Digitale Editionen historischer Reiseberichte öffnen:

Open Text und Open Data mit einheitlicher Textauszeichnung, semantischer Annotation und ontologiebasierter Datenmodellierung

Sandra Balck – balck@ios-regensburg.de
Ingo Frank – frank@ios-regensburg.de
Leibniz-Institut für Ost- und Südosteuropaforschung Regensburg

Einleitung

In unserem Projekt zur Edition historischer Reiseberichte wollen wir den Text für die wissenschaftliche Analyse und Visualisierung (von Zeit, Raum, Ereignissen, Beobachtungen) öffnen. Wir setzen dazu auf Textauszeichnung, semantische Annotation und ontologiebasierte Modellierung. Erprobt wird der Ansatz an der digitalen Edition des Reiseberichts Franz Xaver Bronners (1758 – 1850), der 1810 als Professor für theoretische Physik von Aarau in der Schweiz an die russische Universität Kasan an der Wolga ging und 1817 in die Schweiz zurückkehrte.

Problemstellung

Sowohl der De-facto-Standard TEI (Text Encoding Initiative) zur Textauszeichnung als auch das in den Digital Humanities etablierte CIDOC CRM (CIDOC Conceptual Reference Model) zur Datenmodellierung ermöglichen aufgrund ihrer Ausdrucksmächtigkeit ganz verschiedene Arten der Modellierung von Text und Daten, was zu Problemen mit der Interoperabilität und Nachnutzbarkeit führt. Wir müssen also eine Möglichkeit finden, sowohl die Ausdrücksmöglichkeiten von TEI als auch von CRM einzuschränken und damit zu vereinheitlichen, um dadurch die Interopabilität und Nachnutzbarkeit der digitalen Edition zu verbessern. Datenmodelle wie GeoJSON-T, Linked Places und Linked Traces [6] sind zwar bereits gut etabliert, aber wir sehen sie lediglich als Formate für den Datenaustausch und betrachten sie deshalb nicht als Option für die Datenmodellierung im Editionsprojekt.

Vorgehensweise

Um den Text interoperabler und leichter nachnutzbar zu machen – d. h. um die verschiedenen Textauszeichnungsvarianten mit TEI einzuschränken, verwenden wir nicht das volle TEI P5, sondern das DTA-Basisformat (DTABf) [7] zur Textauszeichnung. Auf Seite der Datenmodellierung verwenden wir CRM als Kernontologie. Um die verschiedenen Modellierungsmöglichkeiten mit CRM einzuschränken, setzen wir auf die Entwicklung von Ontologie-Entwurfsmustern mit OWL (Web Ontology Language) und SHACL (Shapes Constraint Language). Die Ontologie-Entwurfsmuster werden iterativ aufgebaut und zusammen mit SKOS-Vokabularen (Simple Knowledge Organisation System) zur Modellierung und Klassifikation von Reise(teil)ereignissen (wie Abreise und Ankunft) und Reisebeobachtungen (z. B. besuchte öffentliche Orte, Gewohnheiten von Personen) angewandt. Wir modellieren mit den Ontologie-Entwurfsmustern nicht die Erzählung als solche [2], sondern den rekonstruierten und stellenweise interpretierten Reiseverlauf als Repräsentation der Realität. Aus narratologischer Sicht machen wir mit dem ereigniszentrierten Modellierungsansatz von CRM also nur die Fabula (chronologische Reihenfolge der Ereignisse) eines Reiseberichts explizit. Das Sujet (Erzählreihenfolge) kann allerdings bei Bedarf anhand der annotierten Textstellen abgefragt und rekonstruiert werden.

Ontologie-Entwurfsmuster

Zur Erstellung des Annotationsschemas und der damit verbundenen Ontologie-Entwurfsmuster verwenden wir die Frame-Semantik als theoretisches Rahmenwerk. Die Ontologie-Entwurfsmuster dienen uns als ‚Schablonen' zum Anlegen an den Text – wobei deren Orientierung an den Frames sehr hilfreich ist – um Information über Reisedaten, Beobachtungen und Tätigkeiten unterwegs und während der Zwischenstopps zu erfassen. Im ständigen Austausch mit der Bearbeitung von Forschungsfragen am Text werden die Entwurfsmuster iterativ aufgebaut, getestet und ggf. angepasst [9]. Pragmatische Modellierungsentscheidung zum Aufbau der Ontologie DTO (Digital Editions of Historical Travelogues Ontology):

1. Erstellung von speziellen Klassen zur Modellierung von Reiseereignissen (`dto:Travel`, `dto:Stopover` sowie `dto:Transport`)
2. Verwendung allgemeinerer Klassen (als Unterklassen der CRM-Klasse `crm:E7_Activity`) zur Modellierung von Aktivitäten und deren Typisierung mit Begriffen aus SKOS-Klassifikationssystemen

Die Klassen (bzw. die Entwurfsmuster dafür) orientieren sich an FrameNet-Frames. Die Eigenschaften der Klassen basieren auf den entsprechenden Frame-Elementen und stellen somit semantische Rollen dar.

Text-und-Daten

Eide schlägt neben fünf weiteren Möglichkeiten vor, die in einem Text beschriebenen Ereignisse in CRM explizit zu modellieren und mit dem entsprechenden TEI-kodierten Dokument (bzw. einem bestimmten TEI-Element darin) zu verknüpfen [5]. Ein einfacher Ansatz, um die im TEI-Text ausgezeichneten Entitäten (Reisende, Abfahrtsort, Zielort, Straßen, Fahrzeuge usw.) mit den expliziten Datenbankeinträgen zu verknüpfen, ist die Verwendung des XPath-Selektors. Ein anderer Ansatz zur Verbindung von TEI- und RDF/XML-Daten wäre die Verwendung des W3C Annotation Data Model [3].

Visuelle-Annotation

Für die visuelle Annotation der n-ären Relationen von Ontologie-Entwurfsmustern können Annotationswerkzeuge wie brat [10] oder INCEpTION [8] verwendet werden (Abb. 4). Ein Vorteil der n-ären Struktur der Frame-basierten Entwurfsmuster ist deren Gebrauchstauglichkeit: sie hält den manuellen Annotationsprozess einfach. Die Annotation von Ereignissen auf Grundlage einer Ontologie mit umfassender rollenbasierter Modellierung wäre wesentlich umständlicher.

Lösungsvorschlag

Die Unterstützung von EARMARK (Extremely Annotational RDF Markup) [1] in der Editionsumgebung LEAF-Writer (https://leaf-writer.leaf-vre.org/) ermöglicht Standoff-Markup für die semantische Annotation der n-ären Beziehungen – verankert in einem TEI-kodierten Text [4]. Aus ontologischer Sicht können die explizit modellierten Ereignisse und deren Verknüpfung mit dem Text des Reiseberichts wie in Abb. 5 skizziert modelliert werden.

Ergebnisse

Wir lösen die Interoperabilitäts- und Ausdrucksprobleme von TEI und CRM mit Hilfe von DTABf und Frame-basierten Ontologie-Entwurfsmustern. Die Ontologie-Entwurfmuster dienen als Grundlage für die semantische Annotation von Reiseereignissen und -beobachtungen, wodurch die historischen Reiseberichte für weiterführende Analyse und Visualisierung geöffnet werden. Problematisch bei der semantischen Annotation ist, dass es oft schwierig ist, Ort, Zeit usw. explizit an Nennungen im Text festzumachen. Wir brauchen daher eine Möglichkeit, die Ereignisse trotzdem mit ggf. rekonstruierten bzw. interpretierten Angaben zu modellieren und nur lose einem Textabschnitt zuzuordnen.

Schlussfolgerung

Ontologie-Entwurfsmuster ermöglichen und erzwingen eine konsistente Anreicherung von Text mit Daten und helfen so, interoperable und nachnutzbare digitale Editionen von historischen Reiseberichten zu erstellen. SKOS-Vokabulare gewährleisten dabei während des gesamten ontologiebasierten semantischen Annotationsprozesses eine standardisierte Anwendung von Kategorien für die Kuratierung und Recherche. Angesichts der Fortschritte im Bereich bei Frame-Erkennung und Semantic Role Labeling könnte Frame-basierte semantische Annotation langfristig als Zwischenschicht in digitalen Editionsprojekten dienen.

Literatur

[1] Gioele Barabucci et al. "Annotations with EARMARK in Practice: A Fairy Tale". In: Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities. DH-CASE '13. Association for Computing Machinery, 2013. doi: 10.1145/2517978.2517990.
[2] Valentina Bartalesi, Carlo Meghini und Daniele Metilli. "A conceptualisation of narratives and its expression in the CRM". In: International Journal of Metadata, Semantics and Ontologies 12.1 (2017), S. 35–46.
[3] Marta Borriello et al. "From XML to RDF step by step: Approaches for Leveraging XML Workflows with Linked Data". In: XML Prague 2016 – Conference Proceedings. 2016.
[4] Fabio Cioi und Francesca Tomasi. "Formal Ontologies, Linked Data, and TEI Semantics". In: Journal of the Text Encoding Initiative 9 (Sep. 2016). doi: 10.4000/jtei.1480.
[5] Øyvind Eide. "Ontologies, Data Modeling, and TEI". In: Journal of the Text Encoding Initiative 8 (2015). doi: 10.4000/jtei.1191.
[6] Karl Grossner, Merrick Lex Berman und Rainer Simon. "Linked Places: A Modeling Pattern and Soware for Representing Historical Movement". In: Digital Humanities 2017: Conference Abstracts. 2017, S. 463–465. url: https://dh2017.adho.org/abstracts/204/204.pdf.
[7] Susanne Haaf, Alexander Geyken und Frank Wiegand. "The DTA Base Format: A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources". In: Journal of the Text Encoding Initiative 8 (2015). doi: 10.4000/jtei.1114.
[8] Jan-Christoph Klie et al. "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation". In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. Santa Fe, USA: Association for Computational Linguistics, Juni 2018, S. 5–9.
[9] Valentina Presui et al. "Paern-Based Ontology Design". In: Ontology Engineering in a Networked World. Hrsg. von Mari Carmen Suárez-Figueroa et al. Berlin: Springer, 2012, S. 35–64.
[10] Pontus Stenetorp et al. "brat: a Web-based Tool for NLP-Assisted Text Annotation". In: Proceedings of the Demonstrations Session at EACL 2012. Association for Computational Linguistics, 2012.

Knowledge-Graph-Extraction

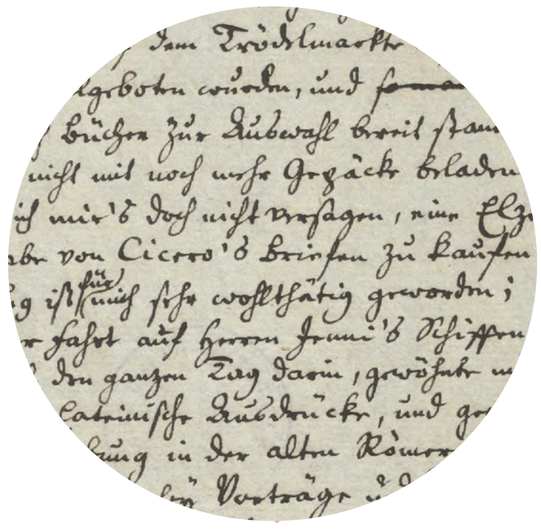

Abbildung 1: Textausschnitt für die Beispiel-Annotation

Frames als Rahmen zur Erfassung und semantischen Annotation von n-ären Relationen mit Ontologie-Entwurfsmustern für Reiseereignisse (Abb. 2) und Beobachtungen (Abb. 3):

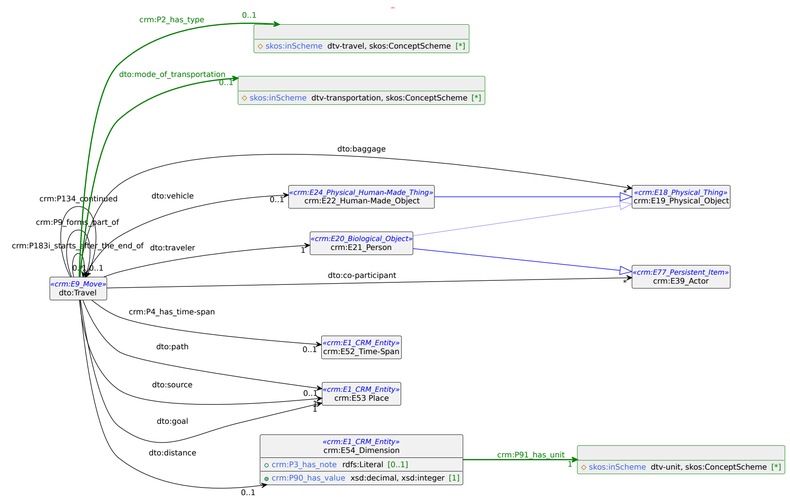

Abbildung 2: Klassendiagramm (OWL/SHACL) des auf dem FrameNet-Frame Travel basierenden Ontologie-Entwurfsmusters zur Modellierung von Reiseereignissen

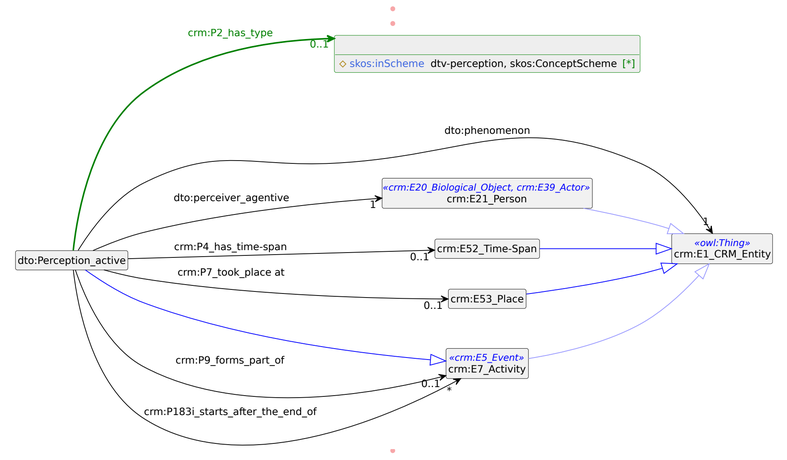

Abbildung 3: Klassendiagramm (OWL/SHACL) des auf dem FrameNet-Frame Perception_active basierenden Ontologie-Entwurfsmusters zur Modellierung von Reisebeobachtungen

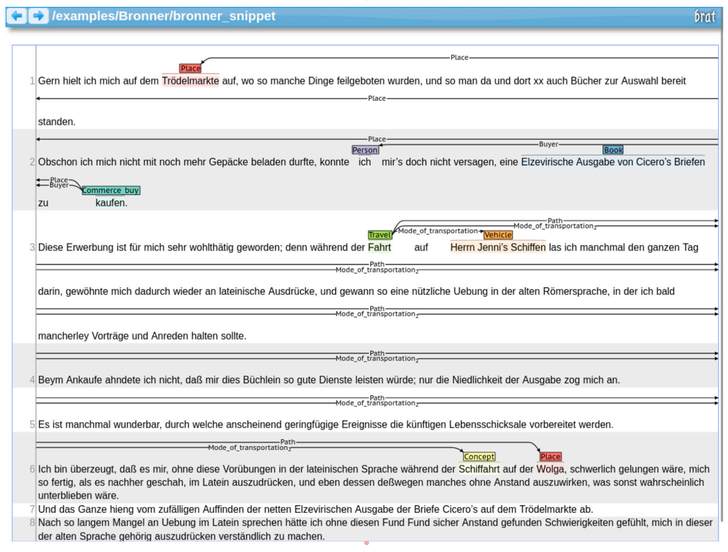

Abbildung 4: Semantische Annotation eines Reiseereignisses und einer Aktivität während eines Zwischenaufenthalts mit Frame-Strukturen im Annotationswerkzeug brat

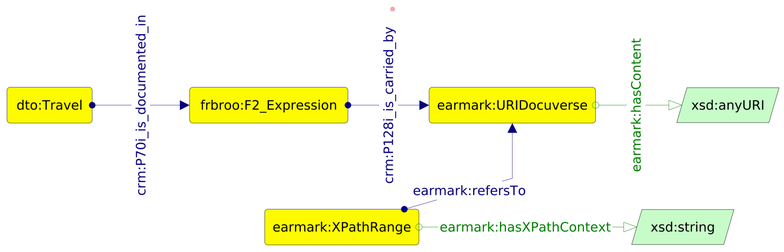

Abbildung 5: Lösungsvorschlag für ein Annotationsschema mit Standoff-Markup in EARMARK zur Verknüpfung von Text und Daten

Leibniz-Institut für OST- UND SÜDOST-EUROPAFORSCHUNG
Leibniz Gemeinschaft
DFG Deutsche Forschungsgemeinschaft

Digitale Editionen Historischer Reiseberichte (DEHisRe): https://dehisre.ios-regensburg.de/
Gefördert furch die DFG: https://gepris.dfg.de/gepris/projekt/453498639
Diagramme in Abb. 1 und Abb. 2 erstellt mit SHACL2PlantUML: https://github.com/rosecky/shacl2plantuml, Diagramm in Abb. 4 erstellt mit Graoo: https://essepuntato.it/graffoo/