TOWARDS THE CURATION OF ENVIRONMENT-RELATED KNOWLEDGE GRAPHS: FINE-TUNING GENERAL-DOMAIN LANGUAGE MODELS FOR BIODIVERSITY NAMED ENTITY RECOGNITION

Geilah T. Tabanao
Department of Computer Science
University of the Philippines
Quezon City, Philippines
geilahtabanao67@gmail.com

Andrew Miguel V. Pagdanganan
Department of Computer Science
University of the Philippines Diliman
Quezon City, Philippines
avpagdanganan@up.edu.ph

Riza Batista-Navarro
Department of Computer Science
University of Manchester, UK
riza.batista@manchester.ac.uk

Roselyn S. Gabud
Department of Computer Science
University of the Philippines Diliman
Quezon City, Philippines
rsgabud@up.edu.ph

ABSTRACT

The availability of climate data fuels timely science-based climate actions. Providing policymakers and regulators with easy-to-digest, structured climate data, e.g., in the form of a knowledge graph, is critical to mitigating the adverse effects of climate change on the natural environment. Natural language processing (NLP) applications that employ Named Entity Recognition (NER) systems can aid in uncovering information hidden in millions of textual documents. In this paper, we evaluated the NER performance of transformer-based Bidirectional Encoder Representations from Transformers (BERT) models that were pre-trained on general-domain data. We fine-tuned BERT-based models on the COPIOUS dataset for the specialist task of biodiversity NER. Our experiments showed that our DeBERTa NER model demonstrated best performance, obtaining a micro-averaged F1-score of 84.18% based on entity-level evaluation. We employed our DeBERTa NER model in a biodiversity Information Extraction (IE) pipeline and applied it on the forestry compendium of the Centre for Agricultural and Biosciences International (CABI) Digital Library. We demonstrate that the pipeline enables the extraction of structured information on reproductive conditions and habitats of tree species.

1 INTRODUCTION

The decision-making process for climate action is systematically driven by information resources that are made available to climate policymakers and regulators. While much of our knowledge and investments about the natural world are reported and disseminated through scientific literature, this information is often buried within natural-language text, making it difficult to search, analyze, and manage compared to structured data, e.g., knowledge bases. Hence, there is a growing need for Information Extraction (IE) systems that have the capability to uncover information that is otherwise obscured within millions of scholarly articles. These IE systems can produce easily digestible data, e.g., knowledge graphs, necessary for timely science-based climate actions administered by policymakers.

Named Entity Recognition (NER) systems that can extract named entities relevant to the identification of species occurrence information (e.g., taxonomic names, geographic locations, temporal expressions, and habitats) in literature, together with Relation Extraction (RE) systems that identify semantic relationships between these named entities could form the basis of natural language

processing (NLP) applications such as the automatic curation of knowledge graphs. Such NLP applications could potentially provide climate stakeholders easy access to geography-based habitat information and time-specific reproductive conditions of tree species. Visual representations of this biodiversity data can aid in crafting efficient science-based land restoration and rehabilitation policies. These are crucial components in mitigating the effects of climate change.

In this paper, we sought to assess the NER performance of Bidirectional Encoder Representations from Transformer (BERT) models (Devlin et al., 2019) that were pre-trained on massive amounts of general-domain data, when fine-tuned on a domain-specific corpus. Specifically, we developed NER models by fine-tuning BERT-base (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), ALBERT (Lan et al., 2020), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2023) on the COPIOUS dataset (Nguyen et al., 2019), a corpus containing gold standard annotations of named entities relevant for extracting species occurrence information from biodiversity literature. Our results can inform biodiversity researchers in deciding whether pre-training language models on domain-specific data—which can be a costly and resource-intensive—is necessary to develop an NER tool, or if fine-tuning general-domain models might suffice.

2 Related Work

Most of the work on NER in the biodiversity domain has been focused on the extraction of taxonomic names, e.g., GNRD (Vanden Berghe et al., 2015), TaxoNERD (Le Guillarme & Thuiller, 2022) and GNFinder (Thessen et al., 2022). Meanwhile, the work of Nguyen et al. (2019) trained Conditional Random Field (CRF)–based and Bi-directional Long Short-Term Memory (Bi-LSTM)-based NER models on the COPIOUS corpus (Nguyen et al., 2019), in order to extract taxonomic names, geographic locations, temporal expressions, habitats, and person names from text. By far, COPIOUS is the largest biodiversity corpus with over 26K sentences from 668 documents sampled from the Biodiversity Heritage Library.[1] Importantly, the corpus provides gold standard annotations for more than 28K entities.

Recently, Abdelmageed et al. (2023) introduced BiodivBERT, a transformers-based language model that adopted the BERT architecture (Devlin et al., 2019). In their work, a BERT model was pre-trained on texts from the domain of life sciences, specifically, abstracts and full-text scholarly articles that were published by Elsevier and Springer. The resulting pre-trained model, BiodivBERT, was then fine-tuned for the NER task using biodiversity-focused corpora such as COPIOUS (Nguyen et al., 2019) and BiodivNERE (Abdelmageed et al., 2022), which include named entity (NE) annotations. Their experiments, however, showed that the improvement in the NER performance of BiodivBERT on the COPIOUS corpus is negligible, when compared to the performance of the original BERT model that was pre-trained on text drawn from the general domain. In this paper, we aim to investigate and compare the performance of various BERT-based models that were pre-trained on general-domain text, when they are fine-tuned for NER in the biodiversity domain.

3 Methods

BERT transformed the field of NLP by outperforming its predecessors and setting new performance standards across an array of NLP tasks including NER. In this work, we fine-tuned five types of BERT models. These are: **BERT-base** (`bert-base-cased`) (Devlin et al., 2019); **DistilBERT** (`distilbert-base-cased`) (Sanh et al., 2020); **ALBERT** (`albert-base-v2`) (Lan et al., 2020); **RoBERTa** (`roberta-base`) (Liu et al., 2019); and **DeBERTa** (`deberta-v3-base`) (He et al., 2023). All these models are readily available in Huggingface,[2] a publicly accessible repository of models.

The COPIOUS corpus was utilized in fine-tuning each of the above-mentioned BERT-based models. This corpus was chosen because it is the biggest NE-annotated corpus relevant to species occurrence information, and the results of the work by Abdelmageed et al. (2023) showed that this is the one dataset where pre-training a BERT model on domain-specific data, did not lead to any improved per-

[1] `https://www.biodiversitylibrary.org/`

[2] `https://huggingface.co/`

Table 1: F1-scores obtained by the BERT-based NER models on the COPIOUS test set. Key: Geo Loc = Geographic Location; Temp Expr = Temporal Expression.

NE Type	DistilBERT	ALBERT	BERT-base	RoBERTa	DeBERTa	BiodivBERT
Taxon	85.59	83.64	85.72	86.11	**87.60**	86.81
Geo Loc	85.62	84.16	86.74	**87.85**	87.58	86.74
Temp Exp	78.11	73.58	81.50	79.58	70.28	**82.59**
Habitat	69.91	65.70	66.21	68.76	**70.93**	66.99
Person	64.71	63.24	69.15	65.88	68.08	**69.32**
OVERALL	82.77	80.67	83.51	83.87	84.18	84.23

formance, thus prompting the question of whether other BERT-based models could perform better, even when pre-trained on general-domain data only.

In all our experiments,[3] we used the same hyperparameter settings employed by Abdelmageed et al. (2023) in their implementation,[4] to make our work comparable with theirs. Batch size was set to 8 for model fine-tuning, and set to 64 for evaluation. The number of epochs was set to 5 for all models. The number of warmup steps was set to 500, and weight decay was fixed at 0.01. We trained all the models using an NVIDIA T4 GPU in Google Colaboratory.

The performance of the models were assessed in terms of standard metrics such as precision, recall and F1-score, measured at the entity level with the help of the `seqeval` library[5]. It is worth noting that this is different from how Abdelmageed et al. (2023) evaluated their models, as they applied the metrics at the token level (rather than at the entity level). This means that matches between model predictions and gold standard annotations were counted based on tokens having the same label (e.g., `B-Taxon`, `I-Taxon` and `O` if following the IOB token tagging scheme), rather than full spans having matching boundaries and entity labels (e.g., `Taxon`). We opted for entity-level evaluation as it makes for a stricter matching strategy, and is considered to be the de-facto technique for evaluating models developed for sequence labelling tasks such as part-of-speech tagging, chunking and NER.

4 Evaluation Results and Discussion

All the general-domain, pre-trained BERT models described in Section 3 were fine-tuned for the biodiversity NER task using the training subset of the COPIOUS corpus. Table 1 presents the results of evaluating the fine-tuned models on the test subset of COPIOUS. Additionally, we provide the results of fine-tuning BiodivBERT, the BERT-based model pre-trained on life science documents, that was developed by Abdelmageed et al. (2023), to allow us to compare the performance of the general-domain models with that of a model pre-trained on biodiversity-relevant data. For brevity, we provide the F1-score obtained by each model for each NE type, and report overall performance in terms of micro-averaged F1-score. For more detailed results, i.e., the precision and recall obtained by each model, we refer the reader to Table 2 in the Appendix.

Lightweight versions of BERT, i.e., DistilBERT and ALBERT, obtained F1-scores that are lower than that of the original, BERT-base. This holds true for every NE type (except for the `Habitat` type, where DistilBERT performed better than BERT-base), as well as in terms of overall performance. RoBERTa, meanwhile, obtained an overall F1-score that is only marginally higher than BERT-base (i.e., with an improvement of 0.36). DeBERTa produced the best overall performance among the models that we fine-tuned, with a micro-averaged F1-score of 84.18. That both RoBERTa and DeBERTa outperformed BERT-base overall is not surprising, given that the two are improved versions of the original BERT model, with enhancements incorporated into the architecture and pre-training process.

In contrast to BERT-base which was pre-trained on 16GB of text from the English Wikipedia and the Google Books corpus, RoBERTa and DeBERTa were pre-trained on a much bigger amount of text (160GB), including additional data from OpenWebText, Stories, and the CC-News corpora. The data

[3] Code publicly available at `https://github.com/BiodivNER/BiodivNERModels`

[4] `https://github.com/fusion-jena/BiodivBERT`

[5] `https://github.com/chakki-works/seqeval`

used to pre-train RoBERTa and DeBERTa is thus 10-fold bigger in size, which likely contributed to the increased F1-scores of both models over BERT-base on Taxon (+0.39 and +1.88), Geographic Location (+0.84 and +1.11), and Habitat (+2.55 and +4.72) entities.

One can observe that BERT-base yields the highest F1-scores for entity types that are generic (i.e., those that very frequently appear even in general-domain documents), namely, Temporal Expression and Person. Presumably, the additional documents from the OpenWebText, Stories and CC-News corpora did not improve the model's vocabulary with respect to person names. Meanwhile, tokens pertaining to temporal expressions are relatively limited, as references to time and dates (e.g., *'last year'*, *'June 1950'*) tend to come from a constrained vocabulary. Thus, increasing the data for pre-training did not help in broadening the other models' ability to detect temporal expressions. It is worth noting that DistilBERT is the second best model for the Temporal Expression and Habitat types. Despite its down-scaled size, DistilBERT can outperform larger-sized models on some of the named entity types.

Looking at the rightmost column of Table 1, one can find the results of fine-tuning and evaluating BiodivBERT. We note that, although BiodivBERT was fine-tuned in exactly the same manner as Abdelmageed et al. (2023) did, the values of the performance metrics are different from those that they reported due to the fact that we employed entity-level rather than token-level evaluation (as explained in Section 3). In terms of overall performance, BiodivBERT is superior with a micro-averaged F1-score of 84.23%. However, the overall performance of DeBERTa is very close at 84.18%, which is impressive considering that this model was not pre-trained on domain-specific data. Even more impressive is the fact that DeBERTa's performance on the biodiversity-relevant NE types, Taxon and Habitat, is better than that of BiodivBERT (87.60% vs 86.81% for Taxon; 70.93% vs 66.99% for Habitat). Based on its superior performance on these biodiversity-specific NE types, DeBERTa was chosen as the optimal biodiversity NER model and was thus selected for integration into an IE pipeline for knowledge graph curation.

5 Curation of Environment-related Knowledge Graphs

A popular application of NER is the extraction of fine-grained information from text, that can then be leveraged to populate or curate structured databases. In this vein, we set out to explore the extent to which an IE pipeline underpinned by NER and relation extraction (RE) can curate a biodiversity-focused database, based on information buried within textual descriptions of various tree species in the Centre for Agricultural and Biosciences International (CABI) Digital Library.[6] Specifically, we integrated our best performing NER model into the pipeline, and applied an existing RE model (Gabud et al., 2023a) to extract information on the habitats and reproductive conditions of species in the CABI Library forestry compendium. This can potentially enable data-driven discovery of tree species' reproductive patterns and habitats.

Taking a corpus of CABI textual descriptions, our pipeline: (1) applies NER to extract mentions of geographic locations, habitats and temporal expressions; (2) applies RE to identify related habitats and geographic locations (i.e., habitat-geographic location relations) and related reproductive conditions and temporal expressions (i.e., reproductive condition-temporal expression relations); and (3) populates a graph database to store the related entities, to allow for querying and visualization. Further details are provided below.

5.1 The CABI Corpus

The forestry compendium[7] of the CABI Digital Library contains information on tree species that are found worldwide. This information is organized in the form of datasheets, whereby each datasheet contains verbose text describing a particular species in detail. In a datasheet, the following information on a species is typically included: morphological description, importance, distribution, growth, reproduction, phenology, habitat, and ecology. We selected 323 open-access datasheets on tree species to form our corpus. The documents were pre-processed using the Natural Language Toolkit (NLTK),[8] to split up paragraphs into individual sentences.

[6] https://www.cabidigitallibrary.org/

[7] https://www.cabidigitallibrary.org/product/QF

[8] https://www.nltk.org/api/nltk.tokenize.html

5.2 Named Entity Recognition

Utilizing our fine-tuned DeBERTa NER model (described in Section 3), we extracted names of geographic locations, temporal expressions and mentions of habitats. Additionally, we sought to extract expressions pertaining to the reproductive condition of a species, which serve as indicators of its reproductive behaviour. We observed that such expressions, e.g., *"fruited heavily"*, *"mass flowering"*, typically include certain verbs or nouns of a limited range. Thus, we implemented a simple dictionary-based NER method to capture such expressions.

5.3 Relation Extraction

We sought to identify tree species' reproductive conditions and their related temporal expressions, and habitats and their related geographic locations. Specifically, we applied RE on the named entities extracted by NER in the previous step and the sentences that contain them, in order to extract two types of relations: "*reproductive condition* `has_time` *temporal expression*" and "*habitat* `has_location` *geographic location*" relations. We employed our (previously reported) RE approach, that is a hybrid of rule-based and transformer-based methods Gabud et al. (2023b). Firstly, it makes use of regular expressions to identify related entities based on the syntax of the sentence in which they appear. Entity pairs that were predicted as unrelated are then presented to a transformers-based natural language inference (NLI) model, to detect any relations between the entities based on the semantics of the sentence.

5.4 Graph Database Curation

We applied the IE pipeline on the CABI corpus and extracted habitat-geographic location pairs, and reproductive condition-temporal expression pairs from every sentence. We stored the entities and relationships that were extracted by our IE pipeline in a knowledge graph using the Neo4j framework.[9] Neo4j facilitates the straightforward representation of entities as nodes and the relations between them as edges of a graph.

Our approach is a step towards creating intuitive visualizations of environment-related information sourced from unstructured text. Figure 1 in the Appendix shows the results of an example query aimed at retrieving species that were found in habitats located in Australia. Ultimately, the knowledge graph can support timely, science-driven decision-making by policymakers and regulators, helping to inform climate actions aimed at mitigating global climate change.

6 Conclusion

We demonstrated that general-domain, enhanced variants of the original BERT language model that were fine-tuned for the NER task on the COPIOUS corpus obtained performance that is comparable and competitive with that obtained by a BERT model pre-trained on domain-specific data. When integrated into an IE pipeline, our best performing NER model was able to facilitate the curation of a knowledge graph. This allows the retrieval and visualization of fine-grained information on tree species, which otherwise would have remained buried within natural-language descriptions. As part of our future work, larger versions of general-domain language models will be investigated and fine-tuned to evaluate the extent to which they can improve biodiversity NER performance.

References

Nora Abdelmageed, Felicitas Löffler, Leila Feddoul, Alsayed Algergawy, Sheeba Samuel, Jitendra Gaikwad, Anahita Kazem, and Birgitta König-Ries. BiodivNERE: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. *Biodiversity Data Journal*, 10:e89481, October 2022. ISSN 1314-2836. doi: 10.3897/BDJ.10.e89481. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9836593/`.

Nora Abdelmageed, Felicitas Löffler, and Birgitta König-Ries. BiodivBERT: a Pre-Trained Language Model for the Biodiversity Domain. In Atsuko Yamaguchi, Andrea Splendiani, M. Scott

[9] `https://neo4j.com/`

Marshall, Chris Baker, Jerven T. Bolleman, Albert Burger, Leyla Jael Castro, Ole Eigenbrod, Sabine Österle, Martin Romacker, and Andra Waagmeester (eds.), *14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4HCLS 2023), Basel, Switzerland, February 13-16, 2023*, volume 3415 of *CEUR Workshop Proceedings*, pp. 62–71. CEUR-WS.org, 2023. URL https://ceur-ws.org/Vol-3415/paper-7.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

Roselyn Gabud, Portia Lapitan, Vladimir Mariano, Eduardo Mendoza, Nelson Pampolina, Maria Art Antonette Clariño, and Riza Batista-Navarro. A Hybrid of Rule-based and Transformer-based Approaches for Relation Extraction in Biodiversity Literature. In *The Second Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning (Pan-DL), co-located with The 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 2023a.

Roselyn Gabud, Nelson Pampolina, Vladimir Mariano, and Riza Batista-Navarro. Extracting Reproductive Condition and Habitat Information from Text Using a Transformer-based Information Extraction Pipeline. *Biodiversity Information Science and Standards*, 7, 2023b. doi: 10.3897/biss.7.112505. URL https://www.proquest.com/docview/2864584119/abstract/D7D158ADA1DE4B5EPQ/1. Place: Sofia, Bulgaria Publisher: Pensoft Publishers Section: Conference Abstract.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, March 2023. URL http://arxiv.org/abs/2111.09543. arXiv:2111.09543 [cs].

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, February 2020. URL http://arxiv.org/abs/1909.11942. arXiv:1909.11942 [cs].

Nicolas Le Guillarme and Wilfried Thuiller. TaxoNERD: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, 13(3):625–641, 2022. ISSN 2041-210X. doi: 10.1111/2041-210X.13778. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13778. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13778.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL http://arxiv.org/abs/1907.11692. arXiv:1907.11692 [cs].

Nhung T.H. Nguyen, Roselyn S. Gabud, and Sophia Ananiadou. COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity Data Journal*, (7):e29626, January 2019. ISSN 1314-2828. doi: 10.3897/BDJ.7.e29626. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351503/.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, February 2020. URL http://arxiv.org/abs/1910.01108. arXiv:1910.01108 [cs].

Anne Thessen, Dmitry Mozzherin, David Shorthouse, and David Patterson. Improving the discoverability of biodiversity data using the Global Names Finder. *Biodiversity Information Science and Standards*, 6:e90026, December 2022. ISSN 2535-0897. doi: 10.3897/biss.6.90026. URL https://biss.pensoft.net/article/90026/. Publisher: Pensoft Publishers.

Edward Vanden Berghe, Gianpaolo Coro, Nicolas Bailly, Fabio Fiorellato, Caselyn Aldemita, Anton Ellenbroek, and Pasquale Pagano. Retrieving taxa names from large biodiversity data collections using a flexible matching workflow. *Ecological Informatics*, 28:29–41, July 2015. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2015.05.004. URL https://www.sciencedirect.com/science/article/pii/S1574954115000825.

A APPENDIX

Table 2: Precision and recall (P/R) obtained by the BERT-based NER models on the COPIOUS test set. Key: Geo Loc = Geographic Location; Temp Expr = Temporal Expression.

NE Type	DistilBERT	ALBERT	BERT-base	RoBERTa	DeBERTa	BiodivBERT
Taxon	86.90/84.33	83.15/83.11	86.55/84.90	86.25/85.97	87.98/87.23	86.91/86.71
Geo Loc	83.63/87.71	84.74/86.54	84.57/89.04	84.46/91.52	85.44/89.82	84.98/88.57
Temp Expr	78.71/77.53	71.37/65.90	81.34/81.65	78.20/81.00	68.86/71.76	81.68/83.52
Habitat	68.33/71.56	65.26/67.80	64.29/68.25	67.70/69.86	70.52/71.35	68.65/65.40
Person	71.63/59.00	73.09/57.71	72.73/65.90	69.53/62.60	72.60/64.09	72.20/66.67
OVERALL	83.53/82.02	81.40/79.95	83.60/83.43	83.13/84.63	84.11/84.25	84.13/84.33

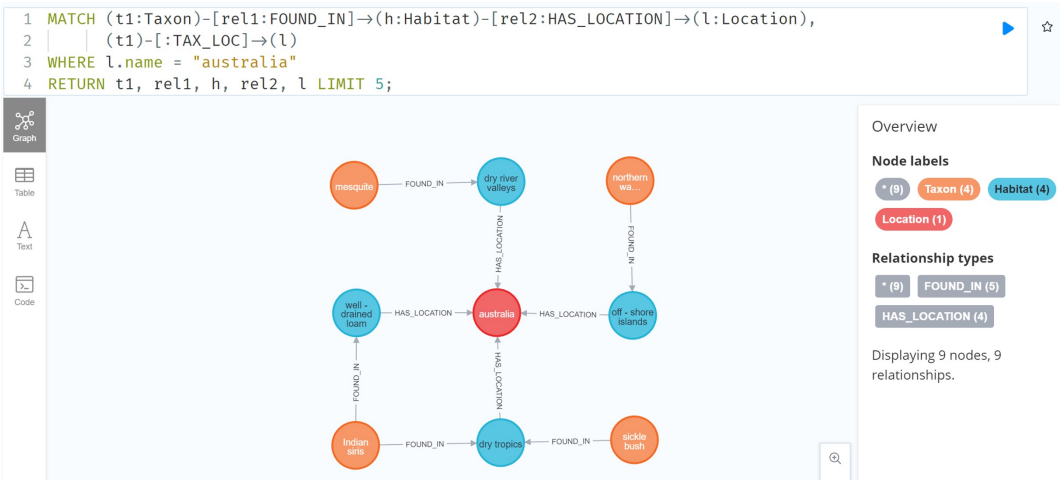

Figure 1: Results obtained by an example query (to retrieve species that were found in habitats located in Australia) that was run on the populated knowledge graph implemented using Neo4j.