeit: öffentlich
Autoren:

Nikolaos Beer (DT/PB)
Germaine Götzelmann (KIT)
Jochen Graf (HKI)
Rainer Stotzka (KIT)
Danah Tonne (KIT)

Versionen:

Date	Author	Comment
19.01.2016	Rainer Stotzka, Nikolaos Beer, Jochen Graf, Danah Tonne, Germaine Götzelmann	Erste Version zusammengestellt aus den Beiträgen der Partner
05.04.2016	Danah Tonne	Korrektur Zahlenwerte DBpedia Spotlight

Inhaltsverzeichnis

1. Einleitung

Im Arbeitspaket 6.3 „Annotationsdienste und Applikationen“ wurden Dienste und Werkzeuge zur Erschließung und Annotation von Forschungsdaten in die DARIAH-DE-Infrastruktur integriert und erweitert. Dies geschah in enger Zusammenarbeit mit dem Arbeitspaket 6.2 „Methoden“ und der AG „Service Lifecycle“, in denen methodisch wichtige Dienste und Demonstratoren ausgewählt wurden. Die Unterstützung und nahtlose Integration in die DARIAH-DE-Infrastruktur wurde durch Mentoren, aktive Weiterentwicklung sowie durch die enge Zusammenarbeit mit „Cluster 2 – eInfrastruktur“ durchgeführt.
DARIAH-DE wurde durch geeignete Methoden und Annotationswerkzeuge, die in AP 6.2 und in externen Projekten erarbeitet wurden, bereichert. Zielgruppenorientierte Visualisierung unterstützt die Auswertung und Präsentation komplexer Zusammenhänge annotierter Forschungsdaten während des gesamten Annotationsprozesses.
Erste konkrete Aufgaben waren die Erweiterung und die Integration des MEI Score Editors (MEISE) und des Werkzeugs Semantic Topological Notes (SemToNotes) in die DARIAH-DE-Infrastruktur.

2. Der MEI Score Editor

2.1. Überblick

Ziel der Fortführung der Entwicklungsarbeit am MEI Score Editor 2 (MEISE) in der zweiten Förderphase von DARIAH-DE war es, die basierend auf der Eclipse Rich Client Platform (Eclipse RCP) mit MEISE 1.1 entstandenen Editier- und Darstellungsmöglichkeiten von MEI-Daten zunächst in einen technologisch zeitgemäßen Rahmen (Webtechnologien) zu überführen und den Umgang und die Handhabung von MEI-Daten im Rahmen einer Editor-Software und auf Basis der in TextGrid und DARIAH-DE I gemachten Erfahrungen weiterzuentwickeln und auszubauen.

Durch den Technologiewechsel sollten darüber hinaus einfachere Anknüpfungspunkte an bestehende geisteswissenschaftliche Infrastrukturen (z.B. DARIAH-DE, TextGrid) und Initiativen (hier: Music Encoding Initiative, MEI) sowie Nachnutzungsmöglichkeiten in anderen musikwissenschaftlichen bzw. musikeditorischen Forschungs- und Toolentwicklungsprojekten geschaffen werden.

2.2. Entwicklungstätigkeit

Zum Start der zweiten Förderphase von DARIAH-DE lag mit MEISE 1.1 die bis dato einzige Editorsoftware vor, mit der sich MEI-Daten außerhalb des Kontexts reiner Texteditoren bearbeiten und darüber hinaus als Notentexte darstellen ließen.

2.2.1. Organisatorische Einbindung der MEI-Community

Im Vorfeld der Music Encoding Conference 2014 (MEC 2014) an der University of Virginia (19.–23.5.2014, Charlottesville/USA), auf der die Ergebnisse der Entwicklungsarbeit an MEISE 1.1 präsentiert wurden, zeigte sich, dass in unterschiedlichen Projektkontexten innerhalb der MEI Community an Softwarelösungen zur Bearbeitung und Darstellung von MEI-Daten gearbeitet wurde. Diese deckten jeweils aber nur spezifische Aufgaben ab und hatten keinen vollumfänglichen MEI-Editor zum Ziel. Vor dem Hintergrund der internen Evaluation und den dabei erkannten beiden zentralen Problempunkten der bisherigen MEISE-Entwicklung (technische Gebundenheit an das TextGridLab & hoher Aufwand zur Umsetzung einer eigenen GEF-basierten Notendarstellung) entstand die Idee, die Weiterentwicklung des MEI Score Editors enger mit der MEI Community zu verzahnen. Daraufhin wurde eine grundsätzliche Übereinkunft getroffen, mit Unterstützung von DARIAH-DE und unter der Federführung des Partners DT/PB die jeweiligen Toolentwicklungen zu koordinieren und untereinander abzustimmen. Dazu wurden in der Folge regelmäßige Entwicklertreffen vereinbart und ein eigener DARIAH-Wiki-Space zur Dokumentation eingerichtet.

2.2.2. Technische Entwicklung aus der Community

Bereits auf der MEC 2014 wurden mit *MEI toVexflow* und *Verovio* zwei neue Renderingbibliotheken vorgestellt, deren Darstellungsmöglichkeiten schon weit fortge-

schritten und von vorneherein auf Plattformunabhängigkeit bzw. auf die Verwendung in Webapplikationen ausgelegt waren.

Das *Distributed Digital Music Archives and Libraries Lab* (McGill University, Montreal) hat in unterschiedlichen Forschungsprojekten Tools entwickelt, die MEI-Daten unter verschiedensten Gesichtspunkten handhabbar machen. Von besonderem Interesse für die MEISE-Entwicklung waren hierunter:

- *meix.js* – Eine MEI/XML-Editor-Webapplikation mit MEI-Validierung.
- *vida.js* (Verovio Image Display Assistant) – Eine Webapplikation, mit deren Hilfe mit Verovio gerenderte MEI-Daten dargestellt werden können.
- *verovioEditor* – Eine Webapplikation die meix.js und vida.js zu einem einfachen MEI-Editor kombiniert und damit wechselseitige Bearbeitungsmöglichkeiten sowohl textbasiert als auch direkt in der Notendarstellung ermöglicht.

2.2.3. Entwicklung beim Partner DT/PB

Zu den Features, die sich in MEISE 1 als besonders hilfreich und daher erhaltenswert herausgestellt hatten, gehörte der sog. „Outline View“, der schnelle Navigation und Bearbeitung durch die Visualisierung von MEI-Files in einer Baumstruktur ermöglichte. Dazu wurde der MEITreeViewer entwickelt, der als Modul in MEISE 2 Einzug erhielt.

Eine weitere zentrale Funktion von MEISE ist die Möglichkeit, Kodierungsprinzipien aus musikeditorischen Kontexten (Varianten bzw. allg. Annotationen) verarbeiten zu können. Hier wurden die in Verovio schon vorhandenen Möglichkeiten zur Darstellung von kodierten Varianten über eine grafische Oberfläche zur Auswahl und Steuerung der Notendarstellung zugänglich gemacht.

2.3. Der DARIAH-DE MEI Score Editor

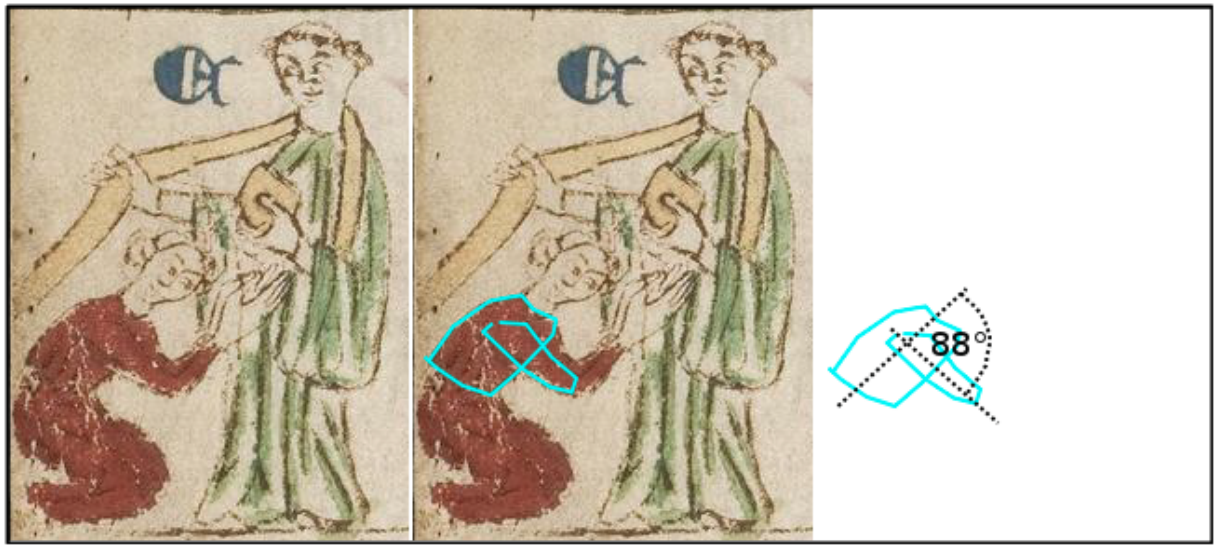
Der finale DARIAH-DE MEI Score Editor (MEISE 2) wird alle Teilentwicklungen aus der MEI Community als auch jene des Partners DT/PB zu einem in die DARIAH-DE-Infrastruktur integrierten Tool zusammenführen. Durch die Nutzung von Synergieeffekten über den europäischen Rahmen von DARIAH hinaus wird damit ein Tool zur Verfügung stehen, dem Interesse aus der gesamten MEI-Community entgegengebracht wird, so dass zu hoffen ist, dass hier insofern eine nachhaltige Lösung gefunden werden konnte als auch nach Ablauf der jetzigen Förderphase eine Weiterentwicklung in anderen Kontexten stattfinden dürfte.

3. Semantic Topological Notes (SemToNotes)

3.1. Einleitung

Im Rahmen des DARIAH-DE Clusters 6 Fachwissenschaftliche Annotationen steht Semantic Topological Notes (SemToNotes) für die methodische Diskussion über semantisch-topologische Annotationen und für die prototypische Realisierung eines semantisch-topologischen Bildannotationstools. Den Ausgangspunkt des SemToNotes Arbeitspakets bildet die kunstwissenschaftliche Methode der Ikonographie. SemToNotes möchte im Besonderen zeigen, dass sich Bildinhalte nicht nur mithilfe von textuellen Annotationen erklären und analysieren lassen, sondern auch durch graphische (nicht-textuelle) Annotationen.

Im DARIAH-II Antrag waren für die SemToNotes Arbeitspakete ursprünglich zwölf Personenmontate (50% / 2 Jahre) vorgesehen. Bedingt durch die Pensionierung von Herrn Professor Thaller (HKI) endete SemToNotes vier Monate früher, sodass insgesamt zehn Personenmonate zur Verfügung standen. Etwa sieben Personenmonate fielen dabei auf das Arbeitspaket 6.3, *Annotationsdienste und Applikationen*, in welchem der Prototyp des Annotationstools realisiert wurde; die restlichen drei Monate fielen auf das Arbeitspaket 6.4, *Vermittlung und Dissemination*, also auf die methodische Diskussion über semantisch-topologische Annotationen.



3.2. Entwicklung

Im AP 6.3 ging es um die prototypische Entwicklung eines topologischen Bildannotationstools, das die Koordinaten und textuellen Beschreibungen von Bildannotationen so abspeichern kann, dass sie mit einem topologischen Retrievalsystem analysiert werden können. Im DARIAH-II Antrag ist das technische Ziel in verkürzter Form wie folgt beschrieben:

- Nicht auf Rechtecke beschränkten Teilflächen eines Bildes, die ikonographisch beschreibbare Elemente enthalten, können inhaltliche Beschreibungen mittels eines Editors zugewiesen werden.

- Diese Teilflächen können durch einen graphischen und einen textuellen Editor so abgespeichert werden, dass sie robust sind gegenüber bildverarbeitenden Operationen (Zooming, Rotation).

- Die Beschreibungen bestehen aus Elementen, die in einer Ontologie verankert sind, somit semantische Queries erlauben. Eine Beschreibung kann auf n Teilflächen bezogen sein. Die Ontologien können innerhalb von SemToNotes ediert werden.

- Die Beschreibungen werden in einer nativen XML Datenbank so beschrieben, dass alle innerhalb der Ontologie ausdrückbaren semantischen Beziehungen retrievalfähig sind

- Im dadurch beschriebenen Retrievalsystem sind auch topologische Bedingungen möglich, die das Aufsuchen räumlich definierter Kombinationen semantischer Merkmale erlauben.

- Alle Operationen sind auf mobilen Geräten durchführbar.

Das SemToNotes Arbeitspaket hat zunächst den technischen Status Quo JavaScript basierter Annotationstools und topologischer Retrievalsysteme evaluiert.

Die Evaluation ergab, dass im Bereich der browser-basierten Annotationstools einerseits ausgereifte Pan-Zoom-Rotate Image Viewer zur Verfügung stehen; andererseits auch relativ ausgereifte Malwerkzeuge für graphische Formen. Jedoch war kein technischer Ansatz zu erkennen, der beide Anforderungen, Pan-Zoom-Rotate Image Viewer *plus* Polygone unterstützendes Malwerkzeug, miteinander verbindet. Deshalb lag ein Schwerpunkt bei der Entwicklung einer JavaScript API, die im Rahmen einiger Pilotprojekte (s. nächster Punkt) sukzessive erweitert und getestet wurde. Die JavaScript API stellt in technischer Hinsicht das Hauptergebnis dar.

Das topologische Retrievalsystem ist als XQuery Modul für die native XML Datenbank *eXist* konzipiert. Neben text-orientierten XML Queries mit der Datenbank Abfragesprache XQuery, lässt das Retrievalsystem auch topologische Queries zu, so wie sie im EXPath geo Standard definiert sind (http://expath.org/spec/geo).

Auch wenn das Modul stabil funktioniert, sind zum jetzigen Zeitpunkt noch viele inhaltliche Erweiterungen möglich. Z.B. ist fraglich, ob sich die für geographische Informationssysteme optimierte Semantik der EXPath Geo Query Sprache eins zu eins auf die Analyse von Handschriften anwenden lässt.

3.3. Kooperationen und Pilotprojekte

3.3.1. Digitale Schriftkunde

Projekt: http://www.gda.bayern.de/DigitaleSchriftkunde

Beispielseite:
http://www.gda.bayern.de/DigitaleSchriftkunde/1496_BayHStA_KU_Frauenchiemsee_88_03.html

Digitale Schriftkunde ist eine von der Generaldirektion der Staatlichen Archive Bayerns initiierte paläographische Übungspattform. Technisch realisiert wurde die Plattform in Kooperation zwischen TextGrid und DARIAH/SemToNotes. Die TEI-basierten Annotationen sind mit dem Text-Image-Linking Editor des TextGridLab gewonnen. Die Webpräsentation basiert auf SemToNotes und der nativen XML Datenbank eXist.

Zentrale Komponente der Plattform ist ein "Leuchttisch", bei dem u.a. das durch Bildannotationen angereicherte Digitalisat mit der Entzifferung in einer interaktiven Text-Bild-Link Ansicht dargestellt ist. Die Verknüpfung geschieht mit Rechtecken und Polygonen, teils zeilenweise teils wortweise. Die zu unterschiedlichen Schreiberhänden gehörenden Bild-Text-Annotationen können verschiedenfarbig visualisiert werden.

3.3.2. SADE Publish Tool (TextGrid)

Demo:
http://141.5.100.153/exist/apps/SADE/foobar/content.html?id=/xml/tile/24hcn.0.xml

SADE ist ein Tool zur Publikation von Editionsprojekten aus der TextGridLab Umgebung. In einem gemeinsamen Workshop zwischen TextGrid und SemToNotes Entwicklern in Göttingen wurden die Ergebnisse aus der Plattform Digitale Schrifkunde in das SADE Publish Tool integriert.

3.3.3. DWork - Heidelberger Digitalisierungsworkflow

Projekt: http://www.ub.uni-heidelberg.de/helios/digi/dwork.html

Beispielseite: http://archivum-laureshamense-digital.de/view/saw_mainz72/0005?sid=f8bb3a8768c9188c8260d3207e7cf8ff

DWork- Heidelberger Digitalisierungsworkflow der Universitätsbibliothek Heidelberg verwendet SemToNotes, anders als die oben genannten Projekte, nicht nur zur interaktiven *Visualisierung* von Text-Bild-Annotationen sondern auch als Annotationstool (Annotieren ist im eingeloggten Zustand möglich). Im Rahmen der Kooperation wurden auf Anregung der Entwickler der UB Heidelberg vor allem Usability Aspekte vorangetrieben. Zu den Ergebnissen gehören:

- Konstante Linienstärke von Bildannotationen bei Skalierung des Bildes
- Individuell konfigurierbare Hover Effekte, auch für überlappende Bildannotationen
- Bildviewer mit umfangreichen View-box Funktionalitäten

- Kompatibilität mit JavaScript Frameworks wie Angular oder jQuery
- Malen von graphischen Formen auch auf mobilen Endgeräten mit Touch Screen
- Dokumentation einer JavaScript API mit wohldefinitierten Funktionsnamen
- Performanzoptimierung für große Bilder, bis zu 3000px Höhe mal Breite

3.3.4. eCodicology

Demo: http://jochengraf.github.io/CodiLab/

In Kooperation mit dem im Arbeitspaket 6.2 angesiedelten Projekt eCodicology wurde das semi-automatische kodikologische Annotationstool *CodiLab* begonnen. CodiLab besteht aus drei Komponenten:

- *CodiKOS*, einem browser-basierten SKOS Editor für eine kodikologische "Web-Ontologie". CodiKOS realisiert den im Antragstext beschriebenen Editor für Ontologie basierte textuelle Beschreibungen.
- *CodiAnn*, einem Bildannotationstool zur semantischen Anreicherung von graphischen Annotationen. Hier sind, wie im DARIAH Antrag beschrieben, 1 zu n Verknüpfungen zwischen graphischer Form und ontologischen Beschreibungen möglich.
- *CodiQ*, einem topologischen Retrievalsystem zur Analyse/Kontrolle von automatischen Bildannotationen (nicht implementiert).

4. Erweiterung DBpedia Spotlight

4.1. Einleitung

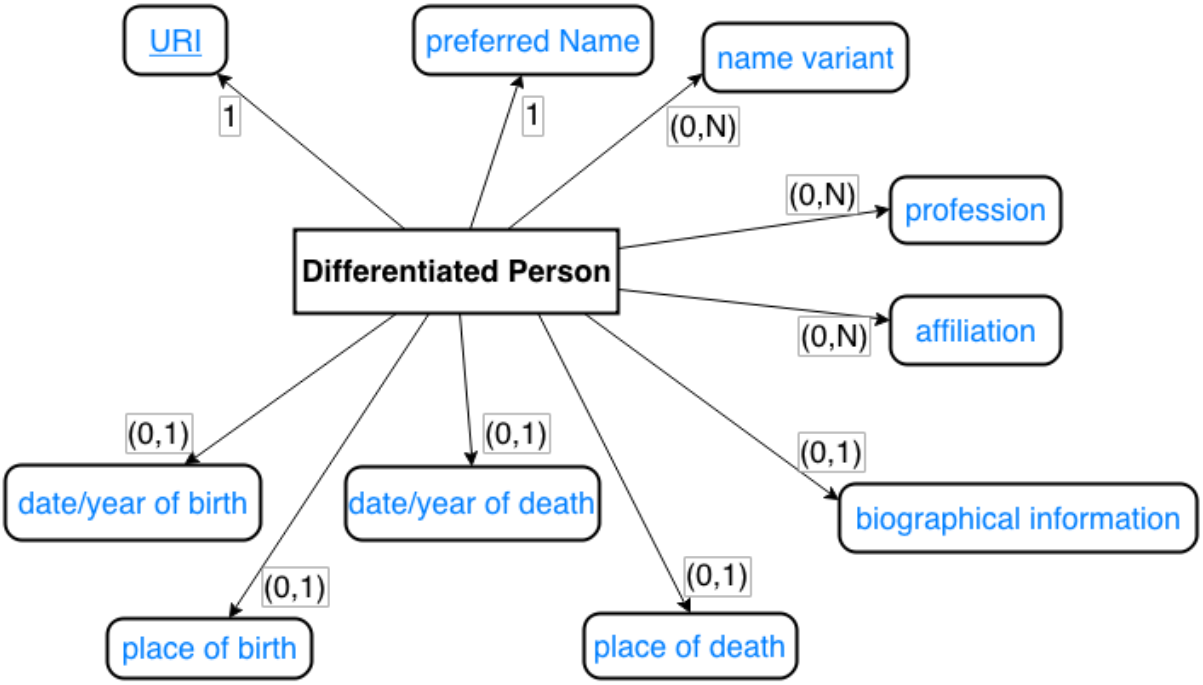
Ausgangsfrage des Projekts war es, inwieweit sich vorhandene Tools zur automatisierten Textannotation für Belange der Digital Humanities adaptieren lassen. Im Fokus standen dabei solche Werkzeuge, die Datenbestände der Linked Open Data Cloud nutzen, um unstrukturierten Text mit strukturierten Informationen zu versehen. Linked Open Data bietet den Vorteil, eine fast unerschöpfliche Quelle öffentlich und frei zugänglicher mit einheitlicher, mensch- und maschinenlesbarer Struktur (RDF) und hoher Verlinkungsdichte zur Verfügung zu stellen. Viele Tools konzentrieren sich jedoch auf den Datenhub Wikipedia/DBpedia, statt andere Datenquellen direkt auszuschöpfen. Mit DBpedia Spotlight adapiert das Projekt ein solches Annotationstool für Normdaten, genauer gesagt für sämtliche Personendatensätze (ca. 3 Millionen Datensätze) der Gemeinsamen Normdatei (GND).

4.2. Entwicklung

DBpedia Spotlight ist ein einfach zu nutzendes Werkzeug in fortgeschrittenem Entwicklungsstand. Es ermöglicht die Annotation und Datenanbindung für verschiedene Sprachen, unter anderem auch für Deutsch. Überdies bietet Spotlight die Möglichkeit, statt direkter Annotation die n-besten Kandidaten für eine im Text gefundene Entität als Liste auszugeben, sodass ein Experte für wissenschaftliche Annotationszwecke die finale Auswahl selbst treffen und damit die Ergebnisse für schwierige Unterscheidungsaufgaben gegenüber vollautomatischer Annotation erheblich verbessern kann.

Die von Spotlight zur Verlinkung von *named entities* im Text herangezogenen Daten aus DBpedia stellen gerade für historische Forschung eine lückenhafte Wissensbasis dar, da das Kriterium der Relevanz für die Enzyklopädie Wikipedia als Filter für die Aufnahme von Entitäten in den Datenbestand sorgt. Normdaten bieten hier gleichzeitig eine vollständigere Datenbasis und eine kontrolliertere Provenienz der Inhalte. Insgesamt sollen die Daten der GND jedoch nur als Beispiel gelten, eine Anbindung domänenspezifischerer Datenbestände ist gleichermaßen möglich, sofern diese per eindeutigem Identifier (URI) im Text verlinkbar sind.

Im Zentrum des Projekts steht die Datenmodellierung von GND-Daten für das Statistical Backend in DBpedia Spotlight. Aus den umfangreichen Namensvarianten der Normdaten werden sog. *surface forms* extrahiert, anhand derer die Entities im zu annotierenden Text identifiziert werden können. Die sonstigen biographischen Informationen werden gestemmed und nach Nennungshäufigkeit gewichtet als *context* herangezogen, der bei Namensgleichheit die Unterscheidung von Entities unterstützt. Darüber hinaus wird für den statistischen Ansatz ein *popularity*-Wert für jede Entity generiert. Die von der GND bereitgestellten DDC-Sachgruppen werden für einen Typfilter herangezogen, der es dem Nutzer ermöglicht, per black- oder whitelisting die Personendaten zur Annotation auf bestimmte Gesellschaftsbereiche/Berufsgruppen einzuschränken.



Mögliche Felder mit Informationseinträgen in der GND Personendatenbank

4.3. Ergebnisse

Bei Testannotation mit Text aus dem Bereich der deutsch-jüdischen Geschichte des 19. Jahrhunderts wurde mit dem neuen Datenmodel eine erhebliche Verbesserung des entity linking gegenüber der Originalimplementierung von DBpedia Spotlight erreicht (Steigerung des F1-Werts gegenüber Original mit Lucene-Backend 28%, mit Statistical Backend und Typfilter 45%). Die Verbesserung des Ergebnisses ist dabei auf eine höhere Anzahl korrekter Annotationen bei gleichzeitig steigender Zahl überhaupt in der Datenbasis vorhandener Personendaten (mit dem Ergebnis eines deutlich verbesserten recalls) zurückzuführen.

System	correctly linked entity mentions	linked overall	entities in text	detectable entities	precision	recall de-tectable	F1
Original Statistical	3	14	43	23	21,43%	13,04%	16,22%
GND Lucene	8	48	43	30	16,67%	26,67%	20,51%
GND Statis-tical	10	122	43	36	8,20%	27,78%	12,66%
GND Statis-tical with types filter	8	46	43	27	17,39%	29,63%	21,92%