Linked Open Data für die Literaturgeschichtsschreibung

Das Projekt "Mining and Modeling Text"

Hinzmann, Maria
hinzmannm@uni-trier.de
Universität Trier, Germany

Schöch, Christof
schoech@uni-trier.de
Universität Trier, Germany

Dietz, Katharina
dietz@uni-trier.de
Universität Trier, Germany

Klee, Anne
klee@uni-trier.de
Universität Trier, Germany

Erler-Fridgen, Katharina
erler@uni-trier.de
Universität Trier, Germany

Röttgermann, Julia
roettger@uni-trier.de
Universität Trier, Germany

Steffes, Moritz
steffesm@uni-trier.de
Universität Trier, Germany

Zielsetzung und Projektstruktur

Im Umgang mit dem stetig wachsenden ‚digitalen Kulturerbe' bietet die Weiterentwicklung der systematischen Datenerschließung und Wissensrepräsentation bisher nicht ausgeschöpfte Potentiale für die Literaturgeschichtsschreibung. Vor diesem Hintergrund werden im Projekt *Mining and Modeling Text (MiMoText)* quantitative Methoden der Informationsextraktion (‚Mining') und Datenmodellierung (‚Modeling') ineinander verschränkt, um ein literaturgeschichtliches Wissensnetzwerk aufzubauen (vgl. Abb. 1).[1] Der Fokus liegt zunächst auf französischen Romanen (1751–1800). Die Übertragbarkeit in andere Domänen wird berücksichtigt. Zentrales Anliegen ist es, den Bereich der quantitativen Methoden zur Extraktion, Modellierung und Analyse geisteswissenschaftlich relevanter Informationen aus umfangreichen Textsammlungen weiterzuentwickeln und aus interdisziplinärer (geistes-, informatik- und rechtswissenschaftlicher) Perspektive zu erforschen.

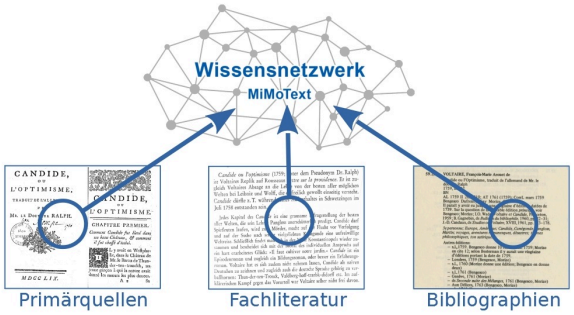

Abb. 1: Informationsquellen des Wissensnetzwerks.

Bezogen auf den Gegenstandsbereich ist ein Ausgangspunkt, dass die über rund zwei Jahrhunderte akkumulierten literaturhistorischen Forschungserkenntnisse größtenteils nicht unmittelbar nutzbar sind, da diese sehr umfangreich sind, nicht digital vorliegen oder auf unterschiedliche Orte und Quellen verteilt und in unterschiedlichen Sprachen publiziert sind. *MiMoText* begegnet diesem Desiderat, indem es Informationen aus drei unterschiedlichen Quellentypen im Aufbau des fachspezifischen Wissensnetzwerks miteinander verknüpft: Metadaten aus Nachweissystemen, Texteigenschaften aus Primärtexten, Sachinformationen aus Fachliteratur. In vier Teilbereichen (Research Areas/RAs) werden Lösungen für die Informationsextraktion, ihre Modellierung, die rechtlichen Rahmenbedingungen sowie die Infrastruktur erarbeitet (vgl. Abb. 2).

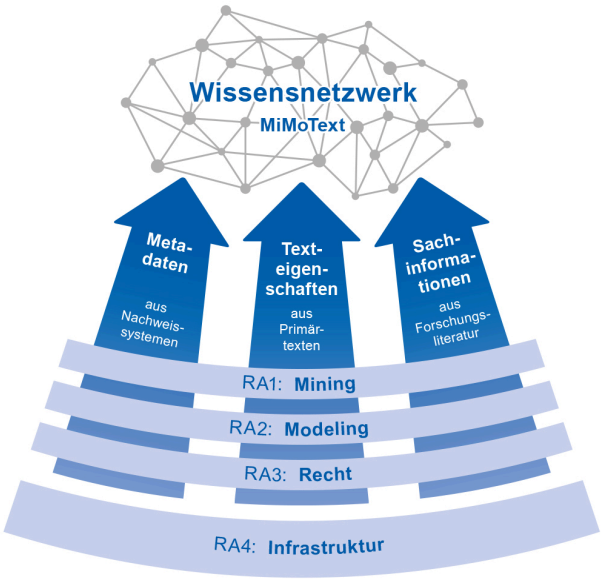

Abb. 2: Projektübersichtsgrafik.

Open Science-Prinzipien

Für die Bereitstellung der Daten sowie die Arbeit und Infrastruktur im Projekt sind Open Science-Prinzipien tragend. Dies betrifft u.a. die Veröffentlichung FAIRer Daten (Röttgermann/Schöch 2020) im Open Access (Schöch 2021), die Nutzung von Open Source-Tools wie *INCEpTION* (Klie et al. 2018)

und *OCR4all* (Reul et al. 2019) sowie Wikibase als Infrastruktur, die dem Linked Open Data (LOD) folgt.[2] Im rechtswissenschaftlichen Teilbereich werden die rechtlichen Rahmenbedingungen von Text und Data-Mining in den Geisteswissenschaften auch im Hinblick auf eine offenere und nachhaltigere Nutzung von Forschungsdaten erforscht (vgl. Erler-Fridgen 2021a; Erler-Fridgen 2021b; Erler-Fridgen 2021c; Raue/Schöch 2020; Schöch et al. 2020).

Mining – Extrahieren von Informationen aus drei Quellentypen

Bibliographische Daten

Für unsere Domäne ist die 1977 von Mylne, Martin und Frautschi veröffentlichte *Bibliographie du genre romanesque français 1751-1800* (BGRF) zentral, da sie die Grundgesamtheit der Romane definiert. Die BGRF wurde von Andreas Lüschow aufwendig erschlossen und nach aktuellen bibliographischen Standards modelliert (Lüschow 2020).

Primärliteratur

Im Teilprojekt zur Primärliteratur wird schrittweise ein Korpus von etwa 200 Romanen aufgebaut, wovon bereits reichlich 100 Texte in XML-TEI verfügbar sind (vgl. Röttgermann 2021; Klee/Röttgermann 2020).[3] Eine Reihe von Analyseverfahren wurde bereits auf diesen Textbestand angewandt, darunter Topic Modeling (vgl. Klee/Röttgermann 2020), NER (Orte und Figuren) sowie explorativ Sentiment Analyse.[4]

Sekundärliteratur

Das mittelfristige Ziel besteht darin, durch überwachtes Lernen die automatische Extraktion von Aussagen (RDF-Tripel) zu ermöglichen. Um Trainingsdaten zu generieren und Tripel in das Wissensnetzwerk einspeisen zu können, werden aktuell literaturgeschichtliche Texte in INCEpTION annotiert. Die Daten sollen als Statements über eine noch zu entwickelnde toolübergreifende Pipeline in unsere projektspezifische Wikibase importiert werden.[5]

Modeling – Repräsentation und Vernetzung von Wissen

Im Aufbau des LOD-Wissensnetzwerks werden literaturgeschichtliche Aussagetypen in einer systematischen Ontologie modelliert und die extrahierten Informationen als RDF-Tripel repräsentiert (vgl. Abb. 3). Der Mehrwert des Netzwerks wird durch exemplarische Nutzungsszenarien in Form von SPARQL-Abfragen konkretisiert. Frageoptionen wie „Tritt Thema x in einem bestimmten Zeitraum y gehäuft auf?“ verdeutlichen, welcher Nutzen daraus für eine datenbasierte Literaturgeschichtsschreibung entstehen kann. Exemplarisch werden Einblicke in die Infrastruktur gegeben (vgl. Abb. 4).

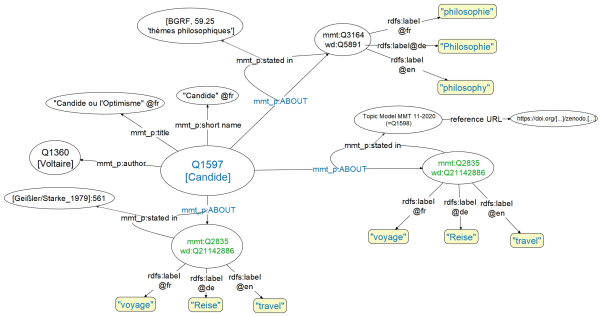

Abb. 3: Thematische Aussagen (Property „about“) über das Werk *Candide* aus den drei unterschiedlichen Informationsquellen.

Domänenspezifische Herausforderungen & Chancen

Ein Standardisierungsprozess wie er sich beispielsweise im *CIDOC CRM* (http://www.cidoc-crm.org) niederschlägt, steht für die Domäne der Literaturgeschichte noch am Anfang. Das Projekt konzentriert sich auf Aussagen über literarische Werke und Autor:innen, bisher vor allem thematische Aussagen (vgl. Schöch et al. 2022) sowie Handlungsorte (vgl. Hinzmann et al. angenommen). Dabei ist das beständig wachsende Wissensnetzwerk in der Zusammenführung von Informationen aus verschiedenen, auch widersprüchlichen Quellen möglichst offen im Hinblick auf unterschiedliche Nutzungsszenarien und Fragerichtungen. Durch die Menge der aggregierten Daten lassen sich literaturhistorische Annahmen bestätigen, revidieren oder präzisieren und neue Fragestellungen sowie eine Metaperspektive auf den literaturwissenschaftlichen Diskurs entwickeln.

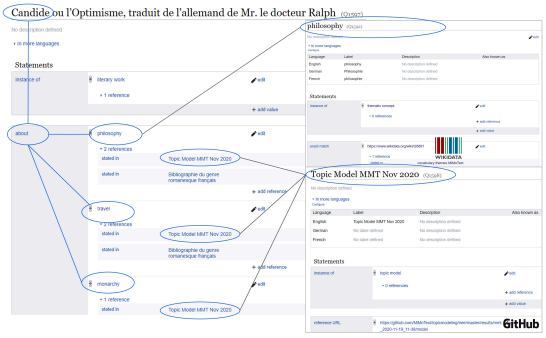

Abb. 4: Screenshot von thematischen Aussagen zu *Candide* in der MiMo-Text-Wikibase.

Präsentationsstrategie

Das Poster veranschaulicht die Integration der verschiedenen Informationsquellen in der Datenmodellierung sowie das Zusammenspiel der verschiedenen Teilprojekte und Tools im Aufbau und in möglichen Nutzungsszenarien des mehrsprachigen Wissensnetzwerks.

Fußnoten

1. Das Projekt wird im Rahmen der Forschungsinitiative Rheinland-Pfalz gefördert und vom *Trier Center for Digital Humanities* koordiniert, vgl. https://mimotext.uni-trier.de.
2. Vgl. zum Datenmodell von Wikibase/Wikidata https://www.mediawiki.org/wiki/Wikibase/DataModel sowie https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format.
3. Vgl. genauer zum Korpusaufbau Röttgermann (angenommen) und zu den Kodierungsprinzipien, die der European Literary Text Collection (ELTeC) folgen: Burnard et al. 2021.
4. Vgl. das GitHub-Repository unter: https://github.com/MiMoText/roman18.
5. Vgl. zum Statement-Begriff in Wikibase/Wikidata https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format sowie https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer#Statements.

Bibliographie

Burnard, Lou / Schöch, Christof / Odebrecht, Carolin (2021): "In Search of Comity: TEI for Distant Reading", in: *Journal of the Text Encoding Initiative* 14 https://doi.org/10.4000/jtei.3500.

Erler-Fridgen, Katharina (2021a): "Die Nutzung wissenschaftlicher Ausgaben für Textanalysen", in: *IRDT PaperSeries 1*. https://irdt.uni-trier.de/die-nutzung-wissenschaftlicher-ausgaben-fuer-textanalysen/ [letzter Zugriff 30. November 2021].

Erler-Fridgen, Katharina (2021b): "Kriterien der urheberrechtlichen Schutzfähigkeit von Texten und Sammelwerken", in: *IRDT PaperSeries 2*. https://irdt.uni-trier.de/kriterien-der-urheberrechtlichen-schutzfaehigkeit-von-texten-und-sammelwerken/ [letzter Zugriff 30. November 2021].

Erler-Fridgen, Katharina (2021c): "Die Präsentation von Textteilen als Ergänzung von Textanalysen", in: *IRDT PaperSeries 3*. https://irdt.uni-trier.de/die-praesentation-von-textteilen-als-ergaenzung-von-textanalysen/ [letzter Zugriff 30. November 2021].

Hinzmann, Maria / Röttgermann, Julia / Klee, Anne / Steffes, Moritz / Schöch, Christof (angenommen): "The French Enlightenment Novel as a Graph? Potentials and Challenges in the Construction of a Knowledge Network", angenommener Beitrag *Graphs and Networks in the Humanities 2022: Knowledge Graphs and Reasoning – Promises, Potentials, and Pitfalls*.

Klee, Anne / Röttgermann, Julia (2020): *Doing Topic Modeling on French 18th Century Novels in the Context of MiMoText Project* [data set] https://github.com/MiMoText/topicmodeling [letzter Zugriff 30. November 2021].

Klie, Jan-Christoph / Bugert, Michael / Boullosa, Beto / Eckart de Castilho, Richard / Gurevych, Iryna (2018): "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation", in: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* 5–9 http://tubiblio.ulb.tu-darmstadt.de/106270/ [letzter Zugriff 30. November 2021].

Lüschow, Andreas (2020): "Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane", in: *Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*, 80–84 10.5281/zenodo.3666690.

Raue, Benjamin / Schöch, Christof (2020): "Zugang zu großen Textkorpora des 20. und 21. Jahrhunderts mit Hilfe abgeleiteter Textformate – Versöhnung von Urheberrecht und textbasierter Forschung", in: *RuZ – Recht und Zugang*, 1.2.: 118–27 10.5771/2699-1284-2020-2-118.

Reul, Christian / Christ, Dennis / Hartelt, Alexander / Balbach, Nico / Wehner, Maximilian / Springmann, Uwe / Wick, Christoph / Grundig, Christine / Büttner, Andreas / Puppe, Frank (2019): "OCR4all -- An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings", in: *ArXiv:1909.04032* [Cs].

Röttgermann, Julia / Schöch, Christof (2020): "FAIRe Daten in den Literaturwissenschaften? Das Beispiel „Mining and Modeling Text“ und der französische Roman des 18. Jahrhunderts", in: *Romanistik-Blog. Blog des Fachinformationsdienstes* https://blog.fid-romanistik.de/2020/11/05/faire-daten-in-den-literaturwissenschaften/ [letzter Zugriff 30. November 2021].

Röttgermann, Julia (ed.) (2021): *Collection de romans français du dix-huitième siècle (1750-1800) / Eighteenth-Century French Novels (1750-1800)* [data set]. Release v0.2.0 10.5281/zenodo.5040855.

Röttgermann, Julia (angenommen): "Établissement d’un corpus de romans français du XVIIIe siècle dans le cadre du projet Mining and Modeling Text" [angenommener Beitrag Frankoromanistentag 2021, Sektion: *Digital, global, transdisziplinär: Impulse für eine transdisziplinäre digitale Romanistik*].

Schöch, Christof / Döhl, Frédéric / Rettinger, Achim / Gius, Evelyn / Trilcke, Peer / Leinen, Peter /Jannidis, Fotis / Hinzmann, Maria / Röpke, Jörg (2020): "Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen", in: *Zeitschrift für digitale Geisteswissenschaften – ZfdG* 10.17175/2020_006.

Schöch, Christof (2021): "Open Access für die Maschinen", in: Kohle, Hubertus / Effinger, Maria (eds.): *Die Zukunft des kunsthistorischen Publizierens*. Heidelberg: arthistoricum.net 79–94 10.11588/arthistoricum.663.c9210.

Schöch, Christof / Hinzmann, Maria / Röttgermann, Julia / Dietz, Katharina / Klee, Anne (angenommen): "Smart Modeling For Literary History", angenommen in: *IJHAC. Linked Open Data in the Arts and Humanities* [special issue] (March 2022).