Revisiting the process of Knowledge Graph generation with the integration of LLMs

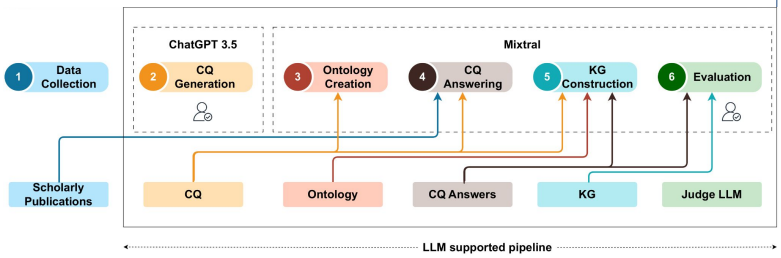
Vamsi Krishna Kommineni[1,2,3], Sheeba Samuel[1,4], Birgitta König-Ries[1,2,4]

[1] Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena, Germany
[2] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany
[3] Max Planck Institute for Biogeochemistry, Jena, Germany
[4] Michael Stifel Center Jena

Introduction

- Ontologies: foundational frameworks for describing and structuring domain knowledge and for constructing comprehensive knowledge graphs.
- Knowledge Graphs (KGs): interlink diverse pieces of information and facilitates sophisticated data analytics and reasoning.
- Knowledge engineering represents a collaborative and interdisciplinary effort, demanding the time and expertise of multiple stakeholders.
- Large Language Models (LLMs): ability to understand and generate human-like natural language.
- Leveraging LLM in Knowledge Engineering, particularly focusing on minimizing the time and human effort involved in these processes.
- We explore the (semi-)automatic construction of KGs facilitated by open-source LLMs.

Method



Data Collection:

- We conducted a systematic literature review to identify publications employing Deep Learning (DL) methods in biodiversity research based on keywords suggested by biodiversity experts.

CQ Generation:

- We prompted ChatGPT-3.5 to get abstract-level questions to describe the provenance of the results of DL pipelines.

Ontology Creation:

- Two-step strategy: (1) Extracted all concepts and their relationships from the CQs. (2) Constructed an ontology using the extracted concepts and relationships.

CQ Answering:

- Retrieved answers for all the CQs using the Retrieval-Augmented-Generation (RAG) approach from the first five selected biodiversity scholarly publications from our dataset that employed DL methods.

KG Construction:

- With the prompt, we instructed the LLM to extract key entities, relationships, and concepts from the answers and map them onto the ontology to generate the KG.

Evaluation:

- Two key outputs produced by the LLM were evaluated:
  - the generated CQ answers and the KG concepts that were automatically
  - extracted from these answers.
- Created multiple KGs in combination with two different prompts and two different RAG-generated CQ answers for five selected publications.

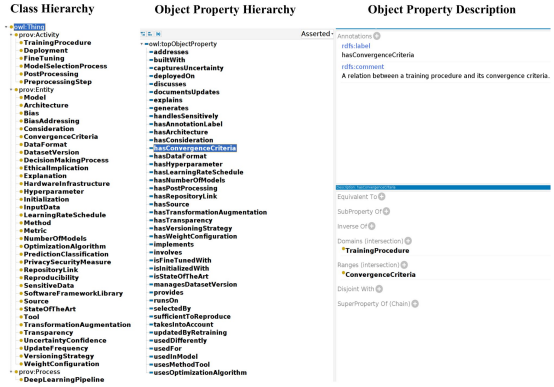
Results

We present here the results generated by the three stages of our pipeline:

Sample Competency Questions

- What data formats are used in the deep learning pipeline?
- What are the sources of input data for the deep learning pipeline?
- How was raw data collected in terms of methods and tools?
- Is the source code openly accessible, and if so, what is the repository link?
- Are there transformations or augmentations applied to the input data?
- Does the paper discuss data bias or ethical implications?
- What is the architecture of the deep learning model in the pipeline?
- What were the considerations in the model selection process?
- How many models are used in the pipeline?
- Are the models considered state-of-the-art?
- Which software frameworks or libraries are used to build the model?
- What hardware infrastructures are used for model training?
- What hyperparameters are used in the model?

DLProv Ontology



Knowledge Graph

Using the CQs, the answers from the CQs generated from the scholarly publications and the ontology, a KG was constructed from the method information of DL pipelines extracted from five scholarly publications using our pipeline.

```
dlprov:DeepLearningPipeline_1 rdf:type dlprov:DeepLearningPipeline ;
    dlprov:hasDataFormat dlprov:DataFormat_1 ;
    dlprov:hasDataFormat dlprov:DataFormat_2 ;
dlprov:DataFormat_1 rdf:type dlprov:DataFormat ;
    rdfs:label 'Audio Spectrogram'.

dlprov:DataFormat_2 rdf:type dlprov:DataFormat ;
    rdfs:label 'Image data'.
```

Conclusion

- We have explored the use of open-source LLMs for the creation of ontologies and knowledge graphs.
- With this, ontology and KG creation require significantly lower effort and less semantic web expertise.
- Plan to run our pipeline on different hardware and evaluate the results using different open-source LLMs to discern potential variations in results.
- Plan to explore methods for mapping the generated ontology with other ML/DL ontologies.

References:

1. Abdelmageed, N., L¨offler, F., Feddoul, L., Algergawy, A., Samuel, S., Gaikwad, J., Kazem, A., K¨onig-Ries, B.: BiodivNERE: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. Biodiversity Data Journal 10 (2022)
2. Ahmed, W., Kommineni, V.K., K¨onig-ries, B., Samuel, S.:How Reproducible are the Results Gained with the Help of Deep Learning Methods in Biodiversity Research?. Biodiversity Information Science and Standards, 7. (2023)
3. Pan, J.Z., et al.: Large Language Models and Knowledge Graphs: Opportunities and Challenges. Transactions on Graph Data and Knowledge 1(1), 2:1–2:38 (2023)
4. Neuhaus, F.: Ontologies in the era of large language models - a perspective. Appl. Ontology (2023)
5. https://github.com/fusion-jena/automatic-KG-creation-with-LLM