Digitale Editionen von historischen Reiseberichten öffnen: Open Text und Open Data mit einheitlicher Textauszeichnung, semantischer Annotation und ontologiebasierter Datenmodellierung

Balck, Sandra
balck@ios-regensburg.de
IOS Regensburg, Deutschland

Frank, Ingo
frank@ios-regensburg.de
IOS Regensburg, Deutschland

Einleitung

Innerhalb der Literatur- und Kulturwissenschaften ist die historische Reiseforschung ein beliebtes Forschungsfeld, dennoch gibt es innerhalb der Digital Humanities wenige nennenswerte Fortschritte bei der Erschließung, Verarbeitung und Visualisierung von Reiseberichten. Konventionellen digitalen Editionen fehlt bisher die notwendige Ausdruckskraft und Flexibilität, um die diversen Anwendungsfälle und Forschungsfragen der historischen Reiseforschung zu erschließen. Anstatt einer starren Auszeichnung mit TEI zu folgen, müssen Informationen in digitalen Editionen erkannt, identifiziert, mit zusätzlichen Daten angereichert und Narrationen der Ereignisse explizit modelliert werden.[1]

Dieser Beitrag beschäftigt sich mit der Frage der Öffnung digitaler Reiseberichte für die wissenschaftliche Analyse und Visualisierung (von Zeit, Raum, Ereignissen u.a.) durch Textauszeichnung, semantische Annotation und ontologiebasierte Modellierung. Hierbei verfolgen wir einen disziplinübergreifenden, iterativen Ansatz, welcher sowohl geistes- als auch informationswissenschaftliche Perspektiven einbezieht.[2] Erprobt wird dieser 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

Datenmodellierung kann laut Flanders und Jannidis (vgl. 2015) in zwei Gruppen unterschieden werden: Curation-Driven und Research-Driven. Curation-driven beschreibt die Praktiken von Bibliotheken und Archiven, Objekte mit Hilfe von Standards einheitlich zu erfassen, um so die Auffindbarkeit und Transparenz zu gewährleisten. Die dafür notwendige Reduktion führt jedoch zu Ungenauigkeiten und Lücken im Datenmaterial. Research-Driven hingegen zielt auf die Beantwortung spezifischer Forschungsfragen ab und die Datenerfassung/Modellierung folgt einem konkreten Forschungsinteresse. Dabei werden nur selten Standards berücksichtigt. Dieser Gegensatz zwischen Forschungs- und Kuratierungspraxis erschwert die Kompatibilität und damit die Vergleichbarkeit der Daten. Um digitale Reiseberichte für wissenschaftliche Analysen verwertbar zu machen, müssen sie beiden Ansprüchen gerecht werden.

Im Bereich der digitalen Editionen haben sich die TEI-Guidelines zum De-facto-Standard entwickelt. “[TEI] is the most systematic effort so far to create standards for scholarly memory in an evolving digital culture.” (tei-c.org 2019) Die Guidelines beinhalten aktuell 585 Elemente und sind flexibel gestaltet, um für ein breites Spektrum von Forschungsfragen anwendbar zu sein. Ein Problem ist jedoch, dass dieselbe Information unterschiedlich kodiert und interpretiert werden kann (z. B. <rs>, <persName>) und der „Standard” damit nicht zwingend interoperabel ist (siehe Unsworth 2011; Burrows et al. 2021; Giovannetti und Tomasi 2022). Die TEI versucht die Brücke zwischen Standardisierung und Granularität zu schlagen, verliert damit aber ihre Eindeutigkeit (vgl. Kudella und Jefferies 2019).

Textauszeichnung und Ontologie-Entwurfsmuster

Einen Lösungsansatz für das Interoperabilitätsproblem auf Textseite bietet das DTA-Basisformat (Haaf et al. 2015), welches sich zum Ziel gesetzt hat: “[...] eine umfassende Textaufbereitung [zu] ermöglichen und dabei gleichzeitig Variationsspielräume bei der Annotation so ein[zu]schränken, dass die Kohärenz [...] untereinander gewährleistet wird.” (DTABf 2011-2020) Das DTABf richtet sich dabei nach den P5-Richtlinien der TEI. Um darüber hinaus auch spezifische Forschungsinteressen der historischen Reiseforschung zu adressieren, entwickeln wir auf Datenseite einen ontologiebasierten Textanreicherungs- und Bearbeitungsworkflow: Ontologie-Entwurfsmuster werden iterativ aufgebaut und für die Klassifizierung von Reise(teil)ereignissen (Abreise, Ankunft usw.) und Reisebeobachtungen (z. B. besuchte öffentliche Orte, Gewohnheiten von Personen) angewandt.

Während TEI für die Textauszeichnung verwendet wird, dient CRM zur Anreicherung des Textes mit explizitem Wissen, welches in einer Datenbank gespeichert wird. Wir modellieren mit den Ontologie-Entwurfsmustern nicht die Erzählung als solche[3] , 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.[4]

Zur Erstellung des Annotationsschemas und der damit verbundenen Ontologie-Entwurfsmuster wenden wir die Frame-Semantik als theoretischen Rahmen an. Frames können als n-äre Relationen[5] repräsentiert und daher zur Entwicklung von Ontologie-Entwurfsmustern für Reiseereignisse und -beobachtungen verwendet und darüber hinaus als "knowledge patterns" zur Validierung der Entwurfsmuster herangezogen werden (vgl. Presutti et al. 2012).[6] Die Ontology Design Patterns dienen uns als „Schablonen" zum Anlegen an den Text – wobei deren Orientierung an den Frames sehr hilfreich ist – um Reisedaten, Beobachtungen und Tätigkeiten unterwegs und während Zwischenstopps zu erfassen. Im ständigen Austausch mit der Bearbeitung von Forschungsfragen am Text werden die Frames und Design Patterns laufend überprüft und angepasst. Diese Form der forschungsgeleiteten Standardisierung (mittels expliziter und einheitlicher Modellierung) macht digitale historische Reiseberichte interoperabel und öffnet sie damit für vergleichende Analysen. Die möglichen Ansätze zur Verknüpfung von TEI-kodiertem Text und RDF-Daten mittels semantischer Annotation evaluieren wir (vgl. Eide 2015 und Borriello et al. 2016) und stellen im Poster Lösungswege mit EARMARK (Barabucci et al. 2013) und NIF (Hellmann et al. 2013) vor.

Verwandte Arbeiten und Schlussfolgerung

Es gibt einige Beispiele wie das Hellespont-Projekt (Mambrini 2016) oder die Semantic Blumenbach-Edition (Wettlaufer 2015), die TEI-kodierten Text und CRM-modellierte Daten miteinander verknüpfen. Unser Ansatz geht jedoch über die bestehenden Projekte hinaus, da wir Ontologie-Entwurfsmuster für eine explizitere Datenmodellierung entwickeln und anwenden.[7] Kurz gesagt, wir lösen die Interoperabilitäts- und Ausdrucksprobleme von TEI und CRM mit Hilfe von DTABf und Frame-basierten Ontologie-Entwurfsmustern, was wiederum die Kategorien von Reiseereignissen und Reisebeobachtungen in historischen Reiseberichten für die weitere Analyse und Visualisierung explizit macht.

Fußnoten

1. Bei der Erstellung der digitalen Edition folgen wir dem Historical Information Life Cycle (Meroño-Peñuela et al. 2014), wobei wir Ontologien nicht nur in der Anreicherungs-, Bearbeitungs- und Retrieval-Phase des Lebenszyklus, sondern auch in der Analyse- und Visualisierungsphase einsetzen.
2. Unser Ontologie-Entwurfsansatz orientiert sich an der eXtreme Design-Vorgehensweise (Presutti et al. 2009), bei der sog. Competeny Questions aus anfänglichen User Stories abgeleitet werden, um die Anforderungen an die Datenmodellierung und das Information Retrieval zu definieren.
3. siehe hierzu den Modellierungsansatz von Bartalesi et al. 2017
4. Das kann interessant für Analyse und Visualisierung sein, weil, wie Maurer (2015, S. 391 f.) anmerkt, Reiseberichte von wissenschaftlichen Forschern oft einer thematischen Organisation folgen, um „Reiseergebnisse" zu präsentieren, anstatt einfach einer chronologischen Reihenfolge zu folgen.
5. Annotationswerkzeuge wie brat (Stenetorp et al. 2012) oder INCEpTION (Klie et al. 2018) bieten Annotations- und Visualisierungskomponenten für n-äre Relationen (Frames oder Ereignisse inkl. der Rollen von Akteuren).
6. Siehe dazu etwa das allgemeine Frame *Travel* mit Frame-spezifischen Eigenschaften zur Beschreibung von reisender Person, Reiseziel, Verkehrsmittel usw. in FrameNet: http://framenet.lexicalsemantics.org/frameIndex.xml?frame=Travel
7. Die Datenmodelle GeoJSON-T, Linked Places und Linked Traces (siehe Grossner et al. 2017) sind bereits gut etabliert, aber es handelt sich dabei nur um Formate für den Datenaustausch und kommen daher in unserem Projekt nicht für den Aufbau der Datenbank in Frage.

Bibliographie

Barabucci, Gioele, Angelo Di Iorio, Silvio Peroni, Francesco Poggi, and Fabio Vitali. 2013. "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, 10.1145/2517978.2517990

Bartalesi, Valentina, Carlo Meghini, and Daniele Metilli. 2017. "A Conceptualisation of Narratives and Its Expression in the Crm." In *International Journal of Metadata, Semantics and Ontologies* 12 (1): 35–46.

Burrows, Toby, Matthew Holford, David Lewis, Andrew Morrison, Kevin Page, and Athanasios Velios. 2021. "Transforming Tei Manuscript Descriptions into Rdf Graphs." In *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing* , 143–54. Norderstedt: BoD, http://www.digitalhumanities.org/dhq/vol/16/2/000605/000605.html

Eide, Øyvind. 2015. "Ontologies, Data Modeling, and TEI." In *Journal of the Text Encoding Initiative* , no. 8 (December), https://doi.org/10.4000/jtei.1191

Füssel, Marian, Tim Neu. 2021. "Akteur-Netzwerk-Theorie und Geschichtswissenschaft". Paderborn: Brill; Ferdinand Schöningh.

Giovannetti, Francesca; Tomasi, Francesca. 2022. "Linked data from TEI (LIFT): A Teaching Tool for TEI to Linked Data Transformation" In *Digital Humanities Quarterly* 16 (2). http://www.digitalhumanities.org/dhq/vol/16/2/000605/000605.html.

Grossner, Karl, Merrick Lex Berman, and Rainer Simon. 2017. "Linked Places: A Modeling Pattern and Software for Representing Historical Movement." In *Digital*

Humanities 2017: Conference Abstracts , 463–65, https://dh2017.adho.org/abstracts/204/204.pdf

Haaf, Susanne, Alexander Geyken, and Frank Wiegand. 2015. "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* , no. 8, doi:10.4000/jtei.1114

Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. "Integrating NLP using Linked Data". In: 12th International Semantic Web Conference, 21–25 October 2013, Sydney, Australia.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation". In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018),* Santa Fe, New Mexico, USA. https://aclanthology.org/C18-2002

Kudella, Christoph, and Neil Jefferies. 2019. "How Do We Model the Republic of Letters?" In *Reassembling the Republic of Letters in the Digital Age* , edited by Howard Hotson and Thomas Wallnig, 41–53. Göttingen: Göttingen University Press, 10.17875/gup2019-1146

Mambrini, Francesco. 2016. "Treebanking in the World of Thucydides. Linguistic Annotation for the Hellespont Project." In *Digital Humanities Quarterly* 10 (2), http://www.digitalhumanities.org/dhq/vol/10/2/000251/000251.html

Maurer, Michael. 2015. "Reiseberichte als Wissensspeicher." In *Wissensspeicher Der Frühen Neuzeit: Formen und Funktionen* , edited by Frank Grunert and Anette Syndikus, 391–412. Berlin, Boston: De Gruyter, 10.1515/9783050086637-015

Meroño-Peñuela, Albert, Ashkan Ashkpour, Marieke van Erp, Kees Mandemakers, Leene Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank van Harmelen. 2015. "Semantic Technologies for Historical Research: A Survey." *Semantic Web* 6 (6). IOS Press: 539–64, 10.3233/SW-140158

Presutti, Valentina, Enrico Daga, Aldo Gangemi, and Eva Blomqvist. 2009. "eXtreme Design with Content Ontology Design Patterns." WOP.

Presutti, Valentina, Eva Blomqvist, Enrico Daga, and Aldo Gangemi. 2012. "Pattern-Based Ontology Design." In *Ontology Engineering in a Networked World* , edited by Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, Enrico Motta, and Aldo Gangemi, 35–64. Berlin: Springer.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topi#, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. "Brat: a Web-based Tool for NLP-Assisted Text Annotation". In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–107.

tei-c.org. 2019. "A very gentle introduction to the TEI markup language." https://tei-c.org/Vault/Tutorials/mueller-index.htm.

Unsworth, John. 2011. "Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the Tei." In *Journal of the Text Encoding Initiative* , no. 1 (June), 10.4000/jtei.215

Wettlaufer, Jörg, Christopher Johnson, Martin Scholz, Mark Fichtner, and Sree Ganesh Thotempudi. 2015. "Semantic Blumenbach: Exploration of Text–Object Relationships with Semantic Web Technology in the History of Science." In *Digital Scholarship in the Humanities* 30 (suppl_1): i187–i198, 10.1093/llc/fqv047