In sechs Stationen rund um MiMoText: Einblicke in das Projekt „Mining and Modeling Text“

Katharina Dietz (Universität Trier), Katharina Erler-Fridgen (Universität Trier), Maria Hinzmann (Universität Trier), Anne Klee (Universität Trier), Julia Röttgermann (Universität Trier), Moritz Steffes (Universität Trier), Christof Schöch (Universität Trier)

6 Stationen und ein „roter Faden“

In sechs Stationen möchten wir den Teilnehmenden Einblicke in das Projekt „MiMoText“ des Trier Center for Digital Humanities geben und einen virtuellen Raum entstehen lassen, indem wir unsere Ziele, Ansätze und Zwischenergebnisse exemplarisch vorstellen, diskutieren und Feedback aus der Community gewinnen. In dem interaktiven Format, welches unsere Forschungsbereiche und Teilprojekte in einzelne Stationen aufgliedert, verschränken wir Projektvorstellung, Dialograum und die Sammlung von neuen Ideen ineinander. Das Format ist so experimentell wie unser Projekt: Die Einzigartigkeit entsteht aus dem Zusammenspiel der verschiedenen Stationen und ihren Verbindungslinien. Als „roter Faden“ dient die Darstellung der Ergebnisse, die in einem Pilotprojekt gewonnen wurden.

Projektvorstellung

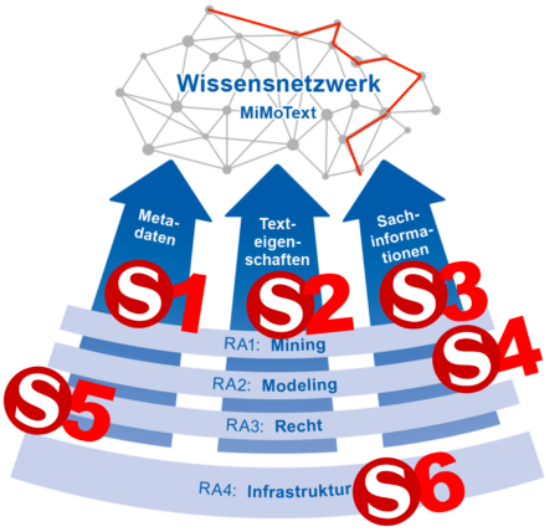
Ziel des Projektes „MiMoText“ ist es, den Bereich der quantitativen Methoden zur Extraktion, Modellierung und Analyse geisteswissenschaftlich relevanter Informationen aus umfangreichen Textsammlungen konsequent weiterzuentwickeln und aus interdisziplinärer (geistes-, informatik- und rechtswissenschaftlicher) Perspektive zu erforschen. Die primäre Anwendungsdomäne ist zunächst die französische Literaturgeschichte der zweiten Hälfte des 18. Jahrhunderts, Übertragungen auf weitere geisteswissenschaftliche Domänen sind geplant. Das innovative, interdisziplinäre Projekt gliedert sich in mehrere Teilbereiche und arbeitet mit verschiedenen Informationsquellen: Metadaten aus Nachweissystemen, Texteigenschaften aus Primärtexten, Sachinformationen aus Forschungsliteratur. Das Team setzt sich aus Forscher:innen aus der Informatik, Computerlinguistik, Romanistik, den Digital Humanities und der Rechtswissenschaft zusammen. Eine auf Text Mining und Datenmodellierung basierende Literaturgeschichtsschreibung kann in dieser interdisziplinären Form an sich schon als Experiment im Sinne des Tagungsmottos verstanden werden. Experimentell ist auch der Raum angelegt, den wir über verschiedene Stationen interaktiv entstehen lassen möchten: Mit einem exemplarischen „roten Faden“, der durch unser Wissensnetz wie auch durch den virtuellen Raum verläuft, werden wir an allen Stationen auf „thematische Aussagen“ eingehen, auf die wir uns im Rahmen eines Pilotprojekts fokussiert haben.

Stationenkonzept

Unsere Idee ist es, einen virtuellen Raum in Form eines Wonder-Raums anzubieten, in dem an sechs verschiedenen Stationen unsere MiMoText-Forschungsbereiche sowie -Teilprojekte präsentiert werden: Mining (Romane, Sekundärliteratur, Bibliographie), Modeling, Recht und Infrastruktur. Wir planen an jeder Station ein kurzes Video (ca. 3 Min) zur Präsentation der Teilbereiche. Außerdem werden die entsprechenden Teammitglieder zum virtuellen Austausch bereitstehen. In Wonder bewegt man sich als eine Art Avatar mithilfe der Maus oder der Pfeiltasten in einem virtuellen Raum und kann sich allen Personen entsprechend nähern. Bei einem nahen Aufeinandertreffen aktivieren

sich Mikrofon/Kamera, sodass in einen Gesprächsdialog getreten werden kann. Neben der Kommunikation mit den Mitarbeiter:innen der Forschungsteilbereiche können sich selbstverständlich auch weitere Kreise bilden, in denen – ähnlich wie auf einer analogen Konferenz – Teilnehmende unabhängig vom Input der sechs Stationen miteinander in den Austausch kommen können. Die URL zu den via Panopto veröffentlichten Videos verlinken wir in einer entsprechenden Ankündigung auf unserer Homepage, sodass der Austausch zwar auf den vorgesehenen Slot fokussiert, aber nicht darauf beschränkt ist. Ziel des Formats ist es, die Inhalte von „Mining and Modeling Text“ in seinen Teilbereichen darzustellen und in einen Dialog mit den Besucher:innen der vDHd zu treten. Der virtuelle Raum versucht somit, gewohnte Formate von Präsenzveranstaltungen mit Informationsstationen und Austauschmöglichkeiten zu simulieren.

Stationeninhalte



MiMoText Stationen

Station 1: Mining – Bibliographie

Die 1977 erschienene „Bibliographie du genre romanesque français 1751–1800“ enthält rund 2600 Einträge und repräsentiert die Grundgesamtheit der Romane, die Gegenstand unserer Untersuchung sind (Mylne, Frautschi und Martin 1977). Die hier dokumentierten Autor:innen und Werke mit ihren grundlegenden Metadaten bilden damit den Anker für das Wissensnetzwerk. Da zudem für zahlreiche (aber nicht alle) Romane eine inhaltliche Verschlagwortung (u.a. zu Erzählform, Handlungsort, Protagonisten, Inhalt der Handlung und Themen oder Stil) vorliegt, können auch weitergehende Aussagen extrahiert werden und mit den aus anderen Quellen erhobenen Informationen verglichen werden. Die Bibliographie wurde von Andreas Lüschow unter Nutzung u.a. der SPAR-Ontologien in RDF modelliert (vgl. Lüschow 2019). Station 3 illustriert die Inhalte der Bibliographie und ihre Nutzung im Projekt anhand eines Beispiels aus dem Pilotprojekt.

Station 2: Mining – Romane

Der Teilbereich von „Mining and Modeling Text“, der sich mit der Informationsextraktion aus Primärtexten auseinandersetzt, stellt seinen Forschungsstand vor: Ein digitales Korpus aus französischen Romanen 1750–1800 befindet sich im Aufbau. Die Texte stammen dabei aus verschiedenen Quellen. Neben eigener Digitalisierung mittels Double-Keying-Verfahren und dem

Einsatz von OCR-Software werden verfügbare Dateien aus dem Internet zusammengetragen. Alle Textdaten werden in TEI-konformes XML überführt. In einem ersten größeren Analyseschritt wurden mithilfe von Topic Modeling in dem Romankorpus vorkommende Themen ermittelt und als thematische Statements in Form von RDF-Tripeln extrahiert.

Station 3: Mining – Sekundärliteratur

Betrachtet wird Sekundärliteratur aus dem 19. und 20. Jahrhundert. Aus den Sekundärwerken werden Informationen extrahiert über besprochene Werke, Autoren etc. Als nächster Schritt nach der reinen Erfassung von Erwähnungen in Form von Named Entities sollen außerdem Aussagen über die Werke erfasst werden. Teilweise werden diese Informationen bereits der Bibliographie entnommen, wie Autor, Werk, Gattung, Erzählperspektive. Aussagen, die modelliert werden, sind z.B. „spricht über“ oder Wertungen. Dafür werden die extrahierten Aussagen als Linked Open Data modelliert. Dies soll später das maschinelle Auffinden von textuellen Zusammenhängen ermöglichen. Station 2 illustriert die Extraktion und Modellierung als Linked Open Data anhand eines Beispiels aus der Sekundärliteratur.

Station 4: Modeling

Die in den ersten 3 Stationen erhobenen Daten aus heterogenen Informationsquellen werden in einer Art „Wikidata für die Literaturgeschichte“ zusammengeführt. Station 4 verdeutlicht die Bedeutung des Linked Open Data-Paradigmas für die Modellierung der Entitäten und Relationen. Neben den konzeptionellen Fragen, was eigentlich die für die Literaturgeschichtsschreibung relevanten und in einer RDF-Tripel-Struktur abbildbaren Aussagetypen sind, gibt die Station exemplarisch Einblick in den bisherigen Stand der Modellierung und die Wechselwirkungen der verschiedenen beteiligten Tools (vgl. Station 6), der Nachnutzung existierender Ontologien und Standards und der angedachten Verbindung mit externen Ressourcen. Station 4 stellt im Kontext des Pilotprojekts einen visualisierten Auszug aus dem Wissensnetzwerk vor, der verschiedene thematische Aussagen, die aus den drei verschiedenen Informationsquellen gewonnen wurden, zusammenbringt.

Station 5: Recht

Das Forschungsprojekt wird von Beginn an rechtswissenschaftlich begleitet. In Station 5 soll der Modus des interdisziplinären Austauschs beschrieben werden: Es werden rechtliche Hürden identifiziert, die im Projektkontext auftauchen, und diese dann in Form von Handreichungen abstrahiert dargestellt. Inhaltlich zielt die interdiziplinäre Zusammenarbeit bisher darauf ab, insbesondere die urheberrechtlichen Fragestellungen bei der Entwicklung eines Wissensnetzwerkes für die Geisteswissenschaften zu ergründen. Dabei verfasste Handreichungen machen es sich zum Ziel, die rechtlichen Rahmenbedingungen beim Einsatz von Text und Data Mining-Verfahren in den Geisteswissenschaften – über den Projektkontext hinaus – darzustellen.

Station 6: Infrastruktur

Das Forschungsprojekt stellt seine Infrastruktur und genutzte Tools vor: Wir setzen zur Digitalisierung der Werke aus dem 18. Jahrhundert OCR4all ein, annotieren die Sekundärliteratur mit dem semantischen Annotationstool INCEpTION und arbeiten mit der Open Source Software Wikibase zur Modellierung eines Wissensgraphen.

Wissensnetzwerk im Dialog

Durch eine Kombination aus kurzweiligen Impulsvideos und einem vertiefenden Dialog an jeder Station wird deutlich, wie die verschiedenen Informationsquellen und Forschungsbereiche in „MiMoText“ ineinandergreifen mit dem Ziel, neue Wege für die Analyse und Modellierung von Literaturgeschichte zu beschreiten. Anhand des „roten Fadens“ thematischer Aussagen, der durch den virtuellen Raum hindurch gelegt wird, kommen alle Partizipierenden ins Gespräch über die Herausforderungen im Aufbau eines literaturgeschichtlichen Wissensnetzwerks, aber auch den möglichen Nutzen und wünschenswerte Abfragemöglichkeiten.

Anm.: Die vDHd 2021 ist eine durch die Community des Verbandes Digital Humanities im deutschsprachigen Raum organisierte virtuelle Konferenz. Die Abstracts haben kein Peer-Review-Verfahren durchlaufen.