## Methode

Wissensgraphen können auf unterschiedliche Arte und Weise erstellt werden. In der Praxis sind die häufigste Variante: manuelle Erstellung, Erstellung aus (semi)strukturierten und unstrukturierten Quellen (vgl. Heist et al. Ege/Paschke 2021: 129). Außerdem gewährleisten neuronale Modelle des Textverstehens wie Large Language Models (LLM) ein großes Potential für die Wissenserfassung (vg. Hamann Hinkelmann/Hoppe/Humm 2025: 68).

Bei der manuellen Erstellung von Wissensgraphen werden die Fakten und Entitäten auf Grundlage von Erfahrung einer oder eines Experter:in definiert. Dies sichert die hohe Qualität des Wissens, aber dieser Ansatz ist zeitaufwendig und arbeitsintensiv (vg. Hamann Hinkelmann/Hoppe/Humm 2025: 74).

Eine Mischform aus manuellem Ansatz und KI-Modellen könnte ein optimaler Ansatz sein. Mithilfe von LLM wird das Korpus nach Entitäten und Fakten durchgesucht, während der manuelle Ansatz zu der Sicherstellung aller relevanten und qualitativ hochwertigen Informationen beiträgt (vg. Hamann Hinkelmann/Hoppe/Humm 2025: 74).

Im Vergleich zu dem manuellen Ansatz hat ein automatisierter Ansatz mehrere Vorteile. Er ist kosteneffizient, skalierbar, flexibel und aktuell. Ein Nachteil des automatisierten Ansatzes ist die Notwendigkeit von ausreichende annotierte Daten für das Modelltraining (vg. Hamann Hinkelmann/Hoppe/Humm 2025: 75).

Die Umsetzung der Implementierung erfolgte in Python unter Verwendung der Standardbibliotheken SpaCy, ….

## OCR-Erkennung

## Informationsextraktion aus Texten

Die Informationsextraktion besteht aus mehrere Teilschritten: Named Entity Recognition (NER) und Linking (NEL), Kohärenzauflösung, Extraktion zeitlicher Informationen, Relationsextraktion (RE). Die Umsetzung der Implementierung erfolgte in Python unter Verwendung der Standardbibliotheken SpaCy, ….

In Bezug auf der Informationsextraktion aus unstrukturierten Daten können unterschiedliche Kategorie von Datentypen identifiziert und extrahiert werden. Man spricht dabei über Entitäten, Beziehungen, Adjektiven und ergänzende Informationen (vgl. Lanquillon/Schacht 2023: 163).

### 5.2.1 Named Entity Recognition (SpaCy, ML, generative LLM)

Entitäten sind hauptsächlich Subjekte und Objekte in einem Text. Die meinst bekannteste sind Named Entities wie Personen, Orte, Firmen usw. Je nach Forschungsfrage können auch andere Subjekte wie Krankheiten, Medikamente usw. als Entitäten bezeichnet werden (vgl. Lanquillon/Schacht 2023: 163).

Eine wichtige Komponente bei der Wissensextraktion im Bereich von Text-Mining ist Named Entity Recognition (NER) (vg. Hamann Hinkelmann/ Hoppe/Humm 2025: 81).

Durch ein NER-Tagger werden Entitäten in einem Text identifiziert und markiert. Er verwendet dabei generische Typen wie *Person* oder *Firma* (vg. Hamann Hinkelmann/Hoppe/Humm 2025: 81).

Bei NER treten verschiedene Schwierigkeiten. Ein Problem besteht darin, ob die einzelnen Wörter im Text zusammen oder getrennt betrachtet werden sollen (*New Yourk*). Ein weiteres Problem ist die Mehrdeutigkeit von Wörtern (*Washington*). Bei solchen Wörtern kann schwierig sein, die richtige Entitätsklasse zu bestimmen (vgl. Lanquillon/Schacht 2023: 172).

Der erste Schritt betrifft die Rohdaten. Das Ziel der Vorerarbeitung (ML-Training) ist es, …

## Relationsextraktion (Ollama)

Relationsextraktion verfolgt das Ziel, die erkannte Entitäten im Text miteinander zu verknüpfen (vg. Hamann Hinkelmann/ Hoppe/Humm 2025: 84). Beziehungen treten als Verben auf und bilden die syntaktische und semantische Verbindung zwischen Entitäten. Sie können mithilfe regelbasierter Verfahren identifiziert werden. Normalerweise werden die Beziehungen allgemein geschrieben: *ist Chef von, ist angestellt bei* (vgl. Lanquillon/ Schacht 2023: 165).

Dadurch werden Tripeln direkt gebildet (vg. Hamann Hinkelmann/ Hoppe/Humm 2025: 84)

Damit Relationsklassifikation effektiv funktioniert, müssen die Entitäten in kurzen Textabschnitten gemeinsam erwähnt werden (vg. Hamann Hinkelmann/ Hoppe/Humm 2025: 85).

Kohärenzauflösung

Während der Datenvorverarbeiten wurden die Texte in Sätze untergeteilt.

Pronomen oder Umschreibungen von Entitäten können Informationen im Text verschleiern, wenn der Kontext nicht berücksichtigt wird.

## 5.3 Datenabgleich

Die automatische Zuordnung von Entitäten zu einem Graph erfolgt durch Named Entity Linking (NEL) (vg. Hamann Hinkelmann/Hoppe/Humm 2025: 83)

Da die Maschinen keine indirekten Bedeutungen erkennen können, daher müssen Relationen zwischen Informationen eindeutig sein. Normdaten wie [GND-Nummer](https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html) (Gemeinsame Normdaten) ermöglichen diese Eindeutigkeit. Wird eine Person über GND-Nummer identifiziert, können Inhalte automatisch mit bspw. [Wikipedia](https://www.wikipedia.de/)-Eintrag oder [Deutschen Biographie](https://www.deutsche-biographie.de/) verknüpft werden. Trotzt der Bemühungen der Deutschen Nationalbibliothek kann die Eindeutigkeit von Normdaten nicht in allen Fällen gewährgeleistet werden. Regelmäßig auftretende Probleme, wie etwa fehlerhafte oder doppelt vorhandene GND-Einträge, lassen sich häufig auf die dezentrale Struktur der Datenerfassung zurückführen, die im Rahmen kooperativer Arbeitsprozesse. In der Praxis wird aus Effizienzgründen oder unter Zeitdruck nicht selten darauf verzichtet, bereits bestehende Datensätze auf ihre Richtigkeit und Vollständigkeit hin zu prüfen. Stattdessen wird ein neuer Eintrag erzeugt, was wiederum zu Redundanzen sowie zu einer Verringerung der Datenkonsistenz und -qualität im Gesamtsystem führen kann. Die Erstellung und Pflege von Normdaten stellt auch die Herausforderung im Bereich von Körperschaften, geographische Orte, Werte, Periodika usw. Ihre eindeutige Identifikation erweist sich oftmals als deutlich komplexer. Insbesondere bei Orten treten spezifische Schwierigkeiten auf: Während georeferenzierende Systeme (GIS) in der Regel lediglich punktuelle Koordinaten erfassen, bleiben flächenhafte Ausdehnungen sowie historische Veränderungen unberücksichtigt. Die historische Dimension eines Ortes – etwa sich verändernde politische Zugehörigkeiten, Ortsnamen oder Grenzverläufe – kann daher nur unzureichend über bestehende textbasierte Normdaten erfasst werden und bedarf ergänzender, oft nicht-textueller Quellen (vgl. Kamzelak 2018: 426-427).

## 5.4 TEI-XML

## 5.5 RDF

## 5.6 OWL (Lanquillon/Schacht 2023, S. 187)

## 5.7 Turtle

## 5.8 Wikidata-Datenmodell

Laube-Artikel

Zur Auffindbarkeit und technischen Interoperabilität von Projektdaten ist es sinnvoll, etablierte Identifikatoren und Normdaten zu verwenden. Im Rahmen dieser Arbeit wurde Wikidata genutzt, um externe Konzepte und Realitäten referenziert und gemeinsamen Bezugsrahmen zu schaffen.

* 1. Zusammenfassung

## SPARQL

Laube-Artikel

## Visualisierung

Netzwerkanalyse biete neue Möglichkeit zur Visualisierung und Analyse von Sammlungswissen. Ihre zentralen Ziele sind die Identifikation wichtiger Verbindungen und Strukturen. (vgl. Werner 2020: 248–249).

Für geistes- und sozialwissenschaftliche Forschung ist die Verbindung zwischen Netzwerken und Visualisierung von großer Bedeutung. Die graphische Visualisierung von abstrakten oder komplexen Strukturen macht die Zusammenhänge, die sich auf rein sprachlichem Wege nur schwer oder unzureichend vermitteln lassen, greifbar. Digitale Werkzeuge bieten hier neue Möglichkeiten, große Datenbestände nicht nur zu verarbeiten, sondern auch visuell aussagekräftige Formen überführen können. Daraus ergibt sich einen Anknüpfungspunk für die Verwendung netzwerkbasierter Modelle. Für die biographische Forschung eröffnet der Netzwerkansatz die Perspektive, individuelle Lebensläufe nicht isoliert zu betrachten, sondern als Teil größer Beziehungsgeflechte (vgl. Dörpinghaus 2022).

## Auswertung / Perspektive

## Zusammenfassung

## Literaturverzeichnis

## Anhang