Accessing the Capabilities of KGs and LLMs in Mapping Indicators within Sustainability Reporting Standards

Yuchen Zhou[1,*,†], Xin Gu[2,†], Junsheng Ding[2], Sirou Chen[2] and Alexander Perzylo[2]

[1]*Independent researcher*
[2]*Fortiss GmbH, An-Institute of Technical University of Munich: Munich, Bavaria, DE*

Abstract

Sustainability reporting is gaining importance in response to climate change and the pursuit of social sustainable development. In preparing these reports, sustainability managers are tasked with identifying sustainability indicators across multiple reporting standards. However, the challenge arises to locate the corresponding indicators in another standard through keyword searches due to the inconsistency of the naming conventions and classifications across different standards. Knowledge graphs(KGs) offer a promising solution for mapping the concepts of sustainability reporting from diverse standards. Nonetheless, traditional approaches to construct KGs are often time and resource intensive. In this context, the advanced natural language understanding capabilities of the Large Language Models (LLMs) could be explored to comprehend the reporting standards. Additionally, the rich knowledge structure of KGs could be leveraged to enhance the retrieval of relevant document snippets that describe the indicators within these standards. Accordingly, we propose a framework aimed at accessing the capabilities of KGs and LLMs in mapping indicators within sustainability reporting standards. This paper presents our framework, details two exploratory experiments, and discuses the preliminary results.

Keywords

Knowledge Graphs, Large Language Models, Sustainability Reporting Standards, Indicators

1. Introduction

The importance of sustainability reporting has arisen in response to the awareness of the social sustainable development [1]. Policymakers have established standards that include sustainability indicators (e.g. CO2 emissions), along with reporting requirements for companies to disclose their sustainability performance [2]. To ensure compliance with diverse jurisdictional requirements of their global customers, companies often utilize multiple standards in combination when preparing their sustainability reports [3]. Consequently, a single reporting value in a sustainability report references corresponding indicators from multiple standards.

NLP4KGC: 3rd International Workshop on Natural Language Processing for Knowledge Graph Creation, September 17, 2024, Amsterdam, Netherlands.
*Corresponding author.
†These authors contributed equally.
yuchen.zhou1994@outlook.com (Y. Zhou); gu@fortiss.org (X. Gu); ding@fortiss.org (J. Ding); schen@fortiss.org (S. Chen); perzylo@fortiss.org (A. Perzylo)
0009-0003-0305-9622 (Y. Zhou); 0009-0008-6143-4028 (X. Gu); 0009-0001-1522-4213 (J. Ding); 0009-0005-8989-2256 (S. Chen); 0000-0002-5881-3608 (A. Perzylo)

In practice, sustainability managers are tasked with identifying overlapping indicators among the multitude of standard documents. However, the challenge arises when trying to locate the corresponding indicators in another standard through keyword searches due to the inconsistency in naming conventions and classifications across different standards. Furthermore, the reporting requirements for the same indicator can differ, which requires sustainability managers to invest considerable time in re-establishing connections between indicators whenever regulations are updated. For example, the Global Reporting Initiative (GRI) standards require to disclose the "total fuel consumption within the organization from renewable sources" in "Joules or multiples" under the "Energy" topic, while the European Sustainability Reporting Standards (ESRS) mandates reporting the "total energy consumption, disaggregated by: fuel consumption for renewable sources" in "MWh" under the "Climate change" topic.

As it is challenging to find the corresponding indicators across multiple standards through direct keyword search, providing the mappings of indicators enhances interoperability among reporting standards. In literature, several studies have proposed the use of Knowledge Graphs (KGs) and ontologies to represent and unify indicators from diverse standards [4, 5, 6]. Nonetheless, traditional approaches for Knowledge Graph Construction (KGC) are time-consuming and struggle to manage the dynamically changing nature of the real world, particularly the regulatory updates [7]. Large Language Models (LLMs) demonstrate impressive capabilities in various natural language processing and tasks [8]. While most studies focus on LLM for KGC on common-sense knowledge, research in the sustainability reporting domain remains limited. Moreover, as these standards also encompass domain-specific concepts, such as Scope 1, 2, and 3 CO2 emissions, it is of interest to evaluate LLMs' capability in KGC tasks for those domain-specific concepts. Additionally, identifying the relevant document snippets that describe corresponding indicators from another standard remains daunting, given that the reliance on keyword searches can often lead to suboptimal performance. In light of this, KGs equipped with richer logical structures and higher knowledge density [7] could be used to enhance the retrieval process. Utilizing KGs incorporating additional layers of knowledge in the retrieval process, rather than relying solely on keywords, could improve the accuracy and efficiency of identifying matching indicators across standards. Consequently, it is worthwhile to investigate the potential of the subgraphs, extracted from the constructed KGs, in the context of retrieving mapping indicators from other standards.

Therefore, we came up with the following two research questions:

- Can we use KGs to retrieve the relevant documents describing the matched indicators from another standard?
- Can we use LLMs to construct a domain KG for mapping sustainability reporting standards?

To address the research questions, we developed a framework to access the capability of KGs and LLMs for mapping sustainability reporting standards, as detailed in Section 3. To evaluate each step and component of our framework, we are conducting several ongoing experiments. In this paper, we elaborate on two early-stage experiments[1] and briefly discuss the preliminary results in Section 4. Finally, we introduce our future work in Section 5.

[1]The dataset, prompts, code, and experiment results are available online at https://github.com/OntoSustain/KGLLM

2. Related Works

Several studies have proposed the use of KGs and ontologies to represent indicators from diverse sustainability reporting standards. [4] presented an ontology for representing information about sustainability indicator systems. Similarly, Diamantini et al.[5] constructed a knowledge graph based on the developed ontology to represent shared indicators from two sustainability reporting standards. In our previous work [6], we constructed a knowledge graph to model quantitative environmental indicators from the GRI and ESRS standards. The indicators derived from both standards were represented as instances of the *rso:Indicator* class. Key properties include, for example, linking an indicator to the unit of measure (hasUnit), the quantity kind it measures (hasQuantityKind), the sustainability topic (hasTopic) and object it measures (hasMeasurementVariable). Additionally, ontology-based mapping rules were designed to map indicators from GRI and ESRS by comparing the values of certain properties.

Comprehensive surveys of Large Language Model (LLM)-augmented Knowledge Graph Construction (KGC) tasks, including entity discovery and relation extraction were presented in [7, 8], with the latter emphasizes the use of LLMs in instruction-driven knowledge extraction tasks. [9] conducted an in-depth evaluation of ChatGPT's capability in general KGC and identified limitations in dealing with domain-specific data. Regarding domain-specific KGC, [10] investigated LLMs' application in the biomedical domain, exploring strategies to enhance their performance for Named Entity Recognition tasks. [11] explored (semi-)automatic construction of a Knowledge Graph in the biodiversity domain, starting with collecting competency questions to creating an ontology and filling data into it. Both studies [9, 10] highlighted the importance of meticulously designed prompts through prompt engineering to improve outcomes. In the sustainability domain, a notable work is [12], where the authors examine specific cases of entity and relation extraction for capturing information about public communication around sustainable development goals using various LLMs.

Recently, the Retrieval-Augmented Generation (RAG) approach has achieved state-of-the-art (SOTA) performances in many Natural Language Processing (NLP) tasks, garnering significant attention from the computational linguistics community [13]. A comprehensive survey paper provided an overview of Retrieval-Augmented Language Models (RALMs), discussing essential components such as Retrievers, Language Models, and Augmentations, and how their interactions result in diverse model structures and applications [14]. The authors highlighted the application of RALMs in Natural Language Understanding (NLU) tasks, such as Knowledge Graph Completion. From an application perspective, [15] proposed a pipeline for fine-tuning and RAG, presenting the trade-offs between both approaches for multiple popular LLMs on an agricultural dataset. This work paves the way for further applications of LLMs in other industrial domains. Additionally, [16] explored the existing constraints of RAG pipelines and introduced methodologies for enhancing text retrieval to refine the RAG process on financial documents.

3. Method

Our objective in mapping sustainability standards is to determine the exact correspondence between indicators from different standards. Specifically, we aim to identify Indicator B within Standard B that exactly matches a given Indicator A from Standard A. Building on our prior research[6], we have established criteria for determining an exact mapping between two indicators. Indicator B is considered to match Indicator A if it: 1) measures the same sustainability topic; 2) pertains to the same sustainability object; 3) evaluates the same quantity kind. To resolve our research questions, we designed a framework comprising four steps: Retrieval, Extraction, Mapping, and Construction, as illustrated in Figure 1. Below, we elaborate on each step in detail.

- ***Retrieval:*** Given that sustainability reporting standards often span over 1,000 pages of PDF files, our initial step is to refine the search scope and retrieve the most relevant document snippets from Standard B using a RAG framework. We first segmented the entire document of Standard B into text chunks, where each chunk corresponds to one page. We then utilized a subgraph of Indicator A extracted from our previous work as input for retrieving pertinent text chunks from Standard B. This subgraph encapsulates detailed information about Indicator A, such as the sustainability topic it addresses, the object it measures, and the quantity kind it measures. The exemplary subgraphs are available at our GitHub repository. Simply using the name of Indicator A for retrieval may yield suboptimal results, due to insufficient search information. Therefore, we aim to leverage the knowledge density of the KG to enhance the search process and improve the retrieval of relevant chunks.

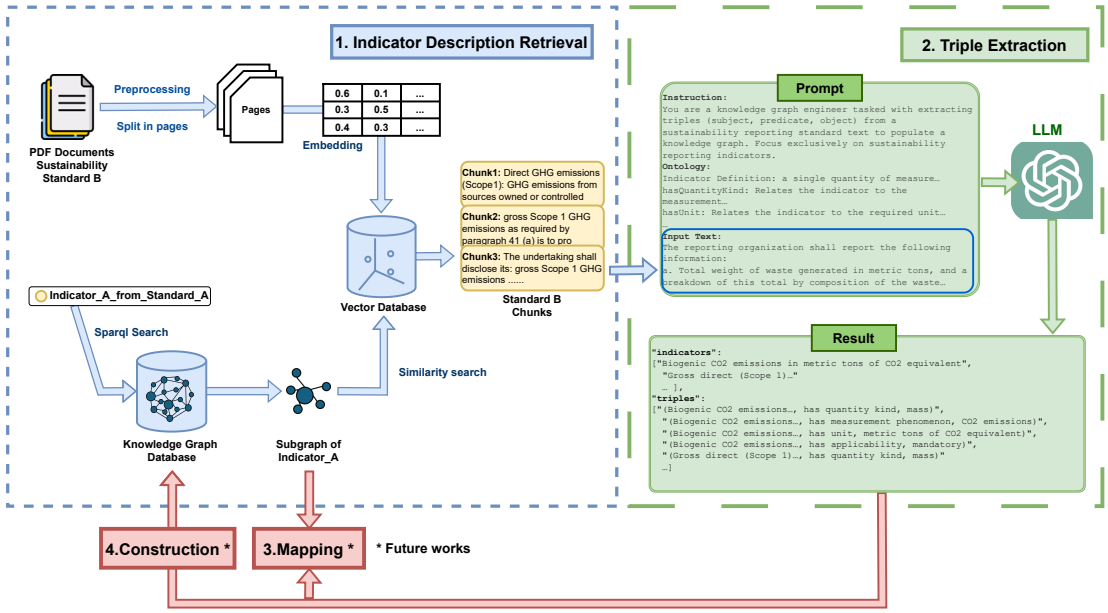

Figure 1: Framework overview. Step 1 and 2 are accomplished, while step 3 and 4 (marked with *) are planned work.

- ***Extraction:*** In this step, we harness the natural language understanding capabilities of LLMs to perform entity recognition and relation extraction tasks. We formulate few-shot prompts tailored for LLMs, generating structured text that organizes indicators and triples in a predefined format. Each prompt comprises four main sections: instructions, retrieved text chunks from the retrieval step, a list of classes and relationships, and their descriptions and definitions. The instruction section delineates the role of the LLM, task specifics, and the desired format of the output. The extraction tasks are defined as extracting entities (e.g., entities of the *Indicator*, *Unit*, *QuantityKind* classes) and relations (e.g. *hasUnit*, *hasQuantityKind*, *hasMeasurementPhenomenon*, etc.) from the given text chunk. Given that these entities and relations encompass domain-specific concepts, we employ SPARQL queries to extract their definitions from the predefined Sustainability Reporting Standards ontology (RSO[2]). This process equips the LLM with domain-specific knowledge that is crucial for accurate extraction.
- ***Mapping:*** The initial retrieval step in our process does not always ensure that all retrieved document segments are relevant. Consequently, the indicators extracted from non-relevant documents might not align perfectly with the target Indicator A. To address this, we have developed a mapping step that serves to eliminate biased indicators. After extracting the desired information for each mapping candidate in a structured format (triples), we prompt the LLM to compare the properties of Indicator A with those of each candidate, guided by a predefined mapping rule established in our previous work, outlined in the problem statement.
- ***Construction:*** In this step, we use LLMs as KG builders to construct a KG in the RDF/OWL formalization based on the extracted triples from the last step. The constructed KG, using classes and properties from the designed ontology can represent indicators from multiple sustainability reporting standards, as well as their mapping relationships. The validated KG will be stored in a knowledge base with appropriate interfaces to be used flexibly.

4. Pilot Experiment and Discussion

We designed several experiments to evaluate each step individually and identify any weaknesses in the overall framework. Since our work is still in progress, we first present two pilot experiments and discuss their preliminary results in this section. Experiment 1 evaluates the Retrieval step, and Experiment 2 assesses the Extraction step.

Experiment 1 assesses the performance of the Retrieval step. The task involved retrieving the most relevant text chunks from Standard B documents containing potential mapping candidates for a specific Indicator A from Standard A. We constructed the RAG pipeline using the Langchain framework, employing GPT-4 and text-embedding-3-small models. For each experimental Indicator A, we used the subgraph extracted by a SPARQL query from the KG in our previous work as input. The desired output was the chunk numbers from Standard B that contained potential mapping candidates. To evaluate the results, we compared the outputs generated by the LLM with those determined by human experts. Since the expected number of results varies for each indicator, we conducted multiple experiments, setting the number of retrieved chunks

[2]The RSO ontology is available at https://github.com/OntoSustain/RSO

Table 1
Retrieved chunk numbers from Standard B for two exemplary indicators from Standard A, where x represents the number of to be retrieved chunks and n equals to the number of the ground truth.

ID	Indicator Name	Ground Truth	Results when x= n	Results when x=2n
IN1	Total weight of renewable materials that are used to produce and package the organization's primary products and services	#7,#17,#12 (n=3)	#5,#9,#12	#5,#9,#12,#17,#16,#18
IN2	Total fuel consumption within the organization from renewable sources	#10,#29,#30,#19 (n=4)	#29,#30,#10,#19	-

to n, 2n, and 3n for each round, where n represents the number of the ground truth instances for each indicator.

As an early-stage experiment, we first searched for relevant chunks of exact matching indicator in the ESRS E1 and E5 standards for six GRI indicators derived from GRI 301, 302, and 305. The matched indicators were identified in our previous work, and the corresponding chunks were identified by human annotators. The complete experimental results are available at our GitHub repository. Table 1 shows two exemplary results, where all relevant chunks were found at x=n for IN2, at x=2n for IN4, IN5, and IN6, and at x=3n for IN1 and IN3. Although the overall results are satisfactory, two significant issues were identified. First, the model failed to distinguish the different semantic meanings of the word "material" in the standards. It is a noun meaning "raw resources" in the indicator name, and it is an adjective means "significant" when it appears in phrases such as "material impact" or "material topic". Therefore, the model retrieved many false positive results for IN1. Furthermore, the results show that a number of relevant chunks were only found when x=3n, which may require processing a large number of tokens to find all relevant chunks. The issue of inaccuracies were also identified in [16], noting that irrelevant answers often stem from suboptimal text chunk retrieval by RAG. In our framework, indicators in irrelevant chunks will be pruned by the subsequent mapping step, and this may not directly affect the final results. However, improving the retriever can effectively reduce the number of tokens processed and enhance overall efficiency.

Experiment 2 was designed to assess LLM capabilities for conducting entity and relation extraction tasks. The objective was to extract indicator entities along with four types of properties for each entity from the given input text: hasQuantityKind, hasUnit, hasMeasurementPhenomenon, and hasApplicability. The prompt we developed contained four parts: instructions, class and property explanations, few-shot examples, and the input text. The class and property explanations were derived from our previous work. The few-shot examples were also sampled from our KG and reformulated for the prompt. We conducted the experiment using the GPT-4 model. To evaluate the results, we compared the extracted entities and triples with those annotated by human experts. We conducted Experiment 2 for input text derived from GRI standards 302, 305, and 306. In total, 52 indicators and 214 triples labeled by human annotators were used as the ground truth. By manually comparing the results derived from the model with

Table 2
Precision, Recall, and F1-Score for Experiment 2

Task	Precision	Recall	F1-Score
Indicator Extraction	0.73	0.67	0.70
Relation Extraction	0.67	0.58	0.62

the ground truth, we computed the precision, recall, and F1-scores using the confusion matrix for entity and relation extraction. These results are shown in Table 2. From the results, two obvious issues were observed:

Wrong indicator entities and missing indicator entities. This is a well-known NER problem in the literature. The LLM model tends to extract wrong indicators that do not conform to the provided indicator definition. For example, “Types of energy included in the reductions” cannot be recognized as a "quantitative sustainability indicator". Meanwhile, the LLM model failed to recognize certain indicators in the input text. For example, we identify three separate indicators from the text “Production, imports, and exports of ODS in metric tons of CFC-11 equivalent”, namely “production of ODS”, “imports of ODS”, and “exports of ODS” as the ground truth, as they measure three different objects. However, the model generated failed to identify them as separate indicators, although we injected similar examples in the prompt. This issue aligns with the viewpoint presented in [17] that large models usually rely on “mechanical memorization” to handle long-tailed or isolated samples rather than actually learning underlying patterns.

Incorrect property value. This mostly occurs when generating values for the “hasMeasurementPhenomenon” relationship. For example, the model generated “Gross energy-indirect (Scope 2) GHG emissions in metric tons of CO2 equivalent, has measurement phenomenon, GHG emissions”, while the correct property value should be “energy-indirect GHG emissions”. Another example is “Total weight of hazardous waste directed to disposal in metric tons, has measurement phenomenon, hazardous waste disposal”, while the correct property value should be ‘hazardous waste directed to disposal”. In both examples, the LLM failed to identify the sustainability domain-specific concepts as correct property value.

5. Future Works

As future work, we aim to validate our framework through extended experiments. We will compare the results obtained from the current setup with the results using pure keywords and using text chunks of the Indicator A as input in Experiment 1, to assess the performance of using subgraphs as input for the retrieval step. For Experiment 2, we plan to extend our evaluation to include additional entity and relation types. For both experiments, we aim to evaluate the performance of different language models, as well. Additionally, we will conduct further experiments for the mapping and construction steps. Meanwhile, we intend to evaluate whether fine-tuning improves the model performance on the retrieval and KGC tasks. Furthermore, we will explore practical applications of the constructed KG, particularly focusing on how to utilize the KG constructed by LLMs to mitigate the hallucination problem within the RAG framework to empower the digitalization and automation of the corporate sustainability reporting process.

References

[1] KPMG, Global Survey of Sustainability Reporting 2022, https://kpmg.com/sg/en/home/insights/2022/10/global-survey-of-sustainability-reporting-2022.html, 2022. [Online; accessed: 24th July 2024].

[2] R. Siew, A review of corporate sustainability reporting tools (srts), Journal of Environmental Management 164 (2015) 180–195. doi:10.1016/j.jenvman.2015.09.010, epub 2015 Sep 15; PMID: 26379255.

[3] GRI and SASB, A Practical Guide to Sustainability Reporting Using GRI and SASB Standards, https://www.globalreporting.org/media/mlkjpn1i/gri-sasb-joint-publication-april-2021.pdf, 2021. [Online; accessed: 13th March 2024].

[4] L. Ghahremanloo, J. A. Thom, L. Magee, An ontology derived from heterogeneous sustainability indicator set documents, in: Proceedings of the Seventeenth Australasian Document Computing Symposium, ACM, Dunedin, New Zealand, 2012, pp. 72–79. doi:10.1145/2407085.2407095.

[5] C. Diamantini, T. Khan, D. Potena, E. Storti, Shared metrics of sustainability: a knowledge graph approach, in: Proceedings of SEBD, 2022, pp. 244–255.

[6] Y. Zhou, Y. Cao, A. Perzylo, Towards digital sustainability reporting: An ontology for mapping of indicators in gri and esrs, in: Proceedings of the 20th International Conference on Semantic Systems (SEMANTICS), Accepted for presentation, 2024.

[7] H. Chen, Large knowledge model: Perspectives and challenges, arXiv preprint arXiv:2312.02706 (2023).

[8] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering (2024).

[9] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, N. Zhang, Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities, arXiv preprint arXiv:2305.13168 (2023).

[10] M. Monajatipoor, J. Yang, J. Stremmel, M. Emami, F. Mohaghegh, M. Rouhsedaghat, K.-W. Chang, Llms in biomedicine: A study on clinical named entity recognition, arXiv preprint arXiv:2404.07376 (2024).

[11] V. K. Kommineni, B. König-Ries, S. Samuel, From human experts to machines: An llm supported approach to ontology and knowledge graph construction, arXiv preprint arXiv:2403.08345 (2024).

[12] A. M. Brasoveanu, L. Nixon, A. Weichselbraun, A. Scharl, Framing few-shot knowledge graph completion with large language models (2023).

[13] H. Li, Y. Su, D. Cai, Y. Wang, L. Liu, A survey on retrieval-augmented text generation, arXiv preprint arXiv:2202.01110 (2022).

[14] Y. Hu, Y. Lu, Rag and rau: A survey on retrieval-augmented language model in natural language processing, arXiv preprint arXiv:2404.19543 (2024).

[15] A. Gupta, A. Shirgaonkar, A. d. L. Balaguer, B. Silva, D. Holstein, D. Li, J. Marsman, L. O. Nunes, M. Rouzbahman, M. Sharp, et al., Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture, arXiv preprint arXiv:2401.08406 (2024).

[16] S. Setty, K. Jijo, E. Chung, N. Vidra, Improving retrieval for rag based question answering

models on financial documents, arXiv preprint arXiv:2404.07221 (2024).
[17] X. Chen, L. Li, N. Zhang, X. Liang, S. Deng, C. Tan, F. Huang, L. Si, H. Chen, Decoupling knowledge from memorization: Retrieval-augmented prompt learning, Advances in Neural Information Processing Systems 35 (2022) 23908–23922.