Ein Web Annotation Protocol Server zur Untersuchung vormoderner Wissensbestände

Tonne, Danah
danah.tonne@kit.edu
Karlsruher Institut für Technologie, Deutschland

Götzelmann, Germaine
germaine.goetzelmann@kit.edu
Karlsruher Institut für Technologie, Deutschland

Hegel, Philipp
philipp.hegel@fu-berlin.de
Freie Universität Berlin, Deutschland

Krewet, Michael
m.krewet@fu-berlin.de
Freie Universität Berlin, Deutschland

Hübner, Julia
julia.huebner@fu-berlin.de
Freie Universität Berlin, Deutschland

Söring, Sibylle
ssoering@cedis.fu-berlin.de
Center für Digitale Systeme

Löffler, Andreas
andreas@loeffler-bermayer.de
Karlsruher Institut für Technologie, Deutschland

Hitzker, Michael
mail@hitzker.de
Karlsruher Institut für Technologie, Deutschland

Höfler, Markus
Markus.Hoefler@live.de
Karlsruher Institut für Technologie, Deutschland

Schmidt, Timo
timo.schmidt2@student.kit.edu
Karlsruher Institut für Technologie, Deutschland

Der Sonderforschungsbereich 980 „Episteme in Bewegung“

Der Sonderforschungsbereich 980 (SFB 980) „Episteme in Bewegung“ untersucht Prozesse des Wissenstransfers und Wissenswandels in europäischen und nicht-europäischen Kulturen vom 3. Jahrtausend vor Christus bis etwa 1750 nach Christus. In 27 Teilprojekten aus 21 Disziplinen wird gezeigt, wie gerade dort, wo in den Selbstbeschreibungen der vormodernen Kulturen und aus der Perspektive der Moderne Kontinuität und Stabilität im Vordergrund stehen, Neukontextualisierungen von Wissen fassbar werden. Die konkreten Möglichkeiten digitaler Methoden sowie Auswirkungen auf den Forschungsprozess werden an Hand zweier Anwendungsfälle dargelegt.

Anwendungsfall 1: Prozesse der Traditionsbildung in der de interpretatione-Kommentierung der Spätantike

Die Erforschung der handschriftlichen Überlieferung einer Schrift kann Erkenntnisse über beispielsweise Verwandtschaftsverhältnisse oder Verwendungskontexte einzelner Exemplare liefern. Im Falle von Aristoteles’ Schrift *de interpretatione*, von der 150 griechischsprachige Handschriften erhalten sind, stößt die analoge Untersuchung jedoch an ihre Grenzen (Montanari 1984, Reinsch 2001). Vielfach hatte ein Kopist eines Textes gleich mehrere Exemplare mit unterschiedlichen Textversionen vor sich. Beim Kopieren konnte er diese vermischen oder sogar Vorlagen im Sinne einer für besser befundenen Version korrigieren. Ebenso konnte er beim Kopieren verschiedene Erklärungen unterschiedlichster Provenienz an den Rand des Textes schreiben. Diese Praktik macht es immens schwierig, den genauen Weg, wie die Schrift überliefert wurde, und Traditionen unter den Erklärungen zu erforschen.

Anwendungsfall 2: Vermittlung kommunikativer Alltagsroutinen im Kontext sprachlicher Diversität in der Frühen Neuzeit

Im Zentrum stehen 315 Lehrwerke für Moderne Fremdsprachen mit Deutsch-Anteil (Glück 2002) aus der Frühen Neuzeit (Mitte des 15. Jahrhunderts bis Anfang des 18.Jahrhunderts). Diese Sprachlehrwerke waren in der Regel mehrsprachig angelegt und richteten sich an Reisende aller Art. Einige dieser Drucke wurden über Jahrzehnte immer wieder in überarbeiteter Form herausgegeben und eignen sich daher besonders gut für die Untersuchung von Wandelprozessen. Eine weitere spannende Fragestellung bilden die Autorisierungsstrategien in diesen Lehrwerken: wer ist

autorisiert, Sprache zu beschreiben / zu unterrichten und mit welchen Mitteln legitimieren die Autoren ihre Werke?

Modellierung fachwissenschaftlichen Wissens als Annotation

Zentrale Methodik zur Untersuchung der Forschungsfragen in beiden Anwendungsfällen ist die Anreicherung von Bilddigitalisaten mit zusätzlichen Informationen. Dieses Wissen wird auf technischer Ebene als digitale Annotation modelliert, wobei das verwendete Modell für beliebige referenzierbare Daten nutzbar sein soll. Seit Februar 2017 steht dazu mit dem *Web Annotation Data Model* (WADM) des W3C-Konsortiums (Young et al. 2017) der Nachfolger zum weitverbreiteten *Open Annotation Data Model* zur Verfügung. Mit diesem standardisierten Modell und Format soll der Austausch von Annotationen über Disziplin- und Systemgrenzen ermöglicht werden.

In der Infrastruktur des SFB 980 finden diese Empfehlungen ihre Anwendung, um der großen Heterogenität der Disziplinen, Erkenntnisinteressen und Arbeitsweisen und damit auch der Vielfältigkeit der Annotationstypen, -formate und -praktiken Rechnung zu tragen. Schon bei den zwei betrachteten Anwendungsfällen entstammen die Annotationen drei unterschiedlichen Quellen:

1. *Adobe Acrobat:* Forschende annotieren Digitalisate innerhalb von PDFs, die Annotationen können als XFDF exportiert werden.
2. *Automatische Layoutanalyse:* Algorithmen vermessen Seiten-, Text- und Bildbereiche, die Informationen werden im PAGE-Standard abgelegt.
3. *Projektspezifische grafische Annotationsoberfläche:* Informationen werden direkt als Annotationen gemäß des Web Annotation Data Models gespeichert.

Um eine disziplinübergreifende Auswertung zu ermöglichen, werden die XML-Dateien aus 1. und 2. mit Hilfe eines Parsers in das Web Annotation Data Model überführt und im Annotationsspeicher zugreifbar abgelegt. Abbildung 1 zeigt Auszüge der Modellierung für die automatische Layoutanalyse sowie manueller fachwissenschaftlicher Annotation. Die vermessenen bzw. eingegebenen Informationen werden im so genannten *body* aufgeführt. Zentral ist hierbei die Nutzung des Feldes *purpose* zur Klassifizierung der einzelnen *bodies* mit Hilfe von Kategorien aus dem Web Annotation Data Model (z.B. *classifying*, *tagging*) und der TaDiRAH (Borek et al. 2016) Taxonomie (z.B. *translation*, *transcription*), um eine automatische Auswertung und Visualisierung zu ermöglichen. Auch die aus der Layoutanalyse stammende Unterscheidung von Seiten-, Text- und Bildbereichen wird im *body* modelliert. Die annotierten Bilder werden im *target* referenziert, Bildausschnitte konform mit dem SVG-Standard definiert. Durch die Nutzung des Feldes *creator* sowohl auf Annotationsebene als auch in den *bodies* kann die Herkunft der Informationen nachverfolgt werden und können beispielsweise verschiedene Algorithmusversionen der Layoutanalyse verglichen werden.

```
{
  "@context" : "http://www.w3.org/ns/anno.jsonld",
  "id" : "http://host:port/annotation/w3c/uuid",
  "type" : "Annotation",
  "created" : "2018-06-21T08:01:25.011Z",
  "creator" : [ {
    "type" : "Software",
    "name" : "Akita"
  }, {
    "type" : "Person",
    "name" : "MK"
  } ],
  "modified" : "2018-06-21T08:02:31.048Z",
  "body" : [ ],
  "target" : {
    "type" : "SpecificResource",
    "selector" : {
      "type" : "SvgSelector",
      "value" : "<svg><rect x=\"591\" y=\"2610\"
      width=\"192\" height=\"78\"/></svg>"
    },
    "source" : "http://host:port/pathToImage/947"
  },
  "motivation" : "describing"
}
```

```
"body" : [ {
  "type" : "TextualBody",
  "format" : "text/plain",
  "dc:subject" : "TextRegion",
  "creator" : [ ],
  "value" : "Interlinearglosse",
  "purpose" : "classifying"
}, {
  "type" : "TextualBody",
  "format" : "text/plain",
  "dc:subject" : "TextRegion",
  "creator" : [ ],
  "value" : "16a1",
  "purpose" : "tagging"
}, {
  "type" : "TextualBody",
  "dc:subject" : "TextRegion",
  "creator" : [ ],
  "value" : "bestimmt werden",
  "purpose" : "tadirah:translation"
}, {
  "type" : "TextualBody",
  "dc:subject" : "TextRegion",
  "creator" : [ ],
  "value" : "Spinnmöbel",
  "purpose" : "tadirah:transcription"
```

```
"body" : [ {
  "id" : "urn:anno4j:0431d5b6-23ba-4ad2-85c8-a3dbcb911614",
  "type" : [ "oa:EmbeddedContent", "TextualBody",
  "https://github.com/anno4j/ns#CreationProvenance",
  "https://github.com/anno4j/ns#ExternalWebResource",
  "https://github.com/anno4j/ns#Resource" ],
  "format" : "xsd:string",
  "dc:subject" : "PageRegion",
  "dc:title" : "DigitizedManuscriptPageWidth",
  "creator" : "urn:uuid:c4dbcb3f-f03f-3ff6-8c6d-c0cdb44a00ac",
  "http://qudt.org/vocab/unit" : "cm",
  "value" : "20.999449"
}, {
  "id" : "urn:anno4j:d53b0686-40b8-4065-8c09-464fc707a038",
  "type" : [ "oa:EmbeddedContent", "TextualBody",
  "https://github.com/anno4j/ns#CreationProvenance",
  "https://github.com/anno4j/ns#ExternalWebResource",
  "https://github.com/anno4j/ns#Resource" ],
  "format" : "xsd:string",
  "dc:subject" : "PageRegion",
  "dc:title" : "Area",
  "creator" : "urn:uuid:c4dbcb3f-f03f-3ff6-8c6d-c0cdb44a00ac",
  "http://qudt.org/vocab/unit" : "sq.cm",
  "value" : "544.3107"
```


Abbildung 1: JSON-Repräsentation einer Annotation (links) mit Auszügen möglicher bodies aus fachwissenschaftlicher Annotation (Mitte) bzw. automatischer Layoutanalyse (rechts) im Web Annotation Data Model

Ein Blick unter die Haube: RDF-Server und SPARQL-Anfragen

Nach der Modellierung ist die verlässliche Sicherung und die standardisierte Abrufbarkeit der Annotationen von immenser Bedeutung. Das gemeinsam mit dem *Web Annotation Data Model* veröffentliche *Web Annotation Protocol* (Sanderson 2017) definiert eine REST-Schnittstelle für einen Annotationsspeicher und stellt so eine erfolgreiche Kommunikation zwischen Servern und Clients sicher. Ein im Rahmen der SFB-Infrastruktur entwickelter, generischer Annotationsserver setzt alle obligatorischen und viele der optionalen Vorgaben des Protokolls sowie die für Annotationen relevanten Teile der *Linked Data Platform* (LDP)-Empfehlungen vollständig um. Die javabasierte Serverarchitektur bindet dabei modular einen RDF Triple Store (derzeit Apache Jena TDB2, aber prinzipiell austauschbar) an. Auf diese Weise ist ein standardkonformer und damit an verbreitete Annotationsprogrammen anbindbarer Server entstanden, dessen lose gekoppelte Speicherkomponente bei beispielsweise Softwareobsoleszenz oder gestiegenen Skalierbarkeitsanforderungen mit geringstem Aufwand ausgetauscht werden kann. Nach aktuellem Stand sind ca. 15500 Bildannotationen aus automatischer Layoutanalyse und manueller Transkription und Klassifizierung im WADM-Format in Form einzelner RDF-Graphen abgelegt.

Neben der REST-Schnittstelle ist mit dem Triple Store zugleich auch ein SPARQL 1.1 Endpunkt verbunden, der semantische Anfragen an die Annotationsdaten ermöglicht. Hier rücken neben den formalen Vorgaben der Annotationen zur Erzeugung, Bearbeitung und Anzeige (Annotationstyp, Selektoren, Links zu Ressourcen, etc.), die Annotationsinhalte in den Blick. Diese

sind je nach Forschungsinteresse unterschiedlich ausgestaltet und in einem oder mehreren *bodies* abgelegt. Mit inhaltsspezifischen SPARQL-Anfragen können diese *bodies* miteinander oder mit den *targets* in Beziehung gesetzt und ausgewertet werden. Über die eigenen Serverinhalte hinaus können mit Hilfe sogenannter föderierter SPARQL-Anfragen weitere Metadaten innerhalb der *Linked Open Data Cloud* zur Analyse hinzugezogen werden, ohne dass diese redundant abgelegt werden müssen. So eröffnen sich Möglichkeiten zur Nachnutzung von Norm- und Geodaten sowie zur projektübergreifenden Nutzung von Vokabularen und Taxonomien.

Ergebnisse und Ausblick

Bei *de interpretatione* wurden Textbereiche, besondere Textvarianten, Erklärungen, Diagramme und Interlinearglossen mit zusätzlichen Informationen wie Beschreibungen, Transkriptionen, Übersetzungen und vielfältigen Tags angereichert. Mit Hilfe von SPARQL-Anfragen werden beispielsweise die Transkriptionen von Interlinearglossen auf verschiedenen Digitalisatseiten dahingehend abgefragt, zu welchem Satz des Aristoteles-Textes und zu welchem Wort oder Satzfragment sie gehören (realisiert durch *tagging*). Auf diese Weise können verschiedene Variationen der Interlinearglossen verglichen, gezählt und visualisiert werden. Schon bei Betrachtung eines einzelnen Satzes (vergleiche Abbildung 2) lassen sich im Handschriftenkorpus Gemeinsamkeiten zwischen Manuskriptgruppen aufzeigen. Für die anderen Formen der antiken Kommentierung sowie der Textvarianten gestaltet sich die Auswertung analog. Die so durchgeführte Untersuchung von signifikanten Gemeinsamkeiten und Variationen *aller* Handschrifteninhalte ergänzt die herkömmliche textkritische Methode und stellt damit eine unerlässliche Hilfe für die Erforschung der komplexen Austauschprozesse dar, die bei der Textüberlieferung stattgefunden haben.

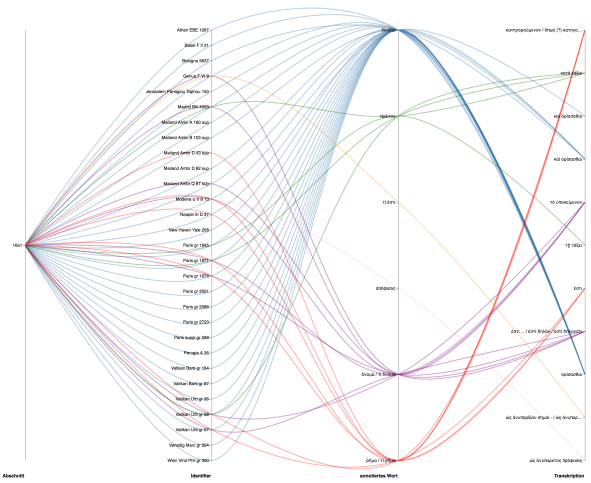

Abbildung 2: Übersicht der Interlinearglossen des Satzes 16a1 aus de interpretatione (1. Achse: Satz des Referenztextes, 2. Achse: Signatur der Handschrift, 3. Achse: glossiertes Wort im Referenzsatz, 4. Achse: Transkription der Glosse)

Im Falle der frühneuzeitlichen Sprachlehrwerke dient ein Annotationstool in einem ersten Schritt dem explorativen Annotieren und dem damit einhergehenden Aufdecken unterschiedlichster Prozesse des Wissenswandels. Im Weiteren wird das Tool dann genutzt, um Einzelanalysen vorzunehmen und um Teile des Korpus quantitativ auszuwerten. So werden Autorisierungsstrategien ausfindig gemacht und in verschiedene Kategorien unterteilt. Diese Ergebnisse werden anschließend mit den Metadaten wie Autor, Zielsprache, und didaktischer Ausrichtung des Lehrwerks korreliert.

Die vorgestellten Anwendungsfälle nutzen bereits die entwickelte Annotationsinfrastruktur und konnten eine neue Herangehensweise an ihre spezifischen Fragestellungen erproben. Die Infrastruktur erlaubt es, Text- und Bilddaten in ihren wechselseitigen Bezügen in den Blick zu nehmen und gemeinsam zu betrachten. Das *Web Annotation Data Model* hat sich als vielseitiges Modell bewiesen, um die heterogenen Annotationen zu modellieren. Dabei sind die Anwendungsmöglichkeiten jedoch weder auf die beiden Anwendungsfälle noch auf den Sonderforschungsbereich in seiner Gesamtheit beschränkt, sondern im Gegenteil Potentiale für alle Disziplinen und digitalen Datensätze erkennbar. Unterschiedliche, disziplinspezifische Praktiken und Anforderungen können im Modell abgebildet werden, die Annotationen verbleiben nach der zugehörigen Ontologie (Ciccarese 2017) strukturiert und damit auswertbar. Für die Inhalte ist allerdings kein semantisches Modell vordefiniert und auch nicht a priori festlegbar. Der Aushandlungsprozess eines solchen Modells ist daher für jede Anwendergruppe erneut und jeweils iterativ durchzuführen. Ebenso verbleibt die manuelle Annotation für die Forschenden arbeitsintensiv, da vielfältige Informationen abgelegt werden. Grafische Benutzeroberflächen können hier unterstützen, den Prozess so komfortabel wie möglich zu gestalten. Zusätzlich werden durch die Verwendung des *Web Annotation Data Models* die Kollaboration mit anderen Fachwissenschaftlerinnen und Fachwissenschaftlern und die Nachnutzbarkeit ermöglicht, so dass die Arbeitslast reduziert werden kann. Größter Vorteil ist jedoch die Möglichkeit, Informationen mit stabilen, in einem Repositorium abgelegten Daten zu verknüpfen und mit dynamisch im Forschungsprozess erlangtem Wissen anzureichern. Auf diese Weise werden Objektmetadaten, automatische und fachwissenschaftliche Annotationen gemeinsam auswertbar und damit gänzlich neue Erkenntnisse möglich.

Bibliographie

Borek, Luise / Perkins, Jody / Schöch, Christof / Dombrowski, Quinn (2016): *„TaDiRAH: A Case Study in pragmatic classification“* in: DHQ: Digital Humanities Quarterly 10/1.

Ciccarese, Paolo / Young, Benjamin / Sanderson, Robert (2017): *„Web Annotation Vocabulary. W3C Recommendation“* https://www.w3.org/TR/2017/REC-annotation-vocab-20170223/ [letzter Zugriff 19.12.2018].

Glück, Helmut (2002): *Deutsch als Fremdsprache in Europa vom Mittelalter bis zur Barockzeit*. Berlin/New York: De Gruyter.

Linked Data Platform (LDP) 1.0: https://www.w3.org/TR/ldp/ [letzter Zugriff 05.10.2018].

Montanari, Elio (1984): *La sezione linguistica del Peri Hermeneias di Aristotele*. Florenz.

PAGE XML-Schema: https://www.primaresearch.org/schema/PAGE/gts/pagecontent/2017-07-15/pagecontent.xsd [letzter Zugriff 05.10.2018].

Reinsch, Diether Roderich (2001): *„Fragmente einer Organon-Handschrift des zehnten Jahrhunderts aus dem Katharinenkloster auf dem Berg Sinai“* in: Philologus 151: 151-177.

Sanderson, Robert (2017): *„Web Annotation Protocol. W3C Recommendation“* https://www.w3.org/TR/2017/REC-annotation-protocol-20170223/ [letzter Zugriff 08.11.2018].

Sonderforschungsbereich 980 „Episteme in Bewegung“: http://www.sfb-episteme.de/ [letzter Zugriff 05.10.2018].

SPARQL: https://www.w3.org/TR/sparql11-query/ [letzter Zugriff 05.10.2018].

SVG: https://www.w3.org/TR/SVG11/ [letzter Zugriff 05.10.2018].

TDB2: https://jena.apache.org/documentation/tdb2/ [letzter Zugriff 05.10.2018].

XML Forms Data Format Specification: https://web.archive.org/web/20160408204348/https://partners.adobe.com/public/developer/en/xml/XFDF_Spec_3.0.pdf [letzter Zugriff 05.10.2018].

Young, Benjamin / Ciccarese, Paolo / Sanderson, Robert (2017): *„Web Annotation Data Model. W3C Recommendation“* https://www.w3.org/TR/2017/REC-annotation-model-20170223/ [letzter Zugriff 08.11.2018].