E.ON Energy Research Center

RWTH AACHEN UNIVERSITY

DAAD

Scientific Progress Report

BY

Charles Chukwuemeka William Emehel

91759575

PhD Student

RWTH Aachen University

November 2023

Ontology Learning
in
Smart Energy Grid

Table of Content

1.0 Introduction

This scientific report is on the progress of my research on knowledge engineering and semantic interoperability in the smart energy grid. It is the building and updating of a Knowledge Graph using ontology learning algorithms which is a process that involves: natural Language processing (NLP), machine learning (ML) and Knowledge representation and reasoning KRR) in the smart energy grid. I am currently developing and testing an ontology learning algorithm and an ontology learning architecture where the implementation can be carried out. I also intend to carry out test with Grover's Algorithm on an IBM-Quantum Experience computer so as to test the possibility of reducing the cost and time of using Graphics Processing Unit (GPU) for fast search lookups and the possibility for pertaining and fine-tuning anytime new information is available. This will subscribable for the Quantum Ontology Learning. I will also mentioned some of the development of Artificial Intelligence on Demand ontology that I developed and the plan to update its metadata for use as a reference standard in ontology development for any energy use case. The significance of the research is looked into based on the need for data integration with the growing amount of data on the traditional data bases and on the semantic web. Ontology is developed and learned to scale and bring quality to data before they are applied to training model for all the required analytics.

2.0 Knowledge Graph

Knowledge Graph is a graph of knowledge that contains interpretable context information with the aid of concept and relations description in hierarchies. Ontology engineering is the different activities that builds, maintains or carries out analytics or inference on ontology data. Ontology engineering helps man and machine to model data in a formal explicit and specified way for data sharing and conceptualization in a domain of interest. [1]. Since knowledge changes over time and domain always evolve, ontology engineering activities has to keep up with this change over the life cycle of the building of the ontology and provides the ontology developer an intelligent change model for expressing the local changes of this ontology in a descriptive way. One of the ontology engineering activities is alignment of links between RDF graphs as illustrated in the figure 1 below.

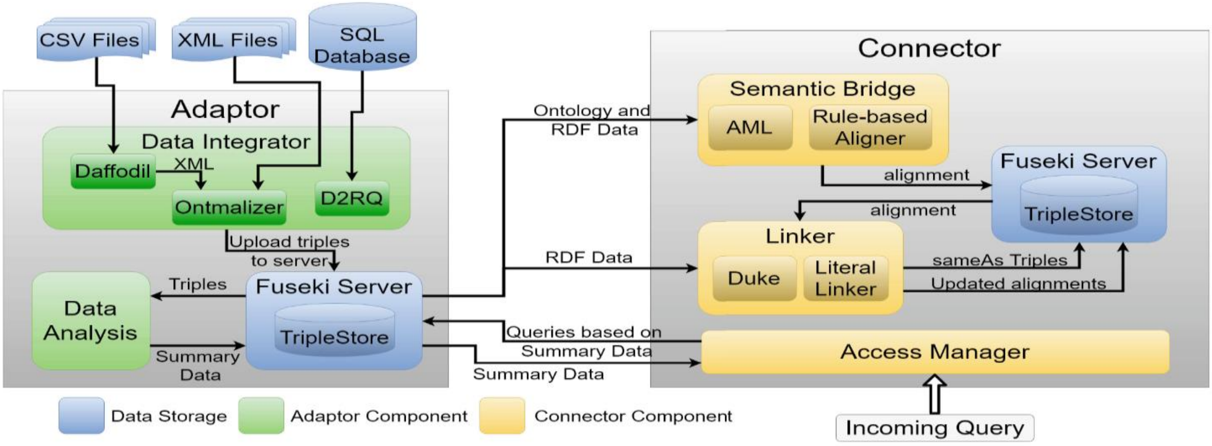

Figure 1 – Ontology Engineering Work flow [2]

Building Knowledge graph with ontology standard and schema is a way for ensuring or guaranteeing a realistic and convergeable multi domain knowledge graph. This will reduce noise and hallucination in making decision and automating a critical infrastructure like the electric grid.

3.0 Ontology Development

From my research I found out that for a realizable semantic interoperability to be achieved, the ontology development process need to follow standards. For instance I have adopted the Basic Formal Ontology BFO (ISO/IEC21838-4) for my top level ontology (TLO). This will help me achieve a correct-by-construction true converges to happen between different domains during ontology learning. One of the important process is to adopt single inheritance so as to be able to achieve: computational performance benefit, Aristotelian definitional structure, ontology structure management, automated ontology alignment and the possibility of achieving multiple inheritance only after reasoning or during inferencing. This is during the running of the application involving the AI agent making decision or in helping a decision support system or during actionable insight on the smart grid. The identified object concept is modeled as a subclass of bfo:object so as to follow the Aristotelian definitional structure for all child concepts on the cim controlled vocabulary concepts. This will enable automatic alignment for equivalent concept as well as achieving taxonomical subsumption on the cim path for all the cim concepts. [3]

To contribute to the smart grid journey towards the energy transition, a cyber-physical smart grid ontology (**cysgo**) with BFO as the TLO cysgo is being developed. This will kick start, with a seed or base ontology on which the learning ontology will be anchored. The seed ontology will be the reference e ontology which will use BFO as the TLO to achieve a realistic semantic interoperable ontology that can easily converge. Then the ontology learning process will generate learned ontologies as axioms or fact that will ride on the seed ontology. The seed ontology will be designed using the Content Ontology Design Patter Content-ODP or CP. The seed ontology is developed with Protégé while the ontology learning will be developed with Python from the loaded ontology exported from Protégé.

The essence of using an ontology schema based knowledge graph is to detect an advanced adversarial attacker that might decide to employ generative AI in creating an utopian or fake scenario that looks so real in other to evade surveillance and detection of its malicious activities. Since the cysgo ontology has ontology learning capability linked to the NVD and Att&ck Mitre data base, updates to the cysgo knowledge base will always happen when new threats is updated on the database. Also using the Grover's algorithm searching capability on the semantic web information will be extracted and updated on the cysgo ontology. Software application development following an object oriented approach always deals with state and behaviors of entities. The state estimation of the converter attributes of voltage, current, frequency, phase angle and power factor by the PMU will be extracted from the measurement data and these data matched with the ontology based knowledge graph. From the instance, from fig 1, when a malicious attacker carries out a threat due to an exploited vulnreabity that was initiated or discovered which exposes a smart grid asset. This could have an adverse impact on the grid. The reasoner could prvide an inference that will enable the detection of this threat before it occurs.

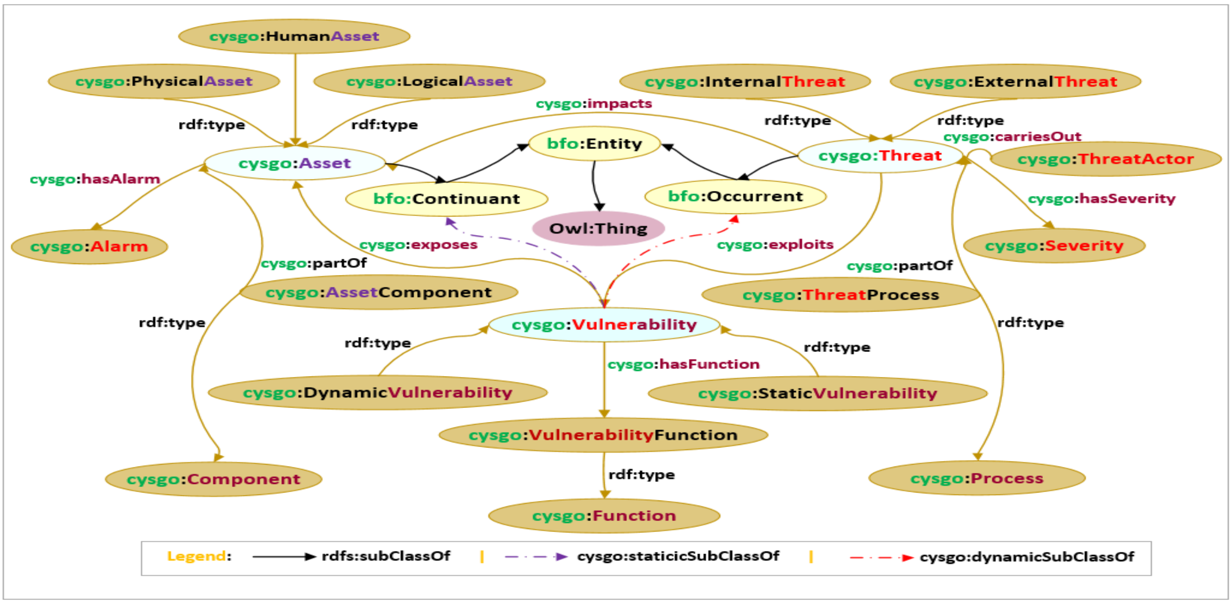

Figure 2 – cysgo Ontology main core entities

4.0 Ontology Learning

Ontology Learning is the use of Grover’s quantum algorithm and ressources to update an ontology for scalability and efficiency. The work of [4] proves the workability of ths approach. See uploaded team presenation on the introduction of this approach. Machine learning algorithms and techniques can be applied to ontology learning as well as for the maintenance of ontology and the performance of analytics on knowledge base. The work of [5] shows the use of machine learning algorithm for searching vectorized RDF data on the semantic web.

The work of [6] shows how a cluster of entity concept can be visualized through embedding learning of linked data. In demonstrating the use of machine learning for link prediction and concept classification the work of [7] and [8] shows that this can be applied in as part of ontology learning. Large corpus of data can be used to train graph neural network with an intermediate embedding being used for the training of the neural network. Distributed training is very is inevitable to handle large corpus and to shorten the training time.[9] Ontology Learning involves the processing of structure, semisturcture or strcture data resource for the purpose of automatically building an ontology or updating an existing ontology. The steps involved in ontology learning are:

- Natural Language Processing (Resource Processing Stage)
- Machine Learning or Deep Learning (semantic learning stage)
- Knowledge Representation and Reasoning (Formation of Ontology Schema and Reasoning)
- User Interface Management (Knowledge Graph Management)

I started my research with using unstructured dataset to use the information on it to update the ontologies developed above. The algorithm for the implementation was based on Jaccard similairity using the coefficient computed from the intersection and union of the characters in the terms from the dataset. The Jaccard index is obtained and the threshold is used to trigger the terms to be selected as related and this to be discarded. The result is then used to form an RDF triple and saved on a triple store as an RDF graph data. Then another approach was the use of information extraction algorithm with the spaCy NLP package. Then I used pre-trained transformer models in the resource processing statge.

Quantum Ontology learning in artificial intelligence is a scalable approach of efficiently scaling the ontology leaning to deal with large data and also have the capability of using large language model algorithms like self Attention Mechanism and Bidirectional Encoder Representations from Transfomers (BERT). The advantage of this approach can be applied to Data Space and Digital Twin research and applications. One of the complex part of ontology learning is the combination of statistics, probability and logic. The following discipline are the foundation of ontology learning:

- Philosophy
- Natural Language Processing
- Machine Learning
- Graph Theory
- Semantic Web
- Data Mining
- Cryptography
- Predicate Logic

An robust ontology learning architecture will ensure a versatile knowledge graph that will scale with the complexity of the information system in an intelligent automation infrastructure.

4.1 Ontology Learning Algorithm

The ontology learning algorithm will have input from unstructured dataset (***𝓓)*** from terms corpus (*tC*) from the web. The output will be and RDF Graph (***G***) that is composed of taxonomical hierarchies, mereological hierarchies and conceptual association from domain vocabularies. The seed Ontology which is developed with Protégé will serve as the base ontology for the learned ontology subgraph will anchor on top to form the output RDF Graph (***G***). As is typical with ontology learning with unstructured dataset, the first stage after crawling web endpoints are natural language processing (NLP).The Named Entity recognition of the subject and object terms as well as the taxonomical, mereological and conceptual association from the control vocabularies of the dataset will be processed. A reward function using the Q() function will be used to assign reward for the first stage before the machine learning stage where the relations will be further process to determine the type of subsumption. The third stage will be the application of Markov's logic Network (MLN) to determine the logical coherence of the triple. Then reasoning will be performed and the RDF triple returned as a graph. Algorithm below shows the step by step procedure of how it will be achieved.

```
Begin
    input: (tC(𝓓) ∧ sO) ∈ 𝒰
    output: RDF Graph (G)
    RDF dataset (𝓓) := empty dataset
    for each termination criteria not true:
        HTTP req := HTTP GET on uri
        nlp := parse(http(req(tC(𝓓))))
        for each uri ∈ default:
                ∃rf() :=  Q(seq(ADS(nlp(t))) * seq(ADS(nlp(m))))
                D[ co(uri)] := rf()
                Set links := co ( uris(G))
        return links
        uri := choose one uri from links
        RDF dataset 𝓓 := uri(Namespace)
        lO := RDF dataset 𝓓
       G := sO + lO
    return G
End
```

Algorithm 1 – Ontology learning algorithm

4.2 Ontology Learning Architecture

The algorithm in section 4.1 is implemented with Protégé and Python in the architecture in figure 3 and achieved by the Ontology Engineer. The seed ontology is developed in Protégé. This stage involves the creation of logical disjoints and partitioning of the axioms as well as application of the OWL constraints.The result is exported to python where the NLP part will be performed. The ontology learning part will be carried out using, OWLready2, RDFlib, PyGeometric, Tensorflow, Sklearn, Pandad, Numpy and matplotlib for plotting and visualization. Quality assurance is applied by using Aristotelian definitional Structure (**ADS**) and aslo to achieve single inheritance and computational performance benefit.Seed ontologies from saref, wordnet and SARGON where used to enrich the ontology learning process and to increase the number of named entities recognized and therefore increased the reward function during the NLP stage. Decision support group can use the information on the Knowledge Graph using the SPARQL Querry endpoint. After inferencing anomalies can be detected depending on what information is learned. Grover's algorithm is proposed to be implemented to scale the ontology learning learning process since its is expensive and takes time in the use of Graphic Processing Unit (GPUs) in pretraining and fine tunning large language models for enrichment of the seed knowledge graphs during the NLP stage. The

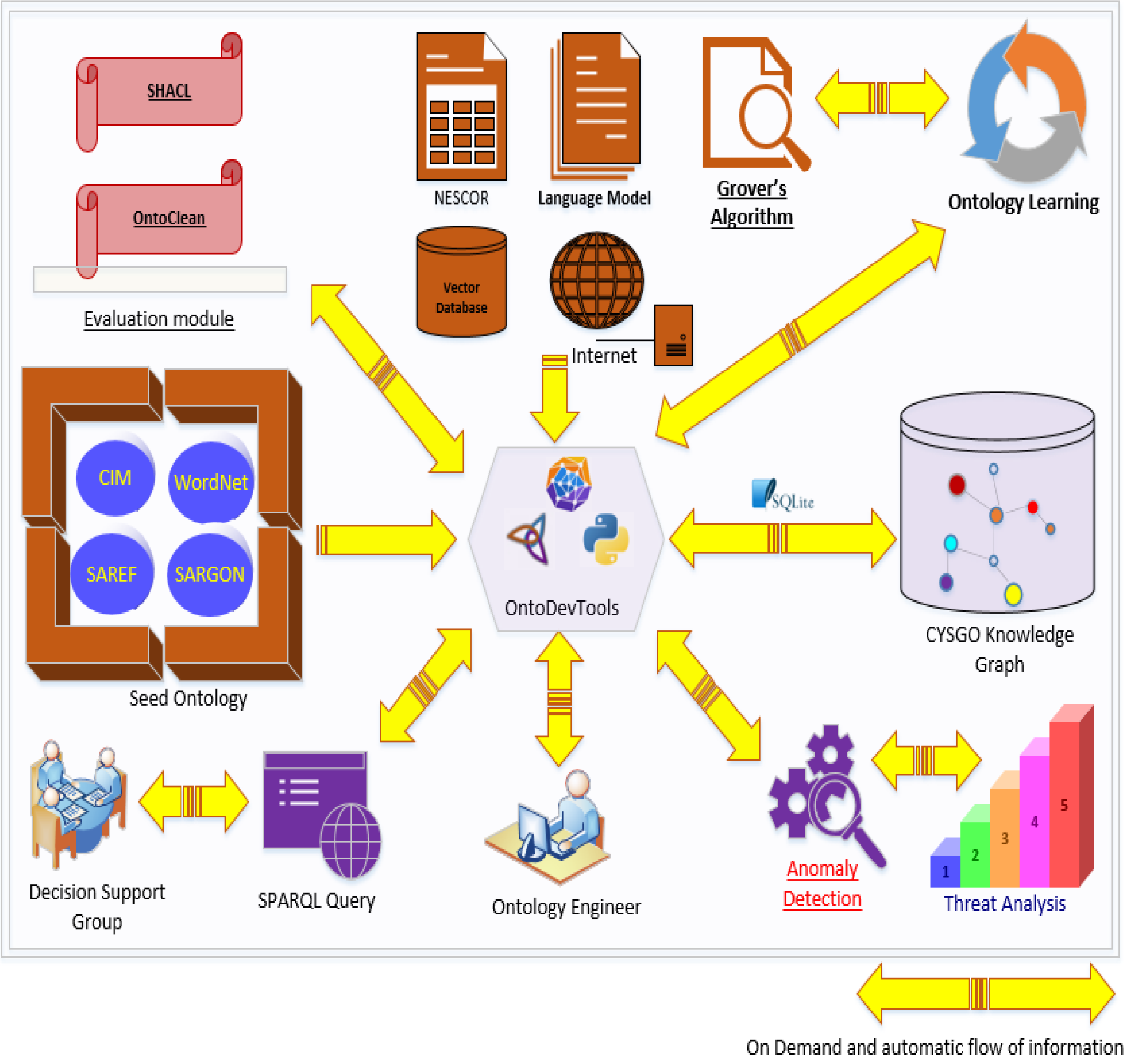

Figure 3 – Ontology learning architecture

5.0 Scalable Ontology Learning

The next step will be to test the use of Grover's algorithm within a rented space on IBM-Q quantum computer and see if speed and less cost could be achieved in lookup or search of corpus on the web during pre-training. This is why Quantum computing is mentioned in the figure 4

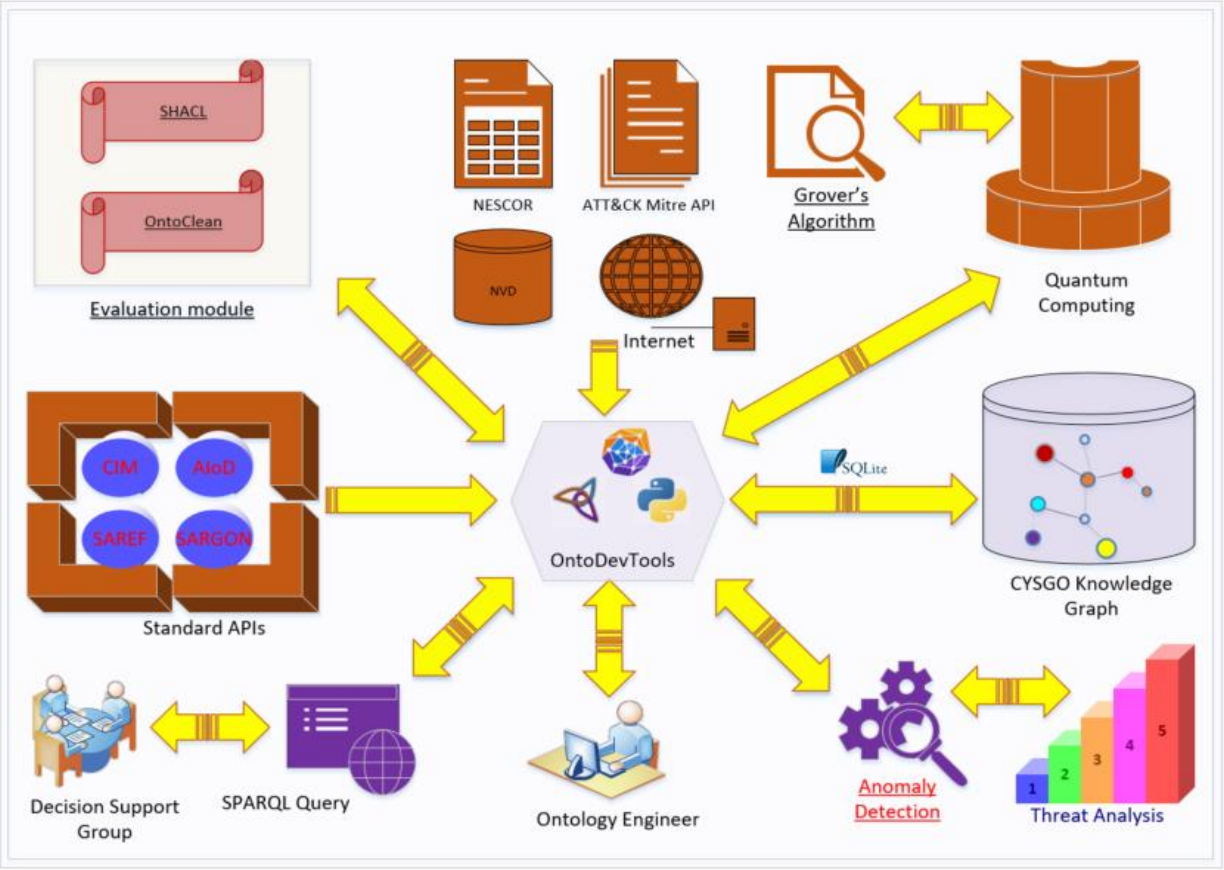

Figure 4 – Ontology learning architecture linked to Quantum Computing

6.0 Use Case Scenarios To Be Considered In My Research

The use case scenario is my proposed Artificial Intelligence energy ontology (AIEO) for the Artificial Intelligence on Demand (AIoD) ontology. This involves renewable and non-renewable energy as some of the core terms or entities of the ontology. The

7.0 Results

7.1 Ontology Learning RDF Graph Output

Figure 5 shows the seed ontology before the ontology learned with python using the OWLready 2 package. The A part is the learned ontology before the balanced distance metric (BDM) and information extraction algorithm were applied. These are algorithms for learning relations. The B part in figure 5 is the seed ontology from Protégé.

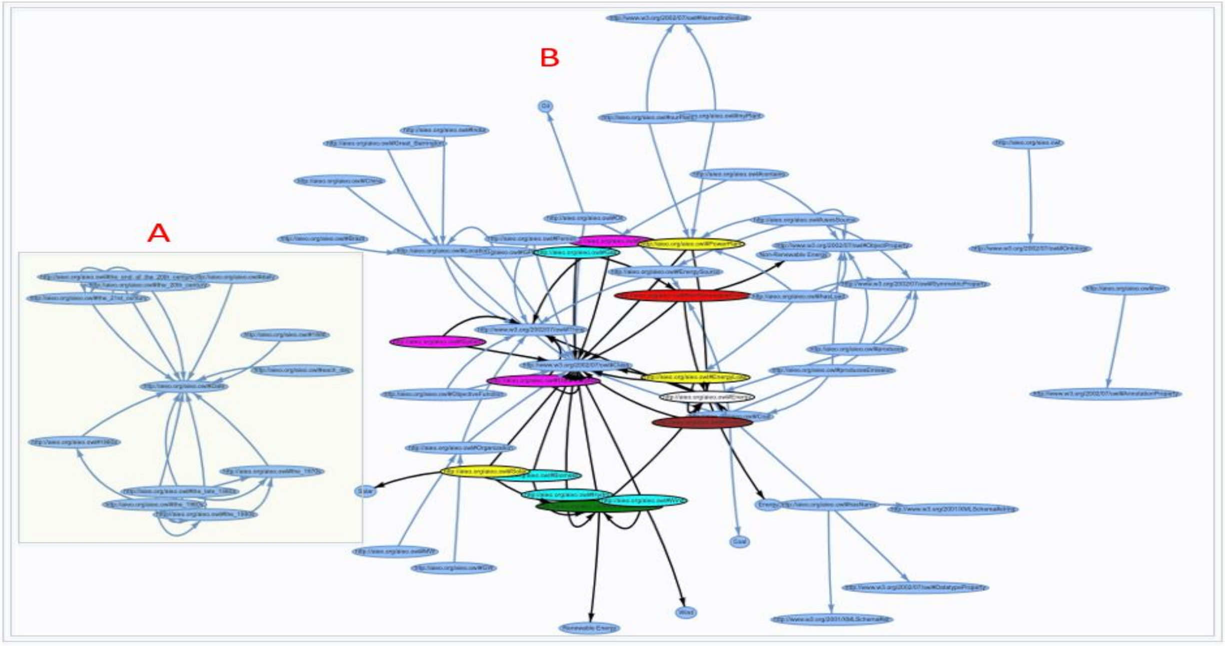

Figure 5 – AIEO Seed Ontology from Protégé.

After the balanced distance metric (BDM) and information extraction algorithm were introduced into the python implementation, there was a link between the learned ontology and the seed ontology. The A part is the seed ontology and the B part is the learned ontology showing the relations between the learned entities and the seed entities.

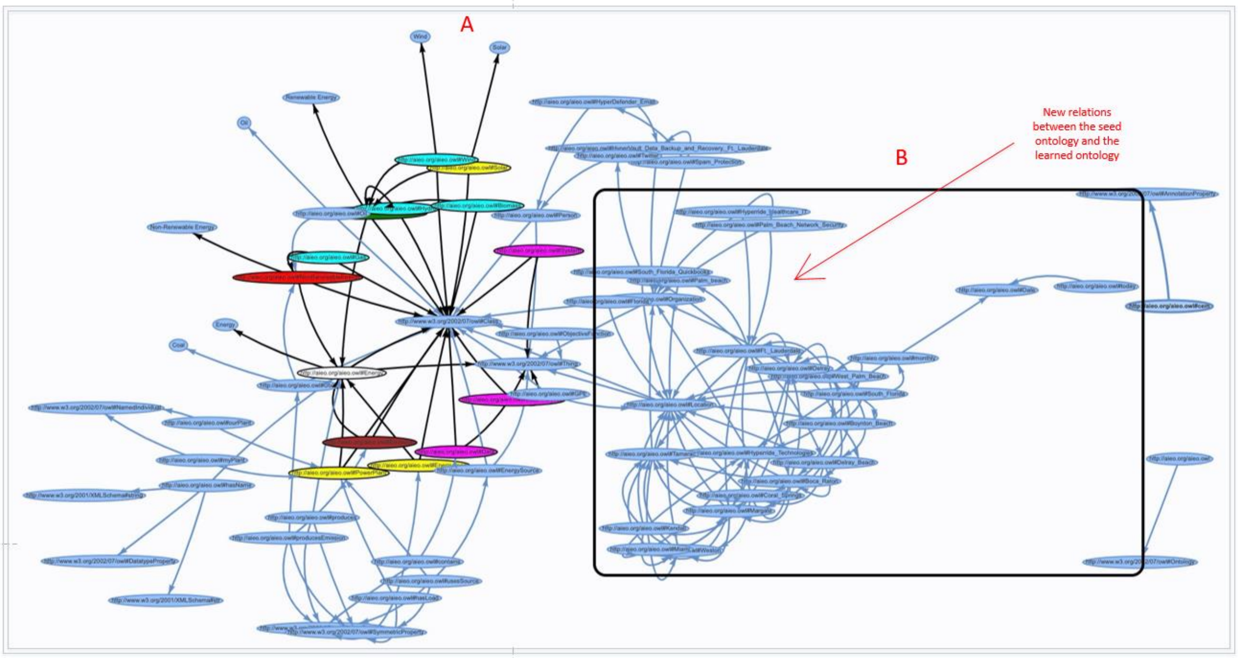
A
B
New relations between the seed ontology and the learned ontology
Wind
Solar
Renewable Energy
Oil
Non-Renewable Energy
Energy
Coal

Figure 6 – Seed Ontology plus ontology leaned form web Terms Corpus

The balanced distance metric (BDM) which shows the distance or relations between the terms in the corpus depicts that most of the relations formed are as a result of the proximity between the terms, lower distance values, and not necessarily due to other semantic distance that can be possible irrespective of the space between the terms or token. This shows the drawback of the of this algorithm which I intend to solve with algorithm 1.

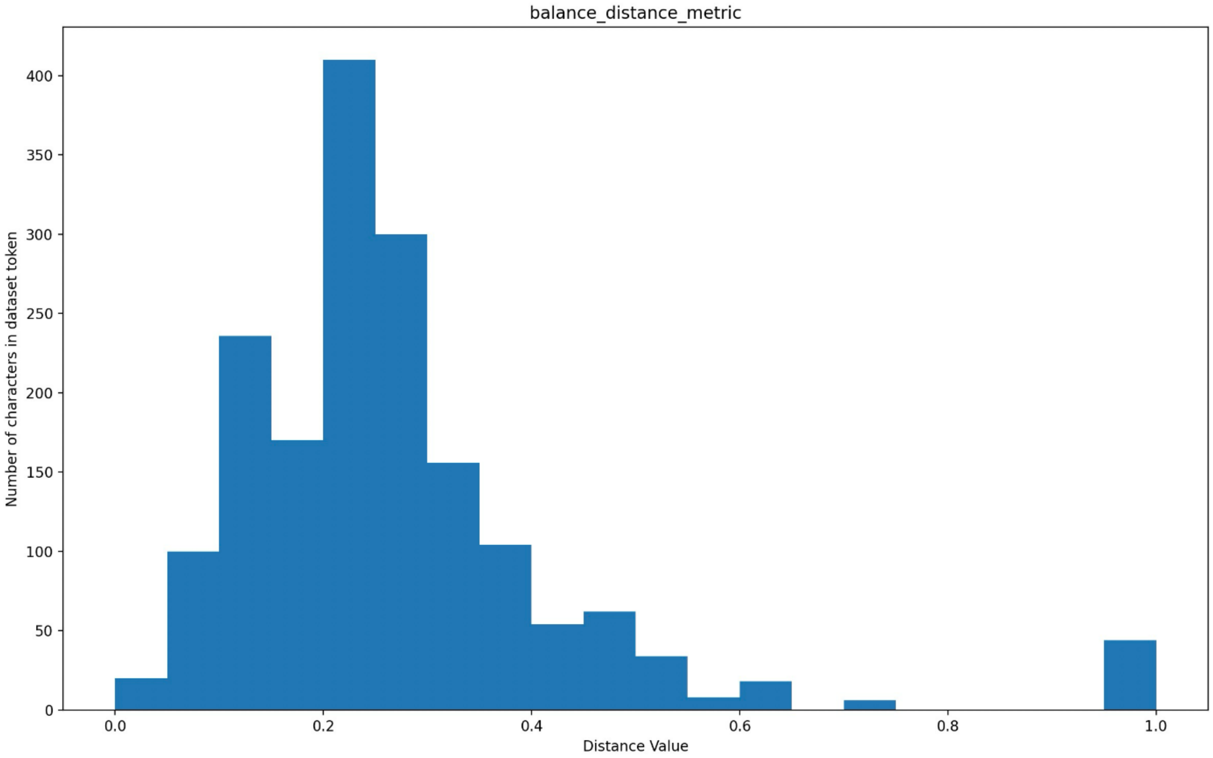

Figure 7– The balanced distance metric from the leaned relations.

Similarly, the use of The information extraction metrics (IEM)nshows similar distribution towards terms that a closer in distance instead of having semantic closeness irrespective of closeness of tokens. This also shows another drawback with the IEM algorithm which will be solved by algorithm 1in the ontology learning algorithm of section 4.1. See figure 8.

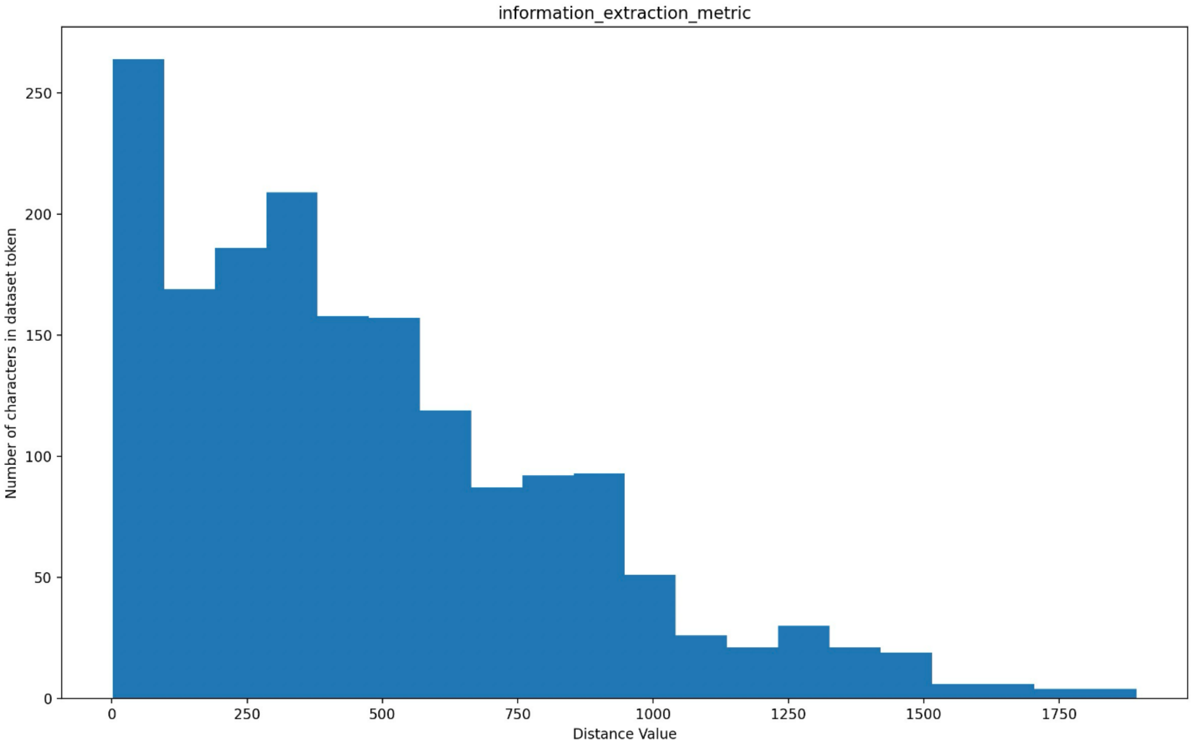

Figure 8– The balanced distance metric from the leaned relations.

7.2 Performance Analysis from Datasets Annotated with learned Ontology

The Performance Analysis of the datasets from the data model annotated by the learned Ontology is described in this section. 1000 samples of Synthetic data was generated based on the object properties and data properties from the learned ontology for prices of houses with high energy efficiency and those with low efficiency as well as for those inbetween. The data was used to train some regression and classification models and the performance were shown in figure 9, 10 ad 11.

Figure shows the performance metric of the datasets on thre regression modes: Random Forest Regressor (RFR), Gradient Boost Regressor (GBR) and Linear Regression (LinR). The idea is to see how a synthetic dataset trained on these models will show the performance with the standard metrics for regression: Mean Square Error (MSE), Root Mean Square Error (RMSE), R2 (for tracking negative scores), Mean Absolute error (MAE), Explained Variance (EV) and Median Absolute Error).

The extreme performance of the Linear regression model is most probably due to overfitting as a result of the synthetic data.Although the R2 score and the explained variance shows that that all three models are explaining the variance in the data very well and this also shows a perfect prediction capability.. GBR perfoms better than the RFR since Gradient Boosting is superior to the estimation of the number of trees in the forest algorithm.

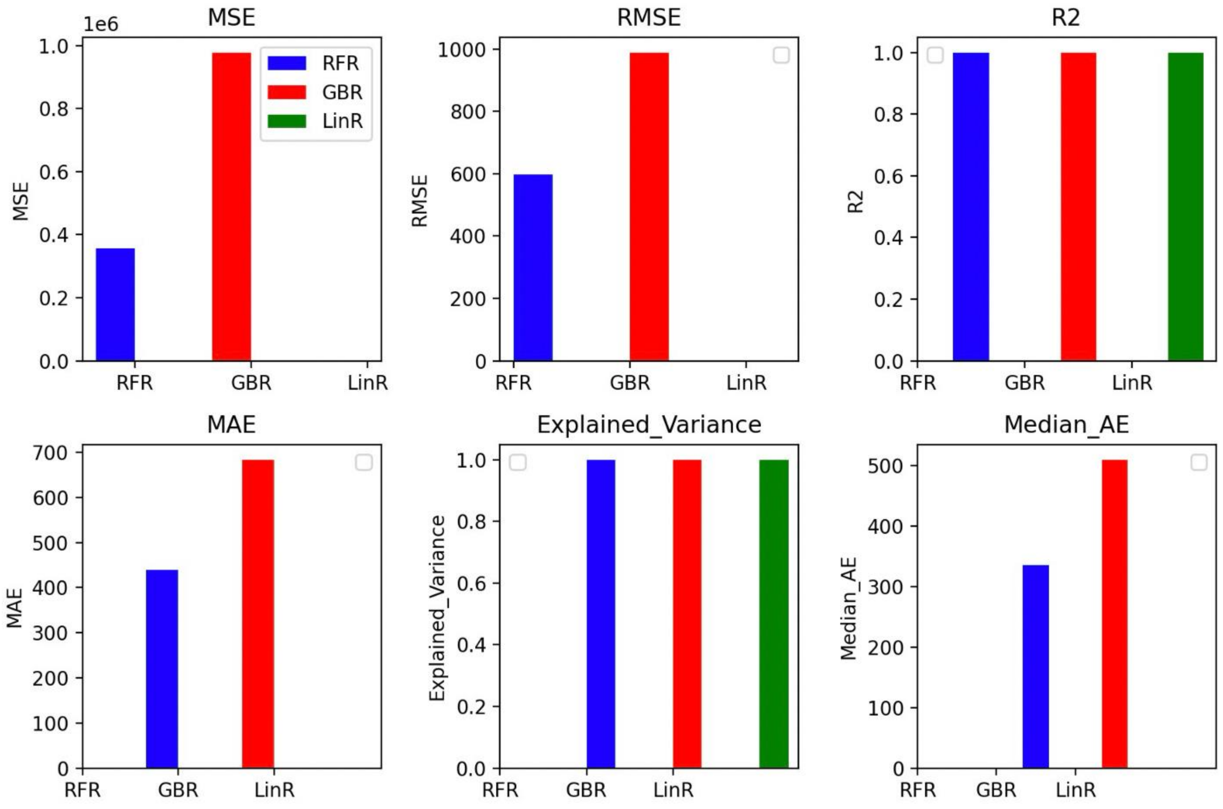

Figure 9– The balanced distance metric from the leaned relations.

The target variable of the dataset was binned or converted to categorical datasets and one-ho encoded. Classification. The models used are Random Forest Classifier (RFC), Gradient Boost Classifier (GBC), Logistic Regression (LogR) and Neural Ntework (NN). The Neural network has 100 hidden layers and the activation function is “ReLu” and “Adam” Solver with maximum iteration of 300 used.

Figure 10 shows the confusion matrix show the predicted and true value based on True positive, True negative False positive and false negative. The performance metric used in his analysis are Accuracy, Precision, Recall and F1-Score.

The models are evaluated based on key metrics: Accuracy, Precision, Recall, and F1 Score. These metrics provide insights into the models' performance in correctly classifying the house prices into the correct categories.

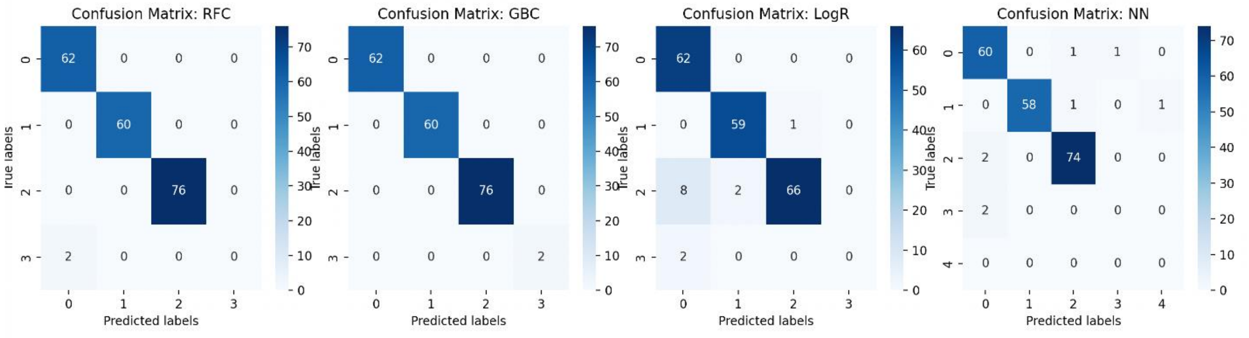

Figure 10– The balanced distance metric from the leaned relations.

From figure 11, it can be shown that RFC and GBC models achieved very high scores across all metrics, with RFC slightly lower than GBC. GBC reached perfect scores (1.0) in Accuracy, Precision, Recall, and F1 Score, indicating it perfectly classified all test instances into their correct price categories. RFC also showed high performance, indicating its efficacy in handling the dataset's complexity.

LogR Showed lower performance compared to RFC and GBC, with Accuracy, Precision, Recall, and F1 Score around 0.935 and below. This suggests that while LogR can model the dataset to a certain extent, it may struggle with the non-linear relationships present in the data compared to the more complex models.

-

The Neural Network (NN) handled by the multi-layer perceptron (MLP) Classifier achieved high scores, with an Accuracy of 0.96, Precision of 0.960625, Recall of 0.96, and F1 Score of 0.9601533494753833. This demonstrates the NN model's strong capability in handling complex, non-linear data structures and **relationships**. However, it did not achieve the perfection of the GBC model but performed comparably to RFC and significantly better than LogR.

For model selection, GBC appears to be the most effective model for this particular dataset, followed closely by the NN model, and then RFC. GBC's perfect scores indicate that it could capture the nuances of the dataset effectively. The NN model, while not reaching the same perfection, still demonstrates the potential of neural networks in handling complex classification tasks and classification task are very important in checking the performance of ontologies learned from unstructured datasets.

For models not performing at the highest level (e.g., LogR and to some extent RFC), further feature engineering, hyperparameters tuning, and exploring more complex or different neural network architectures could potentially improve results. For the NN model, ensuring convergence (as indicated by the warning during the training of the datasets) and adjusting hyperparameters like the learning rate, the number of layers, or the number of neurons could enhance performance.

While GBC and NN models offer high accuracy, their "black box" nature might pose challenges for interpretability. Techniques such as ontology learning and feature importance analysis for GBC and model explainability tools for NNs can help gain insights into how these models make predictions.

In summary, the choice of model should balance between predictive performance, computational efficiency, and the need for model interpretability depending on the application's requirements. GBC stands out for performance, but NN offers a flexible approach that, with further tuning, could potentially match or exceed this performance.

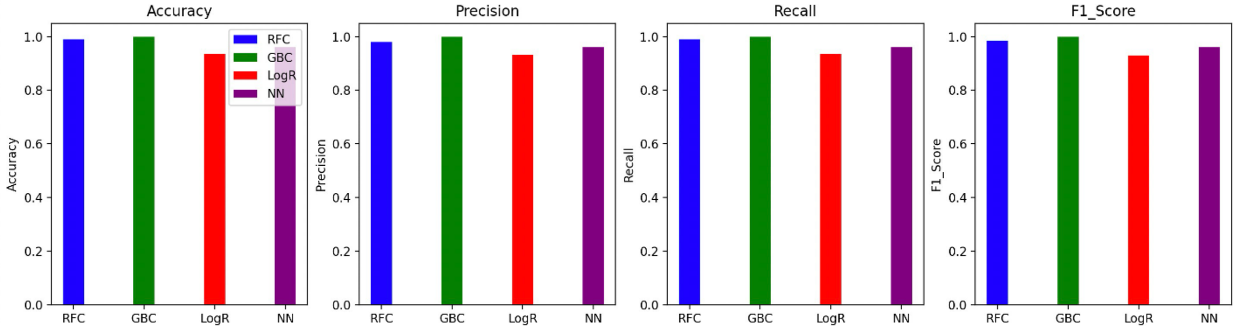

Figure 11– The balanced distance metric from the leaned relations.

8.0 Conclusion

This report and research demonstrate the potential of an ontology based knowledge graph in bringing quality and enhanced dataset feature before model training is embarked upon. This shows that knowledge graph is not only for data sharing and data exchange in information system.

In conclusion, while the reported metrics suggest high performance, the anomaly with the Linear Regression model's perfect scores necessitates a thorough review of the quality of the dataset (synthetic in this case) as well as data preprocessing, feature selection, and model evaluation steps to ensure robust and reliable predictive performance. A better algorithm for ontology learning will boost the feature selection or feature engineering and ence improve in the performance of the model.

The choice of model should balance between predictive performance, computational efficiency, and the need for model interpretability depending on the application's requirements. GBC stands out for performance, but NN offers a flexible approach that, with further tuning, could potentially match or exceed this performance.

8.1 Next Step

The next step is to develop and implement the algorithm in section 4.2 and use the ontology learned to refine a knowledge Graph that will be sued to provision a high quality data model that will be used to harmonize a dataset from different sources and the use of GridLAB-D or OpenModelica to model and simulate the distribution grid with Distributed Energy resoures, (DER), Electronic Vehicle and demand response mechanisms and modules module to co-simulate with the algorithm and architecture I developed above.

9.0 Papers Published and Proposed

9.1 Papers Published

Title	Publisher	Year	Contribution	Type	Link
"Deliverable: 2.3: I-NERGY Components, Services and Architect	Academia	30th May 2022	Wrote section 5.3: Building Blocks for Ontology	Review Paper	https://independent.academia.edu/CharlesEmehel

ure (first version)					
Development and Application of a New Ontology in the Context of Hybrid AC/DC Grids	IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications	17th November 2023	I Modeled the Use case scenario on the ontology developed by project parners and then validate and performed the reasoning and generated inference result. I also wrote the conclsuion and added the result on the Abstract.	Conference Paper	https://www.thinkmind.org/index.php?view=article&articleid=iaria_congress_2023_1_120_50112
…	…	…	…	…	…

9.2 Papers Proposed

Title	Publisher	Deadline	Contribution	Type	Link
Towards Resilient Energy Systems using A Cyber-physical Smart Grid Ontology	IEEE Transact	30th March 2024	Main Author	Journal	https://independent.academia.edu/CharlesEmehel
Automation of Energy Systems using Deep Reinforcement Learning with Ontology	Elsevier	30th May 2024	Main Author	Journal	https://www.thinkmind.org/index.php?view=article&articleid=iaria_congress_2023_1_120_50112

Feedback					
Towards Achieving Smart Energy Grids - Knowledge Graphs and Digital Twins	IEEE Access	30th July 2024	Main Author	Conference Paper	
Computational Ontology of Smart Energy Grid	IEEE Access	30th September 2024	Main Author	Conference Paper	
Ontology Learning for Knowledge Graph Enrichment in Smart Energy Grid	IEEE Access	30th November 2024	Main Author	Conference Paper	
Thesis	RWTH	30th December 2024	Main Author	Dissertation	

10.0 Challenges

Some of the challenges experienced during my research are :

- Lack of dataset due to GDPR
- Lack of validation tools for ontology learning development
- Information system modeling is usual different from the perspective of use case scenario

11.0 References

[1] M. Pritoni *et al.*, ‘Metadata Schemas and Ontologies for Building Energy Applications: A Critical Review and Use Case Analysis’, *Energies*, vol. 14, no. 7, p. 2024, Apr. 2021, doi: 10.3390/en14072024.

[2] F. Herrera, K. Matsui, and S. Rodríguez-González, Eds., *Distributed Computing and Artificial Intelligence, 16th International Conference*, vol. 1003. in Advances in Intelligent Systems and Computing, vol. 1003. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-23887-2.

[3] R. Arp, B. Smith, and A. D. Spear, *Building Ontologies with Basic Formal Ontology*. The MIT Press, 2015. doi: 10.7551/mitpress/9780262527811.001.0001.

[4] S. Naseem, ‘Dynamic ontologies evaluation framework using quantum perceptron neural network’, in *2015 International Conference on Open Source Systems & Technologies (ICOSST)*, Dec. 2015, pp. 151–157. doi: 10.1109/ICOSST.2015.7396419.

[5] A. S. Hadi, P. Fergus, C. Dobbins, and A. M. Al-Bakry, ‘A Machine Learning Algorithm for Searching Vectorised RDF Data’, in *2013 27th International Conference on Advanced Information Networking and Applications Workshops*, Mar. 2013, pp. 613–618. doi: 10.1109/WAINA.2013.204.

[6] Z. Wu, ‘Automatic Clustering and Visualization of Linked Data through Embedding Learning’, in *2019 International Conference on Intelligent Computing, Automation and Systems (ICICAS)*, Dec. 2019, pp. 778–781. doi: 10.1109/ICICAS48597.2019.00167.

[7] K. Juszczyszyn, G. Kołaczek, and D. Dudziak-Gajowiak, ‘Structural Analysis and Link Prediction in Dynamic Networks of Web Services’, in *2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, Jun. 2017, pp. 144–149. doi: 10.1109/WETICE.2017.55.

[8] P. Minervini, N. Fanizzi, C. d’Amato, and F. Esposito, ‘Scalable Learning of Entity and Predicate Embeddings for Knowledge Graph Completion’, in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2015, pp. 162–167. doi: 10.1109/ICMLA.2015.132.

[9] S. Wang, ‘DISTRIBUTED MACHINE LEARNING’, Accessed: Aug. 17, 2021. [Online]. Available: https://www.academia.edu/35877759/DISTRIBUTED_MACHINE_LEARNING_STANLEY_WANG_SOLUTION_ARCHITECT_TECH_LEAD_at_SWANG68