Heinrich-Heine-Universität Düsseldorf

Fakultät für Mathematik und Naturwissenschaften

Institut für Informatik

Exposé zur Bachelorarbeit

Automatische Modellierung von Wissensgraphen für universitäre Veranstaltungen

Vorgelegt von:	Armin Asad-Zadeh-Tabrizi
Email:	arasa100@uni-duesseldorf.de
Studiengang:	Bachelor Informatik
Lehrstuhl:	Lehrstuhl für Datenbanken und Informationssysteme
Erstgutachter:	Prof. Dr. Stefan Conrad
Betreuungsperson:	Nina Amelie Liebrand
Datum:	27.09.2025

Contents

1 Forschungsthema

Das Forschungsthema dieser Bachelorarbeit ist die **automatische Modellierung von Wissensgraphen** für universitäre Veranstaltungen. Ziel ist es, Daten aus verschiedenen Quellen wie Vorlesungsverzeichnissen, Modulhandbüchern oder Webseiten automatisch zu extrahieren, semantisch zu verknüpfen und in einem Wissensgraphen abzubilden. Dadurch sollen inkonsistente oder unvollständige Informationen erkannt und die Datenqualität verbessert werden. Ein besonderer Fokus liegt auf der Entwicklung einer Methode, die sich auch auf andere Universitäten übertragen lässt, ohne dass der Wissensgraph manuell erstellt werden muss. Ein solcher Wissensgraph ermöglicht es Hochschulen und Studierenden, Zusammenhänge zwischen Kursen, Dozierenden, Räumen und Zeiten effizient abzubilden. Dadurch können Planungsfehler reduziert und strukturierte Datenanalysen ermöglicht werden. Außerdem liefert die Arbeit Erkenntnisse für die Integration und Qualitätssicherung heterogener Datenquellen im universitären Umfeld.

2 Zielsetzung und Erkenntnisinteresse

Ziel der Arbeit ist die Entwicklung einer Methode zur **automatischen Erzeugung eines Wissensgraphen** aus heterogenen universitären Datenquellen. Dabei sollen Verfahren zur Datenintegration und -validierung so gestaltet werden, dass sie für verschiedene Hochschulsysteme adaptierbar sind.

Teilziele der Arbeit

- Entwicklung einer Methode zur automatischen Extraktion und Modellierung universitärer Daten in einem Wissensgraphen.
- Untersuchung geeigneter Datenintegrationsverfahren für heterogene Formate (z. B. HTML, PDF).
- Methoden zur Sicherung der Datenqualität evaluieren und dokumentieren.
- Übertragbarkeit der Methode auf andere Hochschulen.

3 Forschungsstand und theoretische Grundlagen

Wissensgraphen haben sich in den letzten Jahren als effektive Methode zur semantischen Modellierung und Vernetzung von heterogenen Daten etabliert [1]. Wissensgraphen stellen strukturierte Darstellungen von Wissen in Form von Knoten und Kanten dar. Dabei nennt man die Knoten auch "Entitäten" und die Kanten "Relationen". Mathematisch lassen sie sich als gerichtete Graphen $G = (V, E)$ mit einer Menge von Knoten V und Kanten E beschreiben. Sie ermöglichen semantische Verknüpfungen zwischen Datenpunkten und werden oft mit RDF (Resource Description Framework) und OWL (Web Ontology Language) umgesetzt [2]. Das Resource Description Framework (RDF) ist die Standard-Wissensrepräsentationssprache des Semantic Web und bietet eine graphenbasierte Semantik zur Beschreibung und Verknüpfung großer, heterogener Datensätze. RDF ermöglicht so eine maschinenlesbare Darstellung von Wissen [3]. Die Web Ontology Language (OWL) ist ein weiterer Standard des Semantic Web, der auf RDF aufbaut und eine formelle, modelltheoretische Semantik bereitstellt. Sie dient der Beschreibung ontologischen und terminologischen Wissens und ermöglicht die Definition von Klassen, Relationen und logischen Regeln, um Bedeutungszusammenhänge zwischen Daten formal abzubilden [4]. Zur automatischen Modellierung werden Ontologien definiert, die Konzepte wie Veranstaltung, Dozierender oder Raumnummer beschreiben. Zur Integration heterogener Quellen können Verfahren zur Textextraktion (z. B. PDFMiner, BeautifulSoup) und Datenparsing (z. B. RDFLib) eingesetzt werden [5]. Zur Sicherung der Datenqualität werden Methoden wie **SHACL-Validierungen** und **Plausibilitätsregeln** verwendet, um fehlerhafte oder widersprüchliche Daten zu erkennen [6].

4 Forschungskonzept

Forschungsfragen

- Wie kann ein Wissensgraph für universitäre Veranstaltungen automatisch aus heterogenen Quellen erzeugt werden?
- Wie lassen sich Konsistenz und Qualität der integrierten Daten sicherstellen?
- Wie kann die entwickelte Methode auf andere Hochschulen übertragbar gemacht werden?

Hypothese

Ein Wissensgraph, der auf Ontologien und semantischen Beziehungen basiert, soll sowohl die strukturierte Darstellung von Veranstaltungen ermöglichen als auch die Abfrage zeitlicher Überschneidungen erleichtern. Durch gezielte Methoden der Datenqualitätskontrolle (z.B. Konsistenzprüfungen und Plausibilitätsregeln) können trotz heterogener Quellen zuverlässige und konsistente Daten erzeugt werden.

Methodik

- Format, Struktur und inhaltliche Vollständigkeit der verfügbaren Datenquellen (LSF-Verzeichnis, Modulhandbücher, Website der Lehrstühle) wird untersucht. Ziel ist es, Unterschiede in der Datenrepräsentation zu identifizieren und daraus Anforderungen für die Extraktions- und Integrationslogik abzuleiten.
- Für die automatische Extraktion werden Bibliotheken wie BeautifulSoup (für HTML) und PDFMiner (für PDF-Dokumente) verwendet. Die extrahierten Daten werden in ein einheitliches Zwischenformat (z.B. JSON oder CSV) überführt, um eine strukturierte Weiterverarbeitung zu ermöglichen. Dabei werden unvollständige Einträge markiert, um sie später in der Qualitätsprüfung zu berücksichtigen.
- Daten mithilfe von RDF in Form von Tripeln (Subjekt–Prädikat–Objekt) modellieren. Die semantische Struktur wird durch eine Ontologie in OWL beschrieben, die Entitäten wie „Veranstaltung“, „Dozent“ und „Raum“ sowie deren Relationen definiert. Die Implementierung erfolgt mit RDFLib oder Apache Jena. Diese Ontologie dient als Schema für die automatische Erzeugung des Wissensgraphen.
- Zur Sicherung der Datenqualität werden SHACL-Validierungen eingesetzt, um strukturelle Konsistenz zwischen Instanzen und Ontologie sicherzustellen. Zusätzlich werden Plausibilitätsregeln definiert, z.B.:

  - Eine Veranstaltung kann nicht gleichzeitig in zwei Räumen stattfinden.
  - Der Startzeitpunkt einer Veranstaltung muss vor dem Endzeitpunkt liegen.
  - Jede Veranstaltung muss mindestens einen verantwortlichen Dozenten oder eine verantwortliche Dozentin besitzen.

- Evaluation der Qualität, Effizienz und Übertragbarkeit des automatisch erzeugten Wissensgraphen, damit der automatisch erzeugte Wissensgraph nicht nur syntaktisch und semantisch korrekt, sondern auch praktisch nutzbar und generalisierbar ist.

  1. Die **Datenqualität** des automatischen Wissensgraphen wird anhand von 4 Kategorien bewertet:
     - **Vollständigkeit**: Anteil der korrekt erfassten Entitäten und Relationen
     - **Genauigkeit**: Anteil der korrekt modellierten Tripel im Verhältnis zur Gesamtzahl der erzeugten Tripel
     - **Fehlerquote**: Anteil der inkonsistenten oder unvollständigen Knoten und Kanten
  2. Die **Effizienz** wird anhand der durchschnittlichen Antwortzeiten von SPARQL-Abfragen und der Speichergröße des erzeugten Graphen bewertet. Typische Anwendungsszenarien, nachdenen getestet wird:
     - Abfragen nach zeitlichen Überschneidungen von Veranstaltungen
     - Suche nach Dozenten bestimmter Fachbereiche
     - Aggregationen (z. B. Anzahl der Veranstaltungen pro Lehrstuhl)
  3. Untersuchung der **Übertragbarkeit der Methode** auf andere Hochschulsysteme. Dazu werden Datensätze einer zweiten, strukturell abweichenden Quelle (z. B. einer externen Fakultätswebseite oder eines anderen Studiengangs) in die Modellierung einbezogen. Bewertet wird, ob die Ontologie und die Extraktionslogik mit minimalen Anpassungen erneut anwendbar sind und vergleichbare Ergebnisse liefern.

5 Vorläufige Gliederung

1. Einleitung
    - Motivation und Relevanz
    - Forschungsziel und Fragestellung
2. Forschungsstand und theoretische Grundlagen
    - Wissensgraphen, Ontologien und RDF/OWL
    - Datenintegration und Datenqualitätsmethoden
    - Bestehende Ansätze für Hochschulveranstaltungen
3. Forschungskonzept und Methodik
    - Fragestellung und Hypothese
    - Modellierung des Wissensgraphen
    - Sicherung der Datenqualität
    - Evaluation der Abfrageeffizienz
4. Implementierung
    - Aufbau des Wissensgraphen
    - Abfrageszenarien zur Erkennung von Überschneidungen
5. Evaluation und Ergebnisse
    - Analyse der Abfrageeffizienz
    - Bewertung der Datenqualität
    - Diskussion der Ergebnisse
6. Fazit und Ausblick
    - Zusammenfassung der Erkenntnisse

6 Zeitplan

1. Woche 1–2: Literaturrecherche und theoretische Grundlagen.
2. Woche 3–4: Analyse der Datenquellen und Entwurf der Ontologie.
3. Woche 5–6: Implementierung der automatischen Datenextraktion und Modellierung.
4. Woche 7–8: Validierung der Datenqualität, Beginn des Schreibens der Einleitung und Theorie.
5. Woche 9–10: Evaluation und Schreiben der Methodik- und Implementierungskapitel.
6. Woche 11: Schreiben der Ergebnisse, Evaluation und Fazit.
7. Woche 12–13: Überarbeitung, Korrekturlesung und finale Abgabe.

7 Literaturverzeichnis

References

[1] Luigi Asprino, & Enrico Daga, & Aldo Gangemi, & Paul Mulholland (2022). *Knowledge Graph Construction with a Façade: A Unified Method to Access Heterogeneous Data Sources on the Web*. ACM Transactions on Internet Technology, 6, 1–31.

[2] Supriya A. Bejalwar (2025). *ENHANCING LIBRARY RESOURCE DISCOVERY AND MANAGEMENT WITH SEMANTIC WEB TECHNOLOGIES*. Gurukul International Multidisciplinary Research Journal.

[3] Mona Alshahrani, & Hussein Almashouq, & R. Hoehndorf (2016). *SPARQL2OWL: Towards Bridging the Semantic Gap Between RDF and OWL*. Semantic Scholar .

[4] P. Cimiano, & C. Chiarcos, & John P. McCrae & Jorge Gracia (2020). *Linguistic Linked Data - Representation, Generation and Applications*. Semantic Scholar.

[5] Alice Rogier & Adrien Coulet & B. Rance (2022) *Using an Ontological Representation of Chemotherapy Toxicities for Guiding Information Extraction and Integration from EHRs* Studies in health technology and informatics Pages: 91 - 95

[6] Jin Ke & Zenon G. Zacouris & Maribel Acosta (2024) *Efficient Validation of SHACL Shapes with Reasoning* Proc. VLDB Endow Vol. 17, Issue 11