# Einleitung

In der Ära der Digitalität sind zunehmend große Textmengen, umfangreiche Metadaten und zahlreiche Hintergrundinformationen digital verfügbar. Geisteswissenschaftliche Forschung stellt besondere Anforderungen an die Verfügbarkeit, Aufbereitung und Durchsuchbarkeit der relevanten Daten. Das Projekt *Mining and Modeling Text: interdisziplinäre Anwendungen, informatische Weiterentwicklung, rechtliche Perspektiven* (im Folgenden kurz *MiMoText*) hat sich zum Ziel gesetzt entsprechend aufbereitete Daten zur Verfügung zu stellen. Um dies zu erreichen wird sich das Projekt von Grund auf mit den erwachsenden Fragestellungen im Bereich der Korpuserstellung, Informationsextraktion, Modellierung und Analyse auseinandersetzen. Dazu bedarf es gebündelter Ressourcen aus den Geistes-, Informatik- und Rechtswissenschaften.

# Ausführliche Projektbeschreibung

Das Projekt beschäftigt sich mit der Bereitstellung einer Graphdatenbank für Literaturwissenschaftler. Konkret geht es um französischsprachige Romane, die im Zeitraum 1750-1800 verfasst wurden und erschienen sind. Es sollen Beziehungen hergestellt werden zwischen den Romanen und literaturwissenschaftlichen Texten, die sich auf diese Romane beziehen. Im Bereich der Sekundärliteratur wird versucht einen möglichst breiten Zeitraum abzudecken, der sich von der ersten Hälfte des 19. Jahrhunderts bis heute erstreckt. Ziel ist es durch ein semantisch detailliert erschlossenes Korpus mit elaborierten Abfragemöglichkeien neue Wege für die Literaturwissenschaft zu eröffnen. Der Romanauswahl des Projekst liegt *Bibliographie du genre romanesque français 1751-1800* zu Grunde. Diese listet die Romane von 1750-1800 in sehr ausführlicher Form auf, u.a. mit kurzen Stichwörtern zu Genre und Protagonisten. Grundlegende Informationen aus der Bibliographie wurden bereits vor Beginn des Projekts im Rahmen einer Masterarbeit von Andreas Lüschow extrahiert. Diese Informationen zu Autor, Werk etc. wurden dann als semantisches Netz modelliert. Auf der Grundlage dieser Bibliographie wurden dann die Primärtexte zusammengesucht. Teilweise wurden diese im Rahmen des Projekts gescannt, teilweise konnten bereits digitalisierte Werke verwendet werden. Passend zu den vorhandenen Primärwerken werden dann Sekundärwerke, die sich mit diesen beschäftigen in das Korpus aufgenommen. Sowohl das Primär- als auch das Sekundärliteraturkorpus befinden sich noch im Aufbau. Eine Voraussetzung für die Auswahl der Sekundärliteratur war, dass diese deutschsprachig sein sollten. Bei der Zusammenstellung des Sekundärtextekorpus mussten Fragen der Kanonisierung, der Multilingualität und des Datenabgleichs mit der Bibliographie bedacht werden. Beide Korpora werden von Scanbild in Text umgewandelt, teilweise von Hand, größtenteils jedoch unter Verwendung von OCR4all. Danach kommt das Preprocessing: Satzsegmentierung, POS-Tagging, durchgeführt mit nltk. Der erste Preprocessing-Schritt, der (kleinere) Probleme aufwirft ist die Named Entity Recognition. Der Natur der Texte entsprechend enthalten auch die überwiegend deutschsprachigen Texte französischsprachige Passagen wie z.B. Zitate oder Werktitel. Um diesem Problem zu begegnen wurde zuerst eine automatisierte Erkennung der Sprache implementiert. Es wird davon ausgegangen, dass die Texte entweder deutsch- oder französischsprachig sind. Eventuell vorhandene Abweichungen (Englisch, Latein) werden nicht berücksichtigt. Die NER wird einmal unter Nutzung von NLTK implementiert und einmal unter Nutzung von BERT.

## Projektbeschreibung allgemein

Das Projekt gliedert sich in die Teilbereiche *Mining*, *Modeling*, *Recht* und *Infrastruktur*. Der Bereich Mining beschäftigt sich mit automatischer Identifikation und Extraktion von Metadaten, Sachinformationen und Texteigenschaften aus den drei Informationsquellen. Die gewonnenen Daten werden als rdf/xml aufbereitet, um so Beziehungen zwischen den behandelten Entitäten darstellbar und abfragbar zu machen. Bei der Bereitstellung von Texten und sei es nur in Ausschnitten sind zahlreiche juristische Probleme zu klären. Diese Fragestellungen werden in Kooperation mit dem Institut für Recht und Digitalisierung der Universität Trier behandelt. Um die genannten Dienste zur Nutzung anbieten zu können, braucht es die entsprechende Infrastruktur. Diese wird im Rahmen des Projekts mit aufgebaut.

## RA 1

RA1 Mining extrahiert Informationen aus drei unterschiedlichen Textarten.

1. Bibliographische Informationen

Die ersten dem Projekt zu Grunde liegenden Daten wurden von Andreas Lüschow im Rahmen seiner Masterabeit *Vom verlegten Buch zu verlinkten Daten* aus der *Bibliographie du genre romanesque français 1751-1800* (im Folgenden als *Bibliographie* abgekürzt) extrahiert. Dabei handelt es sich um eine in den 1960er und -70er Jahren verfasste Bibliographie, die alle französischen Romane dieser Zeitspanne auflistet mit Informationen u.a. zu Verfasser, Titel, Erscheinungsjahr und Gattungsbezeichnung. Diese Angaben wurden von Lüschow (2018) extrahiert und unter Rückgriff auf verschiedene Ontologien und Vokabulare als rdf-xml modelliert.

2. Informationen aus deutschsprachiger Sekundärliteratur

Erstes Ziel ist es hier aus der ausgewählten Sekundärliteratur Statements über Metadata-Items zu gewinnen, die ihrerseits aus der *Bibliographie* gewonnen wurden. Zunächst mussten hierfür geeignete Werke gefunden werden. Die erste Idee mit literaturwissenschaftlichen Aufsätzen zu arbeiten ließ sich nicht umsetzen, da keine deutschsprachigen Aufsätze gefunden wurden, die sich mit den für uns relevanten Werken beschäftigen. Deshalb wurden stattdessen auf die Periode bezogene deutsch-sprachige Literaturgeschichten (Überblickswerke) mit chronologischer Streuung 19. – 21. Jhd. ausgewählt. Später soll das Korpus weiter ergänzt werden. Die Texte werden gescannt und mit OCR4all in Text umgewandelt. Zur weiteren Erschließung werden momentan folgende Schritte durchgeführt: Gewinnung relevanter Aussagen aus der Bibliographie, zusätzliche Erschließungsarbeit durch Aufschlüsselung der Keywords. Überführung in relationale Datenbank um händische Kontrolle und Nachbearbeitung der Keywords zu ermöglichen. Außerdem wird an Visualsierungsmöglichkeiten für die Bibliographie gearbeitet. Zunächst werden Named Entities in der Sekundärliteratur gesucht, dann erfolgt ein Abgleich inwiefern die gefundenen Personen und Figurennamen, Orte etc. mit den Angaben in der Bibliographie übereinstimmen. Zur Bestimmung der Named Entities muss vorher eine Erkennung der jeweils verwendeten Sprache stattfinden. Als im Moment noch nicht ganz ausgereifter Schritt soll ein Topic Modelling stattfinden, um einen Abgleich zwischen den vermeintlichen Hauptthemen, die in der Bibliographie genannt werden, den Themen in der Primärliteratur und den Themen, die in der Sekundärliteratur genannt werden. Diese Aufgabe weist in den Bereich von Julia Röttgermann.

Festlegung relevanter Texteigenschaften

Unter dem Primat der geisteswissenschaftlichen Nachnutzung steht die Festlegung relevanter Texteigenschaften.

Dazu können Informationen gehören, die für stilometrische Analysen relevant sind wie z.B. die Bestimmung der Textlänge in verschiedenen Einheiten wie Zeichen, Wörtern, Absätzen oder Kapiteln.

Von besonderem literaturwissenschaftlichen Interesse sind Gattungsinformationen. Diese beinhalten z.B. die Erzählform. Orientiert an der einfachen Typologie, die die Bibliographie vorgibt kommen die Erzählformen Briefroman, Memoirenroman, Roman erzählt in der dritten Person oder Dialogroman in Frage.

Spannende Perspektiven zur semantischen Erschließung ergeben sich aus der Extraktion von fiktionalen Named Entities und deren Verknüpfung mit solchen, die sich auf die reale Welt beziehen, sowie die Möglichkeit zu untersuchen welche Aussagen über die fikitonalen Named Entities an anderer Stelle getroffen werden .

## RA 2

Die Modellierung soll vorhandene Ontologien nachnutzen, insbesondere diejenige von Dublin Core, SPAR und Wikidata, weil dadurch die Anschlussfähigkeit an vorhandene Knowledge Graphs wesentlich erhöht wird. Zugleich müssen für bestimmte Bereiche spezifische Ontologien entwickelt werden, die den Bedürfnissen des Projekts gerecht werden. Insbesondere im Bereich der Aussagen über Autoren und Werke ("Interpretierende Inhalte") sollen die hier definierten Aussagetypen (Properties und mögliche Values) auch dazu dienen, die Informationsextraktion aus der Fachliteratur zu steuern bzw. zu fokussieren. Zu diesem Zweck wird auf Grundlage des Korpus eine spezifische Ontologie entwickelt, deren Elemente jedoch im Anschluss möglichst im Rückgriff auf bereits existierende Vokabulare ausgedrückt werden sollen.

# Praktisches Vorgehen/aktueller Projektstand

Momentan testen wir Inception zur Annotation von NER als Grundlage für überwachte Lernverfahren. Inception bietet die Möglichkeit über eine graphische Nutzeroberfläche NERs zu markieren, Koreferenzen aufzulösen und den NERs Ids zuzuweisen, z.B. auf Grundlage von Wikibase. Die Ergebnisse sollen als rdf/xml exportiert werden können. Es sind verschiedene Algorithmen zur Klassifikation von NERs in Inception integriert, die als Datengrundlage, die bereits vom Nutzer annotierten Texte nutzt.

# Semantic Web und Linked Data

Im weitesten Sinne bezeichnet Semantic Web alle Techniken, Standards und Methoden, die dazu verwendet werden Daten, die zunächst nur menschenlesbar sind maschinenlesbar zu machen und somit für die automatische maschinelle Verarbeitung und Aufbereitung verwendbar. (Tim Barners Lee, W3c)

Linked Data zeichnet sich durch vier Prinzipien aus (s. Lüschow S. 21 & Berners-Lee)

Verwendung von Uniform Resource Identifiers (URIs)

Verwendung von HTTP-URIs

Die URI verweist auf nützliche Informationen, die unter Verwendung der gängigen Standards (RDF, SPARQL) bereitgestellt werden

Verlinkung zu anderen URIs

##### 2.1.2. Ontologien

Unter Ontologien werden im Kontext des Semantic Web Dokumente verstanden, die anhand der o. g. Standards erstellt werden und in denen das Wissen einer bestimmten Anwendungsdomäne modelliert wird.21 Wichtig ist hierbei, dass nicht nur einzelne Elemente oder Fakten aus dieser Domäne benannt werden, sondern dass Definitionen zentraler Elemente, deren Beziehungen untereinander und mögliche Bedingungen bei der Verwendung bestimmter Beziehungen in der Ontologie beschrieben werden. Beispiele für Ontologien finden sich in Abschnitt 2.5. (Zitat Lüschow S. 22)

##### 2.1.3. SPARQL

SPARQL (SPARQL Protocol and RDF Query Language) ist eine Abfragesprache und W3C Recommendation von 2013, die es möglich macht, Informationen aus RDF-Graphen zu extrahieren und darzustellen. Neben einfachen Anfragen in Tripel-Form lassen sich komplexere Anfragen durch Verwendung von Gruppierungen, Vereinigungen von Teilergebnissen und Filtern, z. B. um Wertebereiche einzuschränken, erreichen. (Lüschow S. 22)

##### 2.2. RDF

Der zentrale Standard zur Repräsentation semantischer Daten ist das Resource Description Framework (RDF), dessen aktuelle Version aus dem Jahr 2014 stammt. In RDF-Dokumenten werden Informationen in Form eines gerichteten Graphen repräsentiert – im Gegensatz zur Baumstruktur, wie sie z. B. in XML-Dokumenten zu finden ist. Die Knoten und Kanten dieses Graphen sind dabei mit eindeutigen Bezeichnern versehen, um sowohl die modellierten Entitäten (Knoten) als auch deren Beziehungen (Kanten) untereinander identifizieren und in unterschiedlichen Kontexten nutzen zu können. Durch Verwendung von URIs und Namensräumen wird sichergestellt, dass gleich benannte Bezeichner in unterschiedlichen Domänen unterschiedliche Bedeutungen haben können. Da ein Hauptanliegen des Semantic Web darin besteht, die Verknüpfung ganz unterschiedlicher Daten zu ermöglichen, ist Disambiguierung in Form von eindeutigen Identifikatoren und durch Verwendung von Normdateien umso wichtiger. Die grundlegende Struktur in RDF modellierter Informationen ist das Tripel, d. h., die Verbindung zweier Knoten durch eine Kante. Dadurch können einfache Aussagen der Form „Subjekt – Prädikat – Objekt“ ausgedrückt werden. Die Kombination vieler Tripel erzeugt einen Graphen, in dem jeder Knoten sowohl als Subjekt als auch als Objekt beliebig vieler Aussagen auftreten kann. Sämtliche Elemente (also alle Subjekte, Prädikate und Objekte) eines Gegenstandsbereiches bezeichnet man auch als Ressourcen und Elemente an Kanten als Properties. An der Stelle des Objektes einer Aussage können neben mit URIs bezeichneten Knoten außerdem Literale auftreten, deren Wert durch eine bestimmte Zeichenkette bestimmt wird. Abb. 2.1 veranschaulicht einen RDF-Graphen

sowie die Verwendung von Literalen und URIs anhand der Aussage „Das Buch Semantic

Web – Grundlagen wird im Springer-Verlag verlegt“. RDF-Graphen können einerseits durch ihre graphische Repräsentation, andererseits auch durch die Aufzählung der in ihnen enthaltenen Tripel dargestellt werden. Für diese Darstellung gibt es unterschiedliche Möglichkeiten und Standards, von denen Turtle und RDF-XML die verbreitetsten sind. Turtle bietet zwar eine recht einfache, intuitive und menschenlesbare Struktur, RDF-XML hingegen erlaubt die Repräsentation komplexerer Inhalte und die maschinelle Verarbeitung (Parsing) in allen gängigen Programmiersprachen, weshalb es die am häufigsten genutzte Art der RDF-Modellierung darstellt. (Lüschow S.22ff.)

# Bibliographische Daten als Linked Data

Zurzeit existiert noch kein einheitlicher, akzeptierter Standard, der in der Bibliothekswelt für die semantische Repräsentation bibliographischer Daten verwendet wird. Stattdessen orientieren sich diejenigen Bibliotheken, die ihre Daten bereits als Linked Data zur Verfügung stellen, an unterschiedlichen Datenmodellen, Schemas und Ontologien, die für sich genommen allerdings teilweise den Status von Standards besitzen. Nichtsdestotrotz existieren zahlreiche Versuche, die bereits entwickelten Modelle in ein möglichst generisches und von vielen Bibliotheken nachnutzbares Modell zu integrieren.

Modellierungsaspekte:

Es wird deutlich, dass die semantische Modellierung bibliographischer Metadaten bislang bei Weitem nicht geklärt ist. Durch das bestehende Nebeneinander unterschiedlicher Modellierungsansätze, die jeweils eigene Perspektiven auf die Daten und die Anforderungen einzelner Institutionen berücksichtigen, lässt sich noch kein (De-Facto-)Standard nutzen, der eine breite Zustimmung in der Bibliotheksgemeinschaft hat. Lüschow S. 30

FaBiO, BiRO, PRO wurde von Andreas wirklich verwendent, s. Lüschow S. 45 (pdf).

# Technische Umsetzung

Als erster Schritt wurde nun ein SPARQL-Endpoint implementiert unter Verwendung von rdflib. Als erster Schritt wurde eine Sparql-Abfrage gestartet, die Autor, Werk, ID und keywords enthält. Diese Keywords sind ein Sammelfeld für Gattungsbezeichnungen, Protagonistennahmen und verschiedenes weiteres. Einfache Vorsortierung RegEx. Außerdem wird in den Sekundärtexten nach den bekannten Personennamen gesucht. Als nächstes wird eine NER durchgeführt. Diese soll reale und fiktionale Personen erfassen sowie Ortsangaben.

# NER/Spracherkennung

# Beschreibung Vorgehen

### Architektur des

### Methodenüberblick

### LOD

Informationsextraktion:

Regelbasierte Verfahren

Überwachtes ML

Unüberwachtes ML

Unüberwachtes ML

Regelbasierte Verfahren im NLTK

Verfahren in Spacy

BERT

**NER mit Bert**

(Devlin et al. 2018)

Es soll versucht werden BERT auf unser NER-Problem anzuwenden

Was ist BERT

Bei BERT handelt es sich um ein new language representation model.

Bidirectional Encode Represantation from Transformers. (s.a. Peters et al 2018)

Besonderheit: bidirektionale Representation

Bis jetzt nur unidirektional, entweder rechts-links oder links-rechts, weil sonst ewige rückkopplungsschreife.

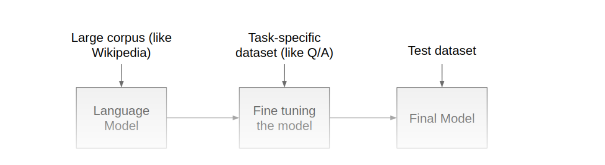
Pretraining

Ungelabelte texte!

Pretrained BERT kann feingetunet werden mit nur einem zusätzlichen Layer.

Paper enthält gute Auflistung an NLP-Benchmarks

Pretraining:



Das Modell wird auf einer sehr großen Datenmenge, die nicht taskspezifisch ist trainiert.

Im Fine-Tuning-Schritt kommt dann Training auf einem taskspezifischen Datenset zustande.

Pre-Training hat sich als effizient herausgestellt, . s1s1

Es gibt zwei unterschiedliche Wege pre-trained sprachmodelle für tasks anzupassen.

1. Feature-based peters et al 2018a, elmo
2. Fine tuning (generativ pretrained transformer openAI GPT radord et al 2018)

BERT ist ein entwickelter Fine-Tuning-Ansatz

Benutzt ein Masked Language Model, inspieriert by the Cloze task Taylor 1953

Masked Language Model: Ein Teil der Input-Token wird zufällig ausgewählt und maskiert dieses Token. Diese Token soll dann an Hand des Kontextes erraten werden. Im gegensatz zu anderen Modellen kann sowohl der vorangehende als auch der nachfolgende kontext verwendet werden, auf Grund der Bidirektionalität des Modells.

BERT soll den Anpassungsaufwand und die Entwicklung von aufwendigen taskspezifischen Architekturen ablösen.

Es gibt BERT für viele unterschiedliche Sprachen

[**BERT-Base, Multilingual Cased**](https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip): 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters

s. GitHub

OK, auf jeden Fall NER Mehrsprachigkeit auf unseren Daten testen <3

**Ansätze, die ähnlich sind wie BERT**

**Unsupervised Feature-Based Approaches**

s.2 u.a. ELMO Peters et al 2017, 2018a

Word embedding + extra dimension

Context sensitive features form l->r und r->l llanguage model

**Unsupervised Fine-Tuning Approaches**

s.2f.

**BERT**

1. Pre-Training

Unlabled data, different pretraining tasks

1. Fine-Tuning

Initialisierung mit pre-trained parametern, dann anpassung aller parameter anhand gelablleter daten speziell für den task

Besonders an BERT ist  die Architektur, die über verschiedene Tasks hinweg unverändert ist.

Nur kleiner Unterschied zwischen pretrained und downstream architektur.

**Architektur**

Multi layer bidirectional transformer encoder based on the original implementatin vaswani et al 2017. (Vaswani et al. 2017) details zu architektur und implementierung dort.

# Tools

## Tools Annotation

Vergleich Annotationstools

* Recogito
* CATMA
* WebAnno
* Inception
* Athen
* FuD
* WissKI
* Neonion

Sonstige Annotationstools:

* Transcribo
* Wikibase
* Scholary Web Annotation Client
* ELAN

Anforderungsprofil:

* Open Source
* Freie Nachnutzbarkeit
* Freie Definitionsmöglichkeiten von Entitäten und Relationen
* XML-Output
* Annotation von Relationen möglich
* Importmöglichkeit für bereits annotierte Daten
* Möglichkeit überlappender und unterbrochener Annotationen
* Speicherung als Linked Data möglich
* Kollaboratives Annotieren
* Verschiedene Quellen von Annotationen
* Organisation durch Nutzung verschiedener Ebenen möglich
* Notizfunktion o.ä. für unklare Stellen
* Integration von Anntationsguidelines
* Größenrahmen: 1 Satz – 300 Romane

## Tools NLP-Programmierung

Implementierung:

* Tensorflow
* Keras
* Bert
* NLTK
* Pytorch
* Spacy
* Hugging Face (Transformers)

# Evaluation

# Fazit und Ausblick