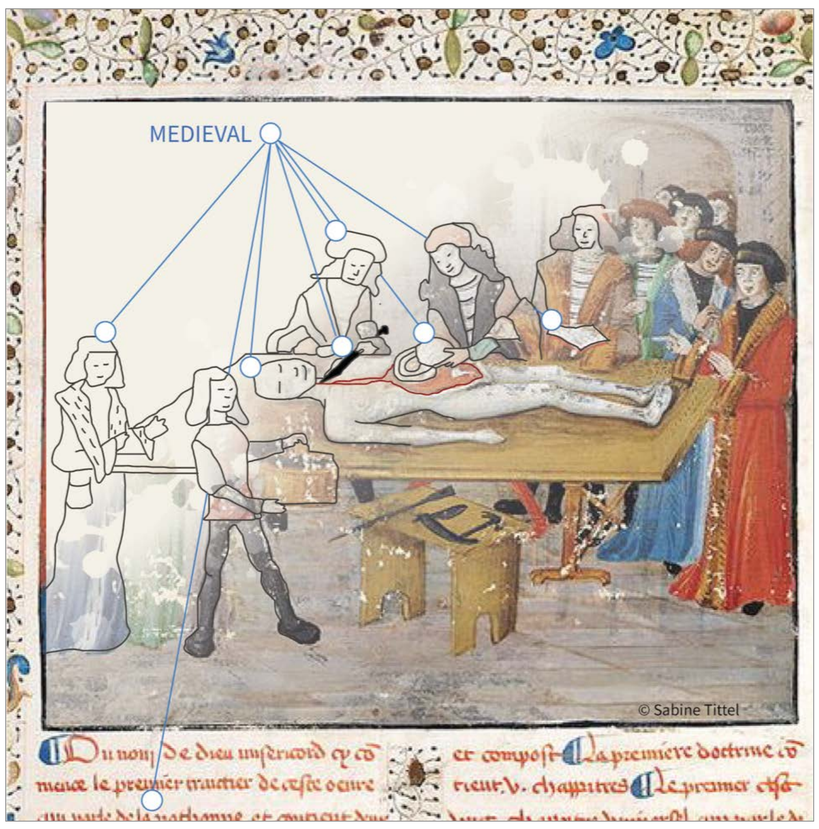
HEIDELBERGER AKADEMIE DER WISSENSCHAFTEN
Akademie der Wissenschaften des Landes Baden-Württemberg

Posterpräsentation im Rahmen des NDFI
Text+ Plenary, 12.-13.9.2022, Mannheim

Modellierung von Texteditionen als Linked Data
edition2LD

Dieta Svoboda-Baas – Sabine Tittel

Editionen als Linked Data



Wissenschaftliche Texteditionen stellen in Hinblick auf die sekundäre, übergreifende Auswertung und ihre Nachnutzbarkeit eine große **Herausforderung** dar:

- Unterschiedliche inhaltliche, zeitliche und geographische Ausrichtung.
- Unterschiedliche Sprachen, darüber hinaus historische Sprachstufen ohne Schreibnorm und schließlich unterschiedliche Schriftsysteme.
- Daten liegen in unterschiedlichen Systemarchitekturen und Datenmodellen (Insellösungen / Datensilos mit Zugänglichkeit im WWW, aber mit proprietärem Format und – davon abhängig – limitierten Zugangsmöglichkeiten).
- Während der Projektlaufzeit sind die erarbeiteten Daten (“hot data”) veränderlich: Daten-Updates und Versionierungen müssen bei der (Multi-Channel-)Publikation berücksichtigt werden. Erst die Langzeitarchivierung der Daten nach Projektabschluss (“cold data”) garantiert die Unveränderlichkeit der Daten.
- Unterschiedliche Kombinationen von Editionstext, Kommentar, Übersetzung, Faksimiles u.a. in der Präsentation der Forschungsergebnisse.
- Unterschiedliche DOI- und URI-Struktur externer Institutionen, die die Langzeitvorhaltung der Daten vornehmen.
- Abläufe und Zeitplanung der inhaltlichen Arbeiten in den Forschungsprojekten dürfen nicht beeinträchtigt werden.

Diese Probleme verlangen nach einer **Lösung für eine Datenkuratierung, die die heterogenen Daten über Ressourcen, Sprachen und Sprachstufen, Datentypen und -formate hinweg interoperabel und zugreifbar** macht. Diese muss zugleich stabil genug für eine langfristige Perspektive und zugleich flexibel genug für die Heterogenität der Daten sein. Darüber hinaus muss sie die Daten der eigenen Institution nahtlos in internationale Forschungskontexte und Zugriffsmöglichkeiten integrieren können. Für die Lösung folgen wir dem Paradigma von Linked Data und der Integration der Daten ins Semantic Web.

Projektidee

Die Idee von edition2LD ist es, anhand bestehender Projektdaten einen Workflow zu erarbeiten, der mit maximal automatisierten Prozessen Editionen in Form von RDF-Tripeln abbildet. Dieser Workflow soll generisch genug sein, um die zukünftige Übertragbarkeit auf weitere Projekte vorzubereiten. Zugleich soll er in der Lage sein, um – wiederholt angestoßen – Daten chargenweise in RDF zu überführen und damit auf die große Herausforderung der veränderlichen “hot data” zu reagieren. Bei der Erarbeitung der automatisierten Abbildungsprozesse ist es daher immens wichtig, den Schritt einer sicherlich in der einen oder anderen Weise nötigen, händischen Nachbearbeitung zu minimieren, im besten Fall soweit, dass er nur einmal, nämlich wenn die Daten keiner Veränderung mehr unterliegen, durchgeführt werden muss.

Ausgangsmaterialien

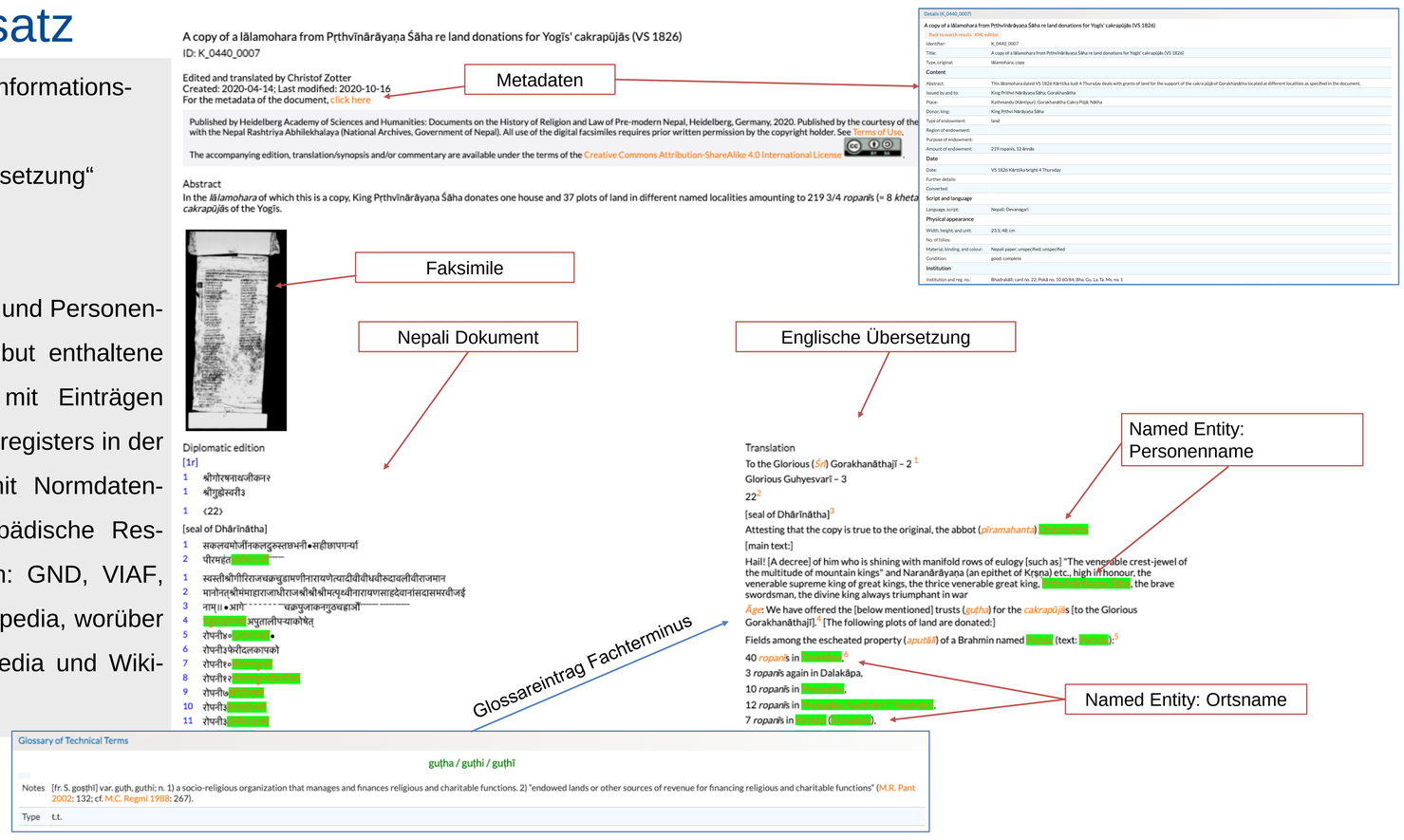
Ausgangsmaterialien könnten in diesem konkreten Fall Datenbestände der Forschungsstelle „*Religions- und rechtsgeschichtliche Quellen des vormodernen Nepal*“ der Heidelberger Akademie der Wissenschaften (HAdW) sein. Dieses Textkorpus vereint ein historisch äußerst wertvolles Material zur Tempel-, Verwaltungs- und Rechtsgeschichte der frühen Śāha- (1769–1846) und Rāṇā-Periode (1846–1951). Das Projekt wird seit 2014 im Akademienprogramm gefördert und hat eine Laufzeit bis 2028. Ziel des Forschungsprojekts ist es, dieses einzigartige Textkorpus in Editionen der Nepali-Texte inklusive Faksimile, englischer Übersetzung und Kommentar zu veröffentlichen. Die Editionen werden auf der Plattform ‘Documenta Nepalica’ und sukzessive unter Vergabe eines DOI zusätzlich in der Universitätsbibliothek Heidelberg veröffentlicht.

Einen durchgehend bereits auf Linked Data aufbauenden Ansatz verfolgt das Projekt „*Sprachdatenbasierte Modellierung von Wissensnetzen in der mittelalterlichen Romania – ALMA*“, das als interakademisches Projekt der HAdW, BAdW und der AdW Mainz am 1. August 2022 im Akademienprogramm gestartet ist. ALMA untersucht auf Basis von Textkorpora den Ausbau der romanischen Vernakularsprachen zu komplexen Wissen(schafts)sprachen. Das Projekt vereint Methoden der Linguistik, Textphilologie und Wissen(schafts)geschichte mit den Digital Humanities und des Ontology Engineerings; die Forschungsergebnisse werden im WWW und in Form von Linked Data und historisierten Domänenontologien zur Verfügung gestellt. Da ALMA Texteditionen (hier von mittelalterlichen Rechts- und Medizintexten) erarbeitet, lässt sich auch dieser Datenbestand gut für einen edition2LD-Ansatz einsetzen.

Inhaltlicher Ansatz

Wir fokussieren auf mehrere Informationseinheiten:

(1) Einheiten „Text“ und „Übersetzung“
(2) Metadaten
(3) Fachtermini
(4) Named Entities, z.B. Orts- und Personennamen. Über die im Attribut enthaltene Information können sie mit Einträgen eines Orts- bzw. Personenregisters in der Projektdatenbank oder mit Normdatenrepositorien und enzyklopädische Ressourcen verknüpft werden: GND, VIAF, GeoNames und engl. Wikipedia, worüber die Verknüpfung mit DBpedia und Wikidata erreicht wird.



Technischer Ansatz

Für die Umsetzung der LD-Modellierung einer Edition gibt es mehrere Herangehensweisen:

(1) Die Integration von RDF in die XML-Elemente der Editionen mittels RDFa (Herman et al. 2015): Informationen werden als Attribute in die bestehenden XML-Elemente eingeschrieben und mithilfe eines bestehenden Tools (z.B. RDFa 1.1 Distiller and Parser; Cimiano et al. 2020, 253-263) in Tripelform extrahiert. Die Integration der Informationen kann bspw. über Python durchgeführt werden.
(2) Die Tripel werden nicht in die Attribute der XML-Elemente integriert, sondern ausgelesen und in einen externen Datensatz geschrieben, entweder (a) über ein Skript in Python, XSLT o.Ä. oder (b) über den Einsatz eines bereits bestehenden Annotations- bzw. Crawling-Tools, z.B. INCEpTION oder XTriples. Im Falle des Einsatzes von CIDOC CRM kann WissKI eingesetzt werden.

Bibliographie:
Cimiano, Philipp, Christian Chiarcos, John P. McCrae und Jorge Gracia (2020). Linguistic Linked Data: Represen-tation, Generation and Applications. Cham (Springer).
Herman, Ivan, Aside, Ben, McCarron, Shane und Birbeck, Mark (2015). RDFa Core 1.1 – Third Edition. URL: https://www.w3.org/TR/rdfa-core [aufgerufen 14.03.2022]

Kontaktinformationen

Heidelberger Akademie der Wissenschaften
www.hadw-bw.de

Religions- und rechtsgeschichtliche Quellen des vormodernen Nepal
www.hadw-bw.de/nepal

Wissensnetze in der mittelalterlichen Romania (ALMA)
www.hadw-bw.de/alma

Die Projekte werden im Rahmen der gemeinsamen Forschungsförderung von Bund und Ländern im Akademienprogramm mit Mitteln des Bundesministeriums für Bildung und Forschung und des Ministeriums für Wissenschaft, Forschung und Kultur des Landes Baden-Württemberg gefördert.