BESCHLEUNIGUNG VON ONTOLOGIEENTWICKLUNG DURCH SPRACHMODELLE IN DEN DIGITALEN GEISTESWISSENSCHAFTEN

M. Sc. Johannes Mitschunas, M A. Clemens Beck, Prof. Dr. Clemens Beckstein, apl. Prof. Dr. Robert Gramsch-Stehfest
Fakultät für Mathematik und Informatik & Philosophische Fakultät, Friedrich-Schiller-Universität Jena

Abstract

Strukturierte Wissensbasen, wie etwa **FactGrid**, ermöglichen eine effiziente Datenabfrage und -analyse in der historischen Forschung. Ontologien sind hierbei zentral, da sie die Extraktion und Wiederverwendung von Informationen aus semistrukturierten Textkorpora unterstützen. Ein wesentliches Hindernis ist jedoch der zeitaufwändige Prozess der Formalisierung von Expertenwissen in Ontologiesprachen wie OWL.

Sprachmodelle (**LLMs**), bieten eine Lösung, indem sie natürlichsprachliches Wissen in formalisierte Ontologiefragmente übersetzen können. Dies ermöglicht eine effizientere Entwicklung von Ontologien, während Experten die Validierung und Verfeinerung vornehmen.

Hintergrund

Im DFG-geförderten Projekt **HisQu** (Laufzeit: 01.01.2025 bis 31.12.2027) ist die Entwicklung von Textparsern zur automatischen Datenextraktion aus semistrukturierten Texten, wie dem *Repertorium Germanicum* (RG), von zentraler Bedeutung. Diese Parser basieren auf einer expliziten ontologischen Grundstruktur, die an die Top-Level-Ontologie CIDOC-CRM anknüpft.

Erste Studien zeigen, dass LLMs wie GPT-4 in der Lage sind, die Formalisierung von Ontologien durch Fachexperten ohne tiefgreifende Modellierungserfahrung zu unterstützen und dabei den Entwicklungsprozess erheblich zu beschleunigen, unter Ausnutzung bereits vorhandener Ontologieteile und des Expertenwissens für einen gegebenen Textkorpus.

Pipeline & Methodik

Im Zentrum steht ein **Retrieval-Augmented Generation** (RAG)-Ansatz:
Als Basis der Modellierung wird ein Ausschnitt des Textkorpus (zB. ein Regest) genommen, der vom Domänenexperten annotiert wird. Ein Sentence Transformer durchsucht dann das vorhandene Ontologiewissen bestehend aus den früheren Modellierungen und einer Basisontologie (z.B. CIDOC-CRM) und extrahiert relevante Klassen, Relationen oder Beispiele. Diese Daten fließen zusammen mit dem ursprünglichen Quellentext und dessen Expertenannotation in das **Prompt** an das LLM (z.B. GPT-4).

Im **Frontend** sehen die Domänenexperten, welche Informationen (etwa Begriffe oder Ontologieausschnitte) dem LLM übergeben werden. Das **Backend** generiert daraufhin aus diesen Informationen und den formulierten Beziehungen ein **Ontologiefragment** (ABox/TBox), das auf syntaktische Korrektheit und Konsistenz (z.B. mithilfe von Reasonern wie HermiT oder Pellet) geprüft wird. Über Prompt-Engineering lässt sich das Verhalten des LLMs steuern und die Erklärungen („Begründungen“) des Modells sichtbar machen. So kann der Experte nicht nur Fehler aufdecken, sondern auch schnell nachvollziehen, ob das LLM die Anweisungen korrekt umgesetzt hat.

Zeigen sich Unklarheiten oder Inkonsistenzen, lassen sich entweder das neue Fragment oder der natürlichsprachliche Input nachbessern. Sobald das Ergebnis valide ist, wird es in die **Wissensbasis** integriert und erweitert diese sukzessive um formal definierte Entitäten und Relationen.

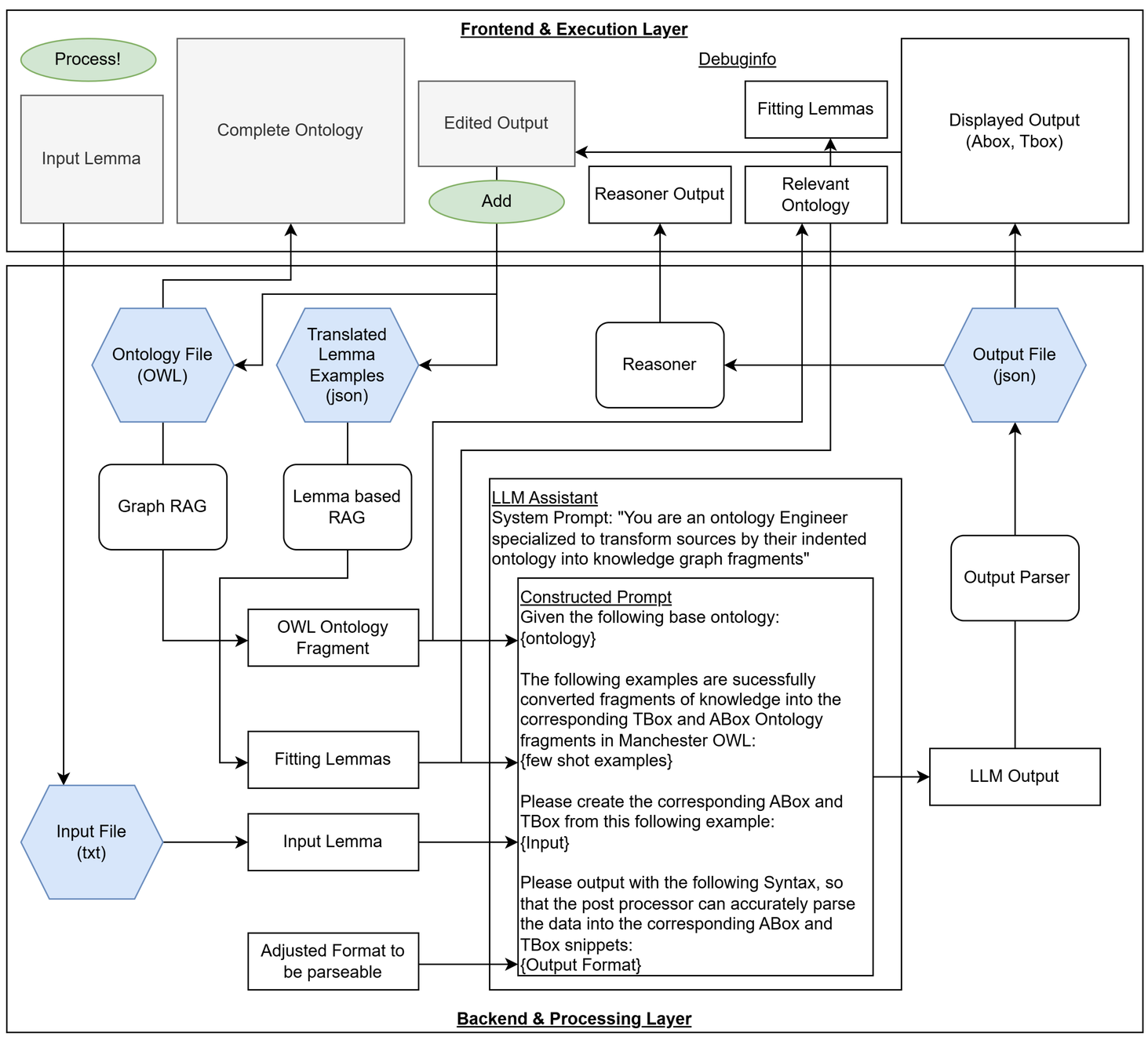

Abb. 1: Ontologiedesign mit dem LLM als Übersetzer zwischen natürlichsprachlich vormuliertem Domänenwissen und formaler Ontologie

Konkretes Beispiel: Regest 57

Aus dem *Repertorium Germanicum* (RG) wurde das Lemma „Regest 57“ ausgewählt, das mehrere Personen (z.B. Hudricus Bernardi), Institutionen (bspw. ein Hospital) und Datumsangaben enthält. Unter Einsatz von GPT-4 und einem **Retrieval-Augmented Generation** (RAG)-System, das bereits einige Ontologie-Bausteine und Beispiel-Lemmata kannte, konnten passende Klassen (**Presbyter**, **Canon**) und Relationen (**hasDate**, **belongsToOrder**) identifiziert werden. Auf diese Weise entstand ein **Ontologieschnipsel** (vereinfachte ABox/ TBox-Struktur), das das „Regest 57“ formal abbildet.

RG III 00057

Henricus de Bocholdia al. d. Foet cler. Traiect., mag. in art. bac. in decr. m. prov. super par. eccl. in Bodegrauen Traiect. dioc. vacat. per transgr. Ghiselberti de Lochorst ad decan. eccl. s. Saluatoris Traiect. 9 apr. 1410. L 138 254v.

9 apr. 1410

Abb. 2: Beispielregest 57 aus dem RG III

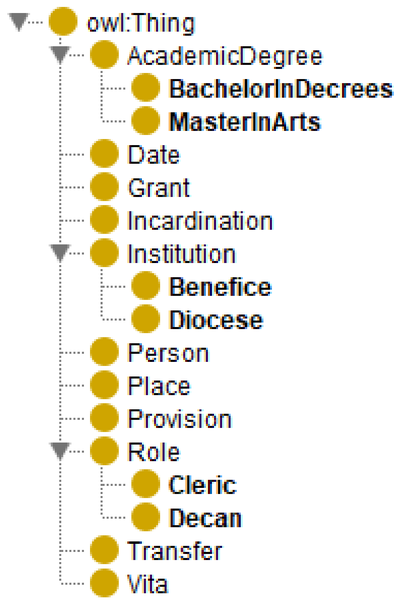

Abb. 3: TBox des generierten Ontologieschnipsels

Von Syntax zu Ontologie: Am Beispiel des Repertorium Germanicum

Die generierte Struktur bleibt häufig **flach**, da tiefere Hierarchien (z.B. via CIDOC-CRM) eine präzisere **Expertenannotation** im Prompt erfordern. Das Fragment wird vor der Integration syntaktisch und per Reasoner (z.B. Pellet) auf Konsistenz geprüft; zeigt sich ein Problem, korrigieren Fachwissenschaftler Text oder Ontologieschnipsel und starten den Prozess neu.

Validierung & Ausblick

Durch wiederkehrende Strukturen im RG lässt sich so eine konsistente, abfragbare Ontologie erstellen, etwa für FactGrid. Trotz Automatisierung bleibt jedoch Domänenexpertise unabdingbar, insbesondere bei größeren Quellentexten (Stichwort Segmentierung). Künftig wird im Projekt **HisQu** der Ansatz weiter erprobt und ausgehend von den Regesten des Repertorium Germanicums eine umfassende Ontologie des mittelalterlichen Kirchenrechts entwickelt.