Visualsierung von semantisch annotierten Forschungsdaten

Version 27.02.2019
Cluster 6
Verantwortlicher Partner TUDa

DARIAH-DE

Überführung der digitalen Forschungsinfrastrukturen für die e-Humanities in die Operational Phase (Betriebsphase)

Dieses Forschungs- und Entwicklungsprojekt wird / wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF), Förderkennzeichen 01UG1610A bis J, gefördert und vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR) betreut.

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Projekt: DARIAH-DE: Überführung der digitalen Forschungsinfrastrukturen für die e-Humanities in die Operational Phase (Betriebsphase)
BMBF Förderkennzeichen: 01UG1610A bis J
Laufzeit: März 2016 bis Februar 2019

Dokumentstatus: Entwurf
Verfügbarkeit: öffentlich
Autoren:

Max Grüntgens, ADWLM
Thomas Kollatz, STI
Danah Tonne, KIT

Revisionsverlauf:

Datum	Autor	Kommentare
10.01.2019	Canan Hastik	Erster Textentwurf
19.02.2019	Danah Tonne	Textbeitrag zu Use Case 2
26.02.2019	Thomas Kollatz und Max Grüntgens	Textbeitrag zu Use Case 1
28.02.2019	Oliver Schmid	Korrekturen
28.2.2019	Canan Hastik	Einarbeitung Feedback des Konsortium

Inhaltsverzeichnis:

1. Einleitung

Ziel des Arbeitspaketes 6.1 *Analyse und Visualisierung annotierter Forschungsdaten* in Cluster 6 *Annotieren, Analysieren, Visualisieren* war es, die digitalen Methoden der fachwissenschaftlichen semantischen Annotation und Visualisierung zu fokussieren und einen Beitrag zur Förderung der Transparenz des gesamten Annotationsprozesses zu leisten. Hierfür wurden zwei Anwendungsfälle definiert, die die Generierung standardisierter Annotationen und ihre Verwertbarkeit als Linked Open Data adressieren. Die in den beiden Anwendungsfällen *Analysemethoden und Visualisierungstechniken für semantisch annotierte Forschungsdaten* und *Informationsvisualisierung hochdimensionaler Annotationsdaten* entwickelten Strategien und auf Basis vorhandener Repositorien implementierten generischen Lösungen wurden ferner um einen dritten Anwendungsfall *Visualisierung von Forschungsaktivität* ergänzt.

Der vorliegende Report gliedert sich daher in zwei Bereiche, die sich auf die beiden ersten Anwendungsfälle beziehen und wird um den weiteren Report 6.1.3 ergänzt, der den Forschungsprozess aus einer Benutzerinteraktionssicht erörtert.

2. Analysemethoden und Visualisierungstechniken von semantisch annotierte Forschungsdaten

Der folgende Teilbericht stellt die Arbeiten und Ergebnisse des Anwendungsfalls *Analysemethoden und Visualisierungstechniken für semantisch annotierte Forschungsdaten* dar. Ausgehend von den TEI-konformen Forschungsdatenrepositorien des Salomon Ludwig Steinheim Instituts Essen (STI) und der Akademie der Wissenschaften und der Literatur Mainz (ADWLM) werden Verfahren erprobt, die eine Anreicherung und Transformation der bestehenden Forschungsdaten als Linked Open Data ermöglichen und auswertbar machen.

2.1. Das Projektvorhaben

In diesem ersten Anwendungsfall wurden exemplarisch die auf dem editorischen Markup-Standard EpiDoc – TEI-XML for Epigraphic Sources basierenden, bzw. in diesem Standard bereitgestellten digitalen Inschriftenedition zu Worms aus den epigraphischen Repositorien epidat[1] (jüdische Grabsteinepigraphik) und DIO[2] (Deutsche Inschriften Online) mit insge-

[1] epidat www.steinheim-institut.de/cgi-bin/epidat

samt über 50.000 Inschriften miteinander vernetzt, als Linked Data aufbereitet und bereitgestellt. Zielsetzung dieses Vorhabens ist es, repositoriumsübergreifende Forschungsansätze zu ermöglichen. Hierfür war es zunächst notwendig, die Datenbestände sowie mögliche gemeinsame Fragestellungen zu formalisieren und zu harvesten, d.h. mittels des XTriples-Webservices[3] in einer prototypischen Weise mögliche Fragestellungen und Fragenkomplexe zu formulieren und diese in ein RDF-Serialisierungsformat zu überführen. Parallel wurden die Daten auf eine CIDOC CRM basierende Domänenontologie gemapped und visualisiert.[4] Durch das beispielhafte Durchexerzieren eines flachen und prototypischen Workflows und das parallele Mapping auf eine etablierte Ontologie wurden beide Forschungsdatenbestände präzisiert, Fragestellungen formalisiert, aber auch weitere Desiderate der Datenbasis aufgezeigt. Beide Repositorien wurden damit in höherem Maße interoperabel gemacht und somit ein Konzept entwickelt, diese übergreifend zu explorieren und zu visualisieren.

2.2. Die Repositorien

Epidat[5] wird seit 2002 am Steinheim-Institut für deutsch-jüdische Geschichte (Essen) entwickelt. Die Datenbank ging 2006 online und konnte seitdem kontinuierlich erweitert werden. Die Bestände sind über die Webseite des Forschungsportals zur jüdischen Grabsteinepigraphik (epidat) frei zugänglich. Topographisch stammen die Inschriften überwiegend aus dem deutschsprachigen Raum, aber es sind auch historische Friedhöfe der Niederlande, Litauens und Tschechiens vertreten. Zeitlich umspannt das Inschriftenkorpus einen Zeitraum, der vom 11. bis ins 20. Jh. reicht. Jeder Datensatz wird unter einer Creative Commons-Lizenz als Open Access-Publikation veröffentlicht. Derzeit enthält die Datenbank 200 historische jüdische Friedhöfe mit mehr als 35.000 Inschriften und etwa 65.000 digitalen Abbildungen. Epidat enthält nahezu ausschließlich Grabinschriften.

Das Projekt Deutsche Inschriften Online (DIO)[6] stellt seit 2009 eine wachsende Zahl von Bänden der Reihe *Die Deutsche Inschriften* (DI) sind ebenfalls online frei zugänglich. DIO ist ein interakademisches Langzeitforschungsvorhaben. Das Sammlungsgebiet ist beschränkt auf das heutige Deutschland, Österreich und Südtirol. Die Bestände der Reihe umfassen Material aus der Zeit um 500 bis etwa 1650 n. Chr. Das Material ist vom sprachlichen, thematischen wie vom kunsthistorischen Standpunkt her heterogener als dies bei epidat der Fall

---

[2] DIO www.inschriften.net
[3] Siehe zum XTriples-Webservice Grüntgens, Schrade (2016).
[4] Zum Zusammenspiel von Epigraphik und CIDOC-CRM, siehe auch Felicetti, Murano, Ronzino, Niccolucci (2015).
[5] Epidat http://www.steinheim-institut.de/cgi-bin/epidat – Siehe zu den Repositorien auch Grüntgens, Kollatz (2018).
[6] Deutsche Inschriften Online http://www.inschriften.net

ist. Der prozentuale Anteil an Funeralinschriften reicht je nach Bestand von ca. 40-80% und fluktuiert aufgrund von Besonderheiten des Sammlungsgebietes, des Überlieferungszufalls und der kulturhistorischen Eigenheiten der Region.

2.3. Die Analysemethoden

Epidat und DIO stellen ihre epigraphischen Forschungsdaten unter anderem auch im dokumentierten Format *EpiDoc: Epigraphic Documents in TEI XML*[7] über Schnittstellen zur Verfügung:

> EpiDoc is an international, collaborative effort that provides guidelines and tools for encoding scholarly and educational editions of ancient documents. It uses a subset of the Text Encoding Initiative's standard for the representation of texts in digital form, which focuses on the history and materiality of the texts[8].

Dieses gemeinsame Format konnte als Ausgangspunkt für die Transformation von TEI-XML zu RDF herangezogen werden. In einem ersten Schritt wurden die für eine RDF-Serialisierung maßgeblichen Schnittstellen zwischen beiden Forschungsrepositorien identifiziert: Lokalisierung, zeitliche Einordnung, Angaben zum Objekttyp und den verwendeten Materialien, Aussagen über Personen, ihre Rollen und Beziehungen. In dieser Phase wurde deutlich, dass das Datenmodell beider Repositorien zwar in sich durchaus schlüssig ist, im Hinblick auf die repositorienübergreifende Analyse jedoch erweitert und in ihrer Modellierung aufeinander abgestimmt werden mussten. Ferner wurden durch den kontrastierenden Vergleich beider Bestände Desiderate sichtbar, die im weiteren Verlauf der Editionsprojekte vertiefend erörtert werden.

Was im Datenmodell jeweils nicht ausgedrückt wurde, waren Aspekte, die repositoriumsintern selbstverständlich waren. Da epidat nur den Objekttyp “Grabmal” behandelt, wurde diese Information bislang nicht explizit angegeben; ebenso wurde in DIO das verwendete Material nur dann angegeben, wenn es sich dabei nicht um Sandstein handelte. In dieser Phase wurde dieses intern als bekannt vorausgesetzte implizite Wissen explizit in das Austauschformat EpiDoc TEI XML aufgenommen und somit das Datenmodell präzisiert. Mit dem generischen XTriples Webservice wurden anschließend RDF-Statements konfiguriert und in einen Triple Store geladen. XTriples ermöglicht in einfacher Weise mittels XPath und XQuery RDF-Statements (Abb. 1) basierend auf Aussagemustern aus beliebigen XML-Formaten oder Schemata zu extrahieren:

[7] EpiDoc Guidelines 9.0 http://www.stoa.org/epidoc/gl/latest/
[8] EpiDoc About https://sourceforge.net/p/epidoc/wiki/About/

```
<xtriples>
  <configuration>
      <vocabularies>
          <vocabulary prefix="dio" uri="http://nbn-resolving.de/"/>
          <vocabulary prefix="epi" uri="http://www.steinheim-institut.de/cgi-bin/epidat?id="/>
          <vocabulary prefix="rdf" uri="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>
          <vocabulary prefix="rdfs" uri="http://www.w3.org/2000/01/rdf-schema#"/>
          <vocabulary prefix="foaf" uri="http://xmlns.com/foaf/0.1/"/>
          <vocabulary prefix="rel" uri="http://purl.org/vocab/relationship/"/>
          <vocabulary prefix="mat" uri="http://www.inschriften.de/material/"/>
          <vocabulary prefix="obj" uri="http://www.inschriften.de/objecttype/"/>
          <vocabulary prefix="con" uri="http://www.inschriften.de/connection/"/>
      </vocabularies>
      <triples>
          <statement> [4 lines]
          <statement> [4 lines]
          <statement>
              <subject prefix="dio">//publicationStmt/idno</subject>
              <predicate prefix="foaf">gender</predicate>
              <object prefix="con" type="literal">//listPerson/person/@sex</object>
          </statement>
          <statement> [4 lines]
      </triples>
  </configuration>
  <collection uri="http://www.steinheim-institut.de/daten/DARIAH-GT/di29_objectType_material.xml">
      <resource uri="{//TEI}"/>
  </collection>
</xtriples>
```

Abb. 1: Konfiguration für die Extraktion der Genderattribution

> With the XTriples webservice you can crawl XML repositories and extract RDF statements using a simple configuration based on XPATH/XQuery expressions. The webservice can be used with direct POST, form-style POST or GET requests.[9]

Damit wird es möglich, repositoriumsübergreifende Forschungsfragen zu formulieren, etwa nach der jeweiligen Verteilung der Materialien und Objekttypen, oder der Genderdistribution in epidat und DIO (Abb. 2).

```
# SPARQL-Request GENDERDISTRIBUTION in DIO and EpiDat
PREFIX dio: <http://nbn-resolving.de/>
PREFIX epi: <http://www.steinheim-institut.de/cgi-bin/epidat/>
PREFIX mat: <http://www.inschriften.de/material/>
PREFIX obj: <http://www.inschriften.de/objecttype/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?gender (COUNT(?gender) as ?counter) ?corpus
WHERE {
        ?s foaf:gender ?gender .
 ?s rdf:type ?corpus . FILTER ( ?corpus = "EPI" || ?corpus = "DIO")

}
GROUP BY ?gender ?corpus
```

Abb. 2: Repositorienübergreifende Sparql Query nach Genderdistribution in DIO und Epidat

[9] XTriples http://xtriples.spatialhumanities.de

Diese Anfragen können weiter in eine RDF-Serialisierung überführt und die Abfrageergebnisse in visualisierungsfreundlichen Datenformaten (CSV, JSON) exportiert werden, um sie mit Webservices, wie beispielsweise rawgraph.io, zu visualisieren. Die Frage etwa, ob es einen signifikanten Unterschied hinsichtlich der Genderdistribution zwischen der jüdischen (epidat) und christlichen (DIO) Inschriftenüberlieferung gibt, lässt sich mithilfe der Visualisierung (Abb. 3) eindeutig beantworten:

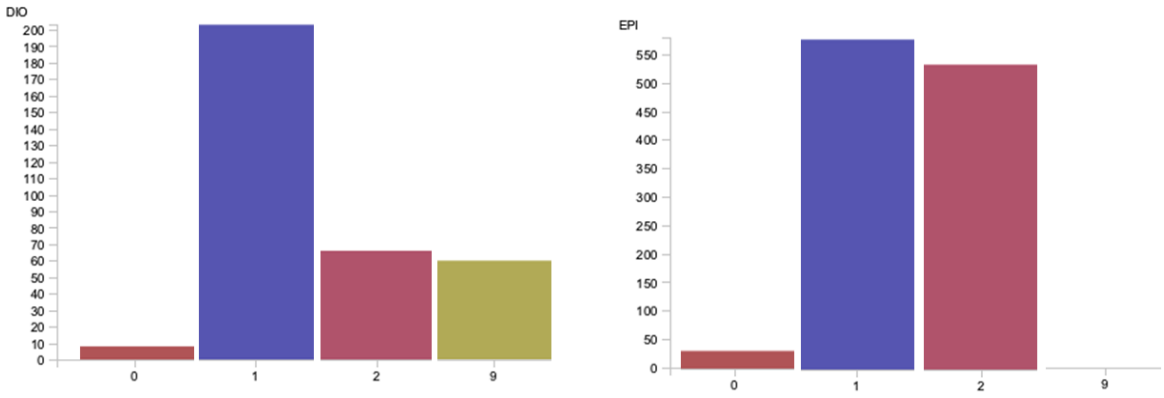

Abb. 3: Visualisierung der Genderdistribution in DIO (links) und epidat (rechts).
1=male; 2=female; 0=unknown (e.g. auf Grund starker Verwitterung ist das Geschlecht der bestatteten Person nicht feststellbar); 9=not defined (e.g. es handelt sich um ein Objekt, das in der Edition keiner Person zugeordnet ist, bspw. ein nicht weiter kontextualisierbares Fragment mit Jahreszahl)

Während im Inschriftenrepositorium jüdischer Grabinschriften (epidat) das Verhältnis zwischen erhaltenen Inschriften über Männer und Frauen nahezu ausgewogen ist, überwiegen in DIO bei weitem die Inschriften über Männer. Mögliche Erklärungen hierfür sind u.a. das Übergewicht an Geistlichen im DIO-Bestand, die es in dieser Form im Judentum nicht gibt.

2.4. Der Workflow

1) Überschneidungen epigraphischer Repositorien und Fragestellungen herausarbeiten:
   a) Grundlegende Schnittmengen: Personen, Orte, Konzepte, Symbole, Material, Objekttyp, u.a.m.
   b) Synergien: Nutzen von und Aufbau auf der bereits getätigten und XML-inhärenten Formalisierungsleistung der Datenbestände.[10]
   c) Einsatz und projektindividuelle Bereitstellung von schablonenartigen Templates zur Extraktion (Bsp.: EpiDoc-Header).
2) Formulierung der Triples:
   a) Formalisierung der Fragestellung
   b) Operationalisierung der Schnittmengen
   c) (Prototypische Konstruktion “flacher” Statements)

[10] Vgl. Grüntgens, Schrade (2016).

   d) Formulierung der Statements
3) Extraktion der Triples in ein RDF-Serialisierungsformat
   a) Bereitstellung als Daten-Dump in verschiedenen Serialisierungsformaten
   b) (Bereitstellung als abfragbare Ressource über einen SPARQL-Endpoint)
4) Import und Aggregation von Extraktionen in einem Triple-Store
   a) Abfragen auf Basis der formalisierten Fragestellung
   b) Einbezug externer Repositorien
   c) Export in einem visualisierungsfreundlichen Format (bspw. CSV)
5) Abfrage (allgemein und federated) und Export der aggregierten Datengrundlage via SPARQL-Endpoint
   a) Beispielabfragen zur Dokumentation des Datenbestandes erstellen.
   b) Bereitstellung von vorgenerierten Queries zur Erkundung des Datenbestandes.
6) Visualisierung über Webservice oder eigene Visualisierung
   a) Bspw. mit Rawgraphs.io
   b) Bspw. mit D3.js
7) (Gegebenenfalls Re-Iteration des Prozesses oder seiner Teile)

2.5. Schlussbetrachtung

Die obigen Ausführungen zum Use Case und dem allgemeinem Workflow zeigen, dass sich die Arbeit mit semantisch annotierten Forschungsdaten sinnvoll und gewinnbringend in bestehende Arbeitsabläufe und Forschungspraktiken integrieren lassen, bzw. im Falle von bereits abgeschlossenen Projekten auf bestehenden Forschungsdaten-Veröffentlichungen aufsetzen kann. Grundlegend für einen sinnvollen Einsatz ist es dabei, neben der Herausarbeitung einer Repositorien-übergreifenden Fragestellung, die den zugrundeliegenden Forschungsdaten in XML bereits inhärente Formalisierungsleistung zu nutzen und im Rahmen des darauf aufbauenden, auf RDF und LOD ausgerichteten Formalisierungsprozesses Synergien zu nutzen. Zu betonen ist, dass die Modellierung in RDF als Erweiterung der bereits bestehenden Forschungsdaten und eine Spiegelung derselben auf eine gemeinsame abstrakte Metadaten-Ebene zu verstehen ist. Damit soll unter anderem auch ein Anknüpfen an weitere, fachübergreifende Fragestellungen ermöglicht werden.[11]

Im Folgenden kann eine erste prototypische, “flache” Triple-Extraktion und Visualisierung den Blick auf Desiderate und Möglichkeiten einer sinnvollen Angleichung der Repositorien aneinander schärfen. In diesem Sinne liegt der Fokus beim obigen Workflow auch auf einer Iteration des Prozesses in Analogie zum Bild der “hermeneutischen Spirale”. Ist nach einigen Iterationen ein als adäquat empfundenes Niveau erreicht, sind die “flachen” Triple-

[11] Vgl. Brodhun, de la Iglesia, Moretton (2015), Grüntgens, Schrade (2016), 61–62; vgl. auch Ciotti, Tomasi (2016), Ciotti (2018); vgl. auch Bender, Rapp, Kollatz (2018), 121–122, 125–127.

Modellierungen im Sinne etablierter Ontologien formal auszuformulieren und können im Anschluss auch anderen zur Verfügung gestellt werden.[12]

Die Projektergebnisse wurden auf diversen Veranstaltungen präsentiert und mit der Fachcommunity diskutiert:

Grüntgens, Max und Kasper, Dominik (2017) “Semantische Annotation und Kodierung: Verstehen – Auszeichnen – Abfragen”, Mainz. [Abrufbar: https://digitale-methodik.adwmainz.net/mod5/5c/slides/annotationen/XML_2017]

Kollatz, Thomas (2018) “Digitale Approaches to Cemeteries: Preservation and Education”, All that remains: Education and Conservation of Jewish Funerary Culture, Utrecht. [Abrufbar: https://kollatzthomas.github.io/20180611_Utrecht]

Grüntgens. Max und Kollatz, Thomas (2018a) “Digital Humanites and Jewish Epigraphy”, EAJS EpiDoc Winterschool, Utrecht University. [Abrufbar: https://digicademy.github.io/2018_EAJS_WS_1]

——— (2018b) “Annotieren, analysieren, visualisieren. Repositorien mergen, analysieren und visualisieren”, DARIAH-DE Grand Tour, Darmstadt. [Abrufbar: https://digicademy.github.io/DARIAH-GT]

——— (2018c) “Two Become One: Lifting TEI EpiDoc Encoded Corpora to RDF – The Case of DIO and EPIDAT”, Linked Pasts IV: Views From Inside The LOD-cloud, Mainz. [Abrufbar: https://digicademy.github.io/Two_become_one]

Kollatz, Thomas, Andreas Kuczera, and Torsten Schrade (2016) “Methods and Tools for Visualising Digital Humanities Data Sets”, DHd 2016, Leipzig. [Abrufbar: http://dhd2016.digitale-akademie.de/workshop]

Schrade, Torsten (2016) “CIDOC-CRM Modellierung epigraphischer Fachdaten mit dem XTriples Webservice”, DARIAH-Expertenworkshop Geisteswissenschaftliche Forschungsinfrastrukturen und CIDOC-CRM Annotation, Darmstadt. [Abrufbar: https://digicademy.github.io/dariah-workshop2016-xtriples-cidoc]

[12] Vgl. Eide, Ore (2019), 194–195; Herausforderungen und mögliche Ansätze bei der Modellierung von Epigraphik in CIDOC-CRM zeigen Felicetti, Murano, Ronzino, Niccolucci (2015) auf.

3. Informationsvisualisierung hochdimensionaler Annotationsdaten

Im Anwendungsfall “Informationsvisualisierung hochdimensionaler Annotationsdaten” werden Forschungsdaten aus dem Sonderforschungsbereich 980 (SFB 980) “Episteme in Bewegung” der Freien Universität Berlin[13] nachgenutzt. Dieser Sonderforschungsbereich untersucht Prozesse des Wissenstransfers und Wissenswandels in europäischen und nichteuropäischen Kulturen vom 3. Jahrtausend vor Christus bis etwa 1750 nach Christus. Konkret liegen der entwickelten Informationsvisualisierung die Daten eines der 27 Teilprojekte zu Grunde, das Prozesse der Traditionsbildung in der *de interpretatione*-Kommentierung der Spätantike untersucht, um Erkenntnisse über beispielsweise Verwandtschaftsverhältnisse oder Verwendungskontexte einzelner Handschriftenexemplare zu gewinnen.

Zurzeit werden für diesen Zweck 46 Aristotelesmanuskripte mit 2406 Einzelseiten inklusive administrativer und inhaltlicher Metadaten im ‘Episteme-Repositorium’ strukturiert verwaltet, nachhaltig gesichert sowie kontinuierlich ergänzt, so dass schlussendlich mindestens 120 der noch ca. 150 überlieferten Manuskripte für eine umfassende Analyse bereitstehen. Im Zentrum des Interesses stehen hierbei kodikologische Gegebenheiten, wobei insbesondere das Layout eine zentrale Rolle einnimmt. Der eigentliche Aristotelestext nimmt in vielen Manuskripten nur einen relativ begrenzten Raum auf der Seite ein, großzügige Ränder und Zeilenabstände sollen vielfältige Kommentare, Erklärungen und Diagramme ermöglichen. Die tatsächliche Ausgestaltung variiert dabei je nach Exemplar zwischen keinerlei Ergänzungen und hochkomplexen, dicht von verschiedenen Schreibern in verschiedenen Orientierungen beschriebenen Seiten. Layouteigenschaften, wie beispielsweise das Verhältnis von Haupttext zur Seite oder Zeilenabstände, erlauben die Aufdeckung von Manuskriptgruppen bzw. Verwandtschaftsverhältnissen und sind mit Hilfe (semi-)automatischer Algorithmen auch objektiv und reproduzierbar messbar.

3.1. Modellierung quantitativer und qualitativer Informationen

Die Aristotelesmanuskripte mit ihrer hochkomplexen Seitengestaltung stellen einen besonders schwierigen Layoutfall dar. Die Erkennung struktureller Elemente wie Seiten-, Text- und Bildbereiche kann durch (semi-)automatische Algorithmen unterstützt werden, gleichzeitig muss eine manuelle, fachwissenschaftliche Korrektur und Ergänzung stets möglich bleiben. Auf diese Weise werden auch semantische Einordnungen sowie höchstes Spezialwissen in

[13] Sonderforschungsbereich 980 „Episteme in Bewegung“: http://www.sfb-episteme.de

der Quellenforschung abgelegt und daher quantitative und qualitative Informationen gleichzeitig für eine automatische Analyse zugänglich.

Zur Klassifizierung und Segmentierung wurde die auf Machine Learning Algorithmen basierende Layoutanalyse[14] des Projekts eCodicology[15] nachgenutzt. Hochauflösende Bilddateien, trainierte Klassifizierer sowie die physische Größe der Handschriftenseite im Rahmen eines Seitenmodells wurden verwendet, um Seiten- und Textbereiche jeder einzelnen Seite zu segmentieren, die entsprechenden Größen zu berechnen und damit Vergleiche zwischen Manuskripten zu ermöglichen. Ebenso wurden weitere quantitative Eigenschaften wie beispielsweise Zeilenanzahl und mittlere Zeilenabstände von Textregionen, Farbeigenschaften (mittlere Sättigungs- und Helligkeitswerte, Standardabweichung, etc.) sowie Größenverhältnisse zwischen Regionen (z.B. Text-Seiten-Verhältnis) erfasst. Die Ergebnisse werden zunächst im PAGE-Standard[16] abgelegt, der im Bereich der Optical Character Recognition weit verbreitet seine Anwendung findet.

Für den gewünschten Verwendungszweck einer Verschränkung von quantitativen und qualitativen Informationen ist der Standard jedoch nicht praktikabel: Informationen sind strukturell und nicht inhaltlich gebündelt sowie eine Anreicherung mit zusätzlichen Informationen jenseits der Layoutanalyse nicht im Fokus des Modells. Aus diesem Grund kommt im beschriebenen Anwendungsfall das so genannte Web Annotation Data Model (WADM)[17] zum Einsatz, das im Februar 2017 mit dem Ziel einer einfachen Austauschbarkeit von Annotationen über Disziplin- und Systemgrenzen hinweg veröffentlicht wurde. In diesem Modell werden Annotationen als Anreicherung von Webressourcen durch andere Webressourcen verstanden und sind damit weitgehend unabhängig von konkretem Annotationsinhalt und -zweck spezifiziert. Auf diese Weise ist sowohl für das zu annotierende Objekt (das sogenannte Target) als auch für die hinzuzufügende Information (der sogenannte Body) größtmögliche Flexibilität gewährleistet.

Alle aus der Layoutanalyse gewonnen Informationen konnten durch die vielfältigen Kombinationsmöglichkeiten von Targets und Bodies in verschiedenen Kardinalitäten modelliert und automatisch ins WADM überführt werden. Der entwickelte Parser nutzt dabei ein allgemei-

---

[14] Chanda, Swati; Tonne, Danah; Jejkal, Thomas; Stotzka, Rainer; Krause, Celia; Vanscheidt, Philipp; Busch, Hannah; Prabhune, Ajinkya (2015): „Software workflow for the automatic tagging of medieval manuscript images (SWATI)“, in: *Document Recognition and Retrieval XXII. International Society for Optics and Photonics*, S. 940206.

[15] eCodicology - Algorithmen zum automatischen Tagging mittelalterlicher Handschriften: http://www.ecodicology.org

[16] PAGE XML-Schema: https://www.primaresearch.org/schema/PAGE/gts/pagecontent/2017-07-15/pagecontent.xsd

[17] Young, Benjamin / Ciccarese, Paolo / Sanderson, Robert (2017): „Web Annotation Data Model. W3C Recommendation“ https://www.w3.org/TR/2017/REC-annotation-model-20170223

nes Mapping von PAGE zu WADM, so dass die Umwandlung grundsätzlich, sofern noch keine zusätzlichen Informationen hinzugefügt wurden, reversibel bleibt.

Die folgende Abbildung (Abb. 4) zeigt Auszüge einer JSON-Repräsentation einer Annotation im WADM mit möglichen qualitativen (von Fachwissenschaftlern manuell eingeben) und/oder quantitativen (mit Hilfe der Layoutanalyse berechnet) Informationen, die in verschiedenen Bodies modelliert wurden. Zur Unterstützung einer automatischen Auswertung und Visualisierung werden die einzelnen Bodies mit Hilfe des Feldes Purpose kategorisiert. Das Target entspricht in diesem Anwendungsfall der Referenz auf eine Manuskriptseite, wobei die Referenz auf einen Ausschnitt durch einen Selektor im SVG-Standard[18] realisiert wird. Sowohl auf Annotationsebene als auch in den einzelnen Bodies wird die Herkunft der Informationen durch Angabe eines Creators belegt.

Abb. 4: JSON-Repräsentation einer Annotation (links) mit Auszügen möglicher Bodies aus fachwissenschaftlicher Annotation (Mitte) bzw. (semi-)automatischer Layoutanalyse (rechts) im Web Annotation Data Model

3.2. Entwicklung der notwendigen Annotationsinfrastruktur

Im beschriebenen Anwendungsfall mussten insbesondere zwei technische Komponenten infrastrukturell verankert werden, um die erforderlichen Funktionalitäten beim Umgang mit den digitalen Annotationen zu ermöglichen: eine grafische Benutzeroberfläche zur einfachen Erzeugung und Modifikation von Annotationen sowie ein verlässlicher Annotationsspeicher.

[18] SVG: https://www.w3.org/TR/SVG11

Die Annotationsoberfläche ist eine JavaScript-Anwendung (vgl. Abb. 5) und nutzt die in DARIAH-DE II entwickelte Bibliothek SemToNotes[19] nach. Den Fachwissenschaftlern wird die Möglichkeit geboten, die zu untersuchende Manuskriptseite mit sämtlichen vorhandenen Annotationen anzuzeigen. Ebenso können neue Bildbereiche markiert (wahlweise als Rechteck oder Polygon), Bildbereiche vorhandener Annotationen modifiziert sowie Begriffe standardisierter und/oder projektspezifischer Vokabulare oder Freitext hinzugefügt werden.

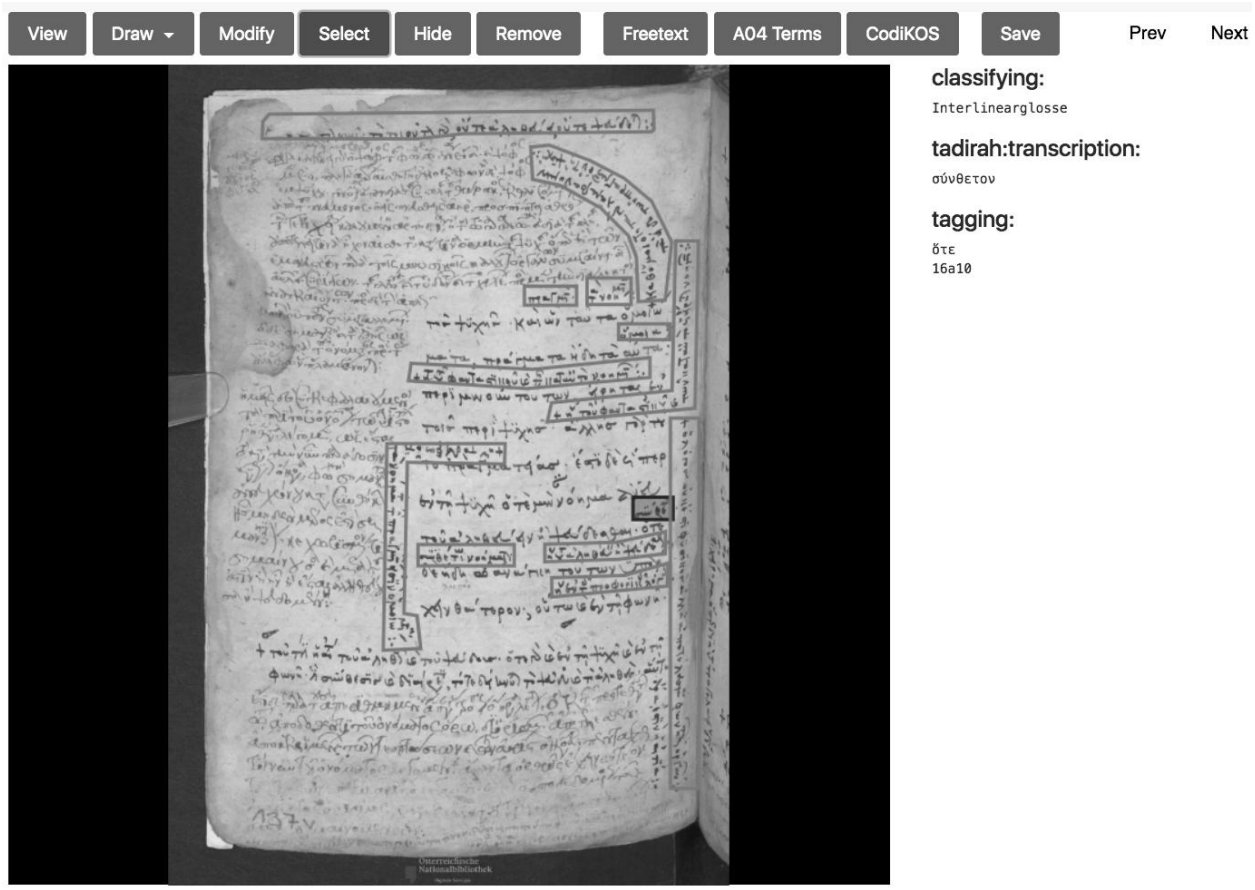

Abb. 5: Screenshot Annotationsoberfläche mit Cod Phil. gr. 300, 137v.
Quelle: Österreichische Nationalbibliothek, URL: http://data.onb.ac.at/rep/1002E2A7

Im konkreten Fall der *de interpretatione*-Kommentierung wird die Oberfläche von den Fachwissenschaftlern zum einen genutzt, um einer quantitativen Untersuchung bisher nicht zugängliche Interlinearglossen und erklärende Diagramme zu markieren, zu klassifizieren, zu transkribieren, zu übersetzen und mit Schlagwörtern zu versehen. Zum anderen bewerten und ergänzen, falls notwendig, die Fachwissenschaftler die aus der Layoutanalyse gewonnen Ergebnisse. Für eine Auswertung stehen also ebenso die Informationen zur Verfügung, ob entweder die Erkennung aus fachwissenschaftlicher Sicht zufriedenstellend erfolgt ist oder welche Bereiche idealerweise hätten erkannt werden sollen. Aus informatischer Sicht

[19] SemToNotes: https://hkikoeln.github.io/SemToNotes

lassen sich auf diese Weise sowohl die verwendeten als auch zukünftige Algorithmen auf einem heterogenen Datenbestand evaluieren und weiter verbessern.

Der entwickelte Annotationsspeicher ist eine Java-basierte Entwicklung, die die im Web Annotation Protocol[20] definierte REST-Schnittstelle bereitstellt. Hauptaugenmerk lag auf der modularen Anbindung eines RDF Triple Store (derzeit Apache Jena TDB2[21], aber prinzipiell austauschbar), so dass ein SPARQL 1.1[22] Endpunkt für semantische Anfragen an die Annotationsdaten bereitsteht. Auf diese Weise werden auch die Annotationsinhalte der Bodies auswertbar, indem Bodies untereinander oder auch mit Targets in Beziehung gesetzt werden. Darüber hinaus stehen weitere Metadaten der Linked Open Data Cloud durch föderierte SPARQL-Anfragen für die Analyse zur Verfügung, so dass Norm- und Geodaten nachgenutzt sowie projektübergreifende Vokabulare und Taxonomien verwendet werden können.

3.3. Visualisierung und Ausblick

Aktuell sind ca. 15.500 manuelle und automatische Annotationen im WADM im Annotationsspeicher abgelegt, werden kontinuierlich ergänzt und stehen erstmals für eine gemeinsame Auswertung zur Verfügung.

Exemplarisch zeigt die Abbildung (Abb. 6) die implementierte Visualisierung eines einzelnen Satzes für die bereits annotierten Aristoteleshandschriften. Mit Hilfe von SPARQL-Anfragen werden maßgebliche Informationen (beispielsweise Transkription und zugehöriges Wort oder Satzfragment) der zu diesem Satz gehörenden Interlinearglossen abgerufen. Anschließend werden verschiedene Variationen der Interlinearglossen verglichen, gezählt und visualisiert, um Gemeinsamkeiten zwischen Manuskriptgruppen aufzudecken. Dieses Verfahren ist problemlos für andere Abschnitte, andere Kommentierungsformen sowie die Untersuchung von Textvarianten nutzbar und unterstützt damit maßgeblich die Erforschung der hochkomplexen Austauschprozesse in der *de interpretatione*-Kommentierung.

[20] Sanderson, Robert (2017): „Web Annotation Protocol. W3C Recommendation“ https://www.w3.org/TR/2017/REC-annotation-protocol-20170223
[21] TDB2: https://jena.apache.org/documentation/tdb2
[22] SPARQL: https://www.w3.org/TR/sparql11-query

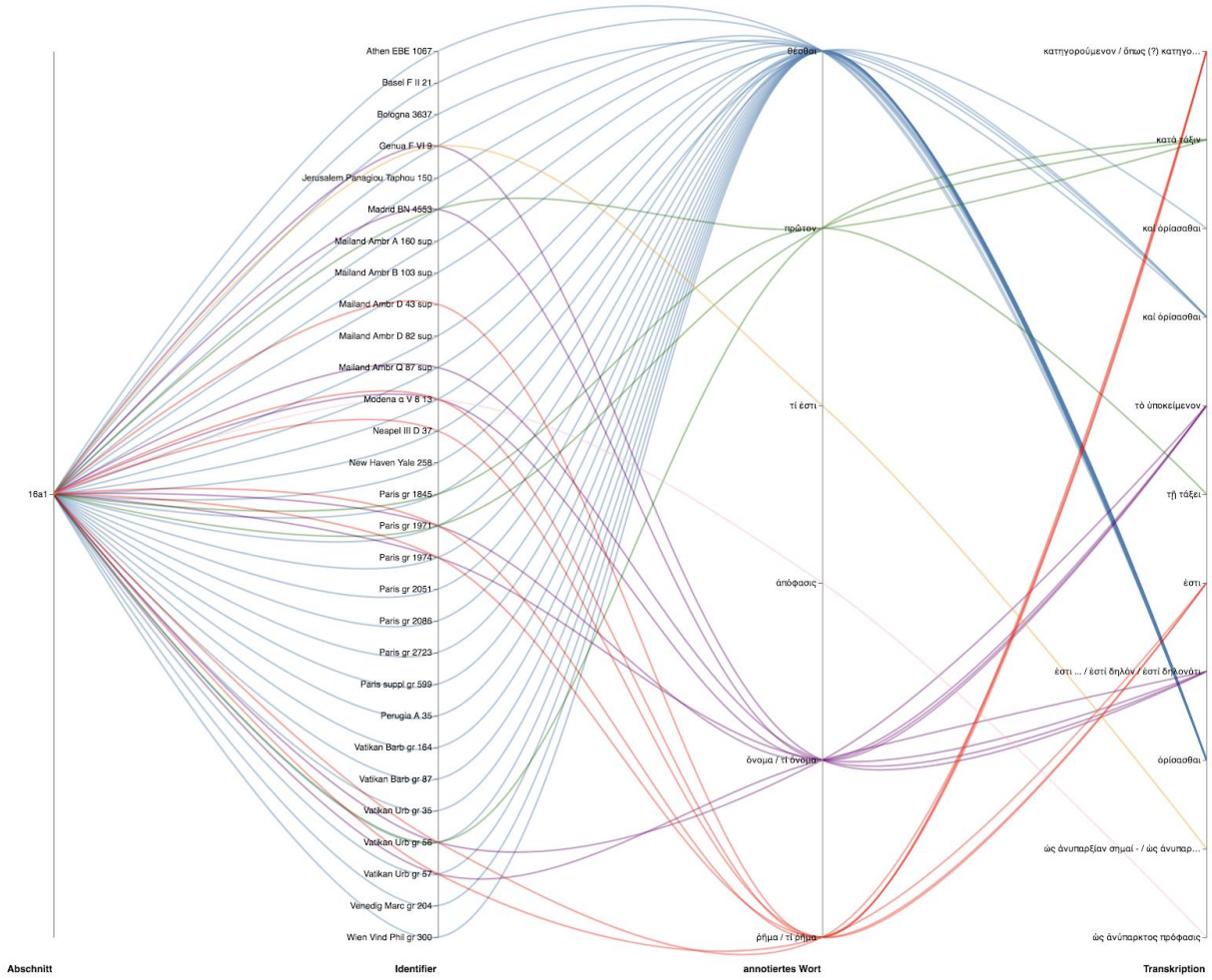

Abb. 6: Übersicht der Interlinearglossen des Satzes 16a1 aus *de interpretatione*
(1. Achse: Satz-ID des Referenztextes, 2. Achse: Sigle der Handschrift, 3. Achse: glossiertes Wort im Referenzsatz, 4. Achse: Transkription der Glosse)

Die erzielten Projektergebnisse wurden und werden auf folgenden Veranstaltungen der Fachcommunity vorgestellt und publiziert:

Tonne, Danah: Aristoteles annotieren. Cluster 6 Workshop „Annotieren, analysieren, visualisieren“ im Rahmen der DARIAH-DE Grand Tour 2018, 20.09.2018

Götzelmann, Germaine; Tonne, Danah: Aristoteles annotieren - Vom Handschriftendigitalisat zur qualitativ-quantitativen Annotation in „Digitale Bilddaten in den Geisteswissenschaften“ in der Reihe Episteme (SFB 980) beim Harrassowitz Verlag (erscheint 2019)

Tonne, Danah; Götzelmann, Germaine; Hegel, Philipp; Krewet, Michael; Hübner, Julia; Söring, Sibylle; Löffler, Andreas; Hitzker, Michael; Höfler, Markus; Schmidt, Timo: Ein

Web Annotation Protocol Server zur Untersuchung vormoderner Wissensbestände, DHd 2019 - 6. Tagung des Verbandes “Digital Humanities im deutschsprachigen Raum (DHd)” (akzeptiert)

4. Literaturverzeichnis

Bender, Michael, Rapp, Andrea, and Thomas Kollatz. 2018. “Objekte Im Digitalen Diskurs – Epistemologische Zugänge Zu Objekten Durch Digitalisierung Und Diskursive Einbindung in Virtuelle Forschungsumgebungen Und -Infrastrukturen.” In *Berlin Studies of the Ancient World. Objektepistemologien*. Berlin.

Brodhun, Maximilian; de la Iglesia, Martin; Moretto, Nicolas: Metadaten, LOD und der Mehrwert standardisierter und vernetzter Daten, in: Neuroth/Rapp/Söring: TextGrid: Von der Community – für die Community, Göttingen 2015, 91–102, URL: http://dx.doi.org/10.3249/webdoc-3947

Chanda, Swati; Tonne, Danah; Jejkal, Thomas; Stotzka, Rainer; Krause, Celia; Vanscheidt, Philipp; Busch, Hannah; Prabhune, Ajinkya (2015): „Software workflow for the automatic tagging of medieval manuscript images (SWATI)“, in: *Document Recognition and Retrieval XXII. International Society for Optics and Photonics*, S. 940206

Ciotti, Fabio: A Formal Ontology for the Text Encoding Initiative. 2018, URL: https://doi.org/10.6092/issn.2532-8816/8174

Ciotti, Fabio und Tomasi, Francesca: Formal Ontologies, Linked Data, and TEI Semantics, Journal of the Text Encoding Initiative [Online], Issue 9 | September 2016 - December 2017, Online since 24 September 2016, connection on 10 January 2019. URL: http://journals.openedition.org/jtei/1480
eCodicology - Algorithmen zum automatischen Tagging mittelalterlicher Handschriften: http://www.ecodicology.org/

Eide, Øyvind und Ore, Christian-Emil: 8. Ontologies and Data Modeling, in: Flanders, Jannidis (ed.): The Shape of Data in Digital Humanities Modeling Texts and Text-based Resources, London 2019, 178–196.

Felicetti, Achille; Murano, Francesca, Ronzino, Paola und Niccolucci, Franco 2015: CIDOC CRM and Epigraphy: a Hermeneutic Challenge. CEUR Workshop Proceedings 1656: 55–68, URL:http://ceur-ws.org/Vol-1656/paper5.pdf

Grüntgens, Max, and Torsten Schrade. 2016. “Data repositories in the Humanities and the Semantic Web: modelling, linking, visualising.” CEUR Workshop Proceedings 1608: 53–63, URL: http://ceur-ws.org/Vol-1608/paper-07.pdf

Kollatz, Thomas, and Max Grüntgens. 2018. “Korpusbasiertes Arbeiten und epigraphische Datenbanken. Möglichkeiten und Herausforderungen am Beispiel von EPIDAT und DIO.” *Osnabrücker Beiträge zur Sprachtheorie* 92: 157–74.

PAGE XML-Schema: https://www.primaresearch.org/schema/PAGE/gts/pagecontent/2017-07-15/pagecontent.xsd

Sanderson, Robert (2017): „Web Annotation Protocol. W3C Recommendation“ https://www.w3.org/TR/2017/REC-annotation-protocol-20170223/

SemToNotes: https://hkikoeln.github.io/SemToNotes/

Sonderforschungsbereich 980 „Episteme in Bewegung“: http://www.sfb-episteme.de/

SVG: https://www.w3.org/TR/SVG11/

SPARQL: https://www.w3.org/TR/sparql11-query/

TDB2: https://jena.apache.org/documentation/tdb2/

Young, Benjamin / Ciccarese, Paolo / Sanderson, Robert (2017): „Web Annotation Data Model. W3C Recommendation“ https://www.w3.org/TR/2017/REC-annotation-model-20170223/

5. Abbildungsverzeichnis