Semantic Enrichment of the Quantum Cascade Laser Properties in Text- A Knowledge Graph Generation Approach

Deperias Kerre[1,3,*], Anne Laurent[1], Kenneth Maussang[2] and Dickson Owuor[3]

[1]LIRMM, Univ Montpellier, CNRS, Montpellier, France
[2]IES, Univ Montpellier, CNRS, Montpellier, France
[3]SCES, Strathmore University, Nairobi, Kenya

Abstract

A well structured collection of the various Quantum Cascade Laser (QCL) design and working properties data provides a platform to analyze and understand the relationships between these properties. This analysis can result in insights into how different design features impact laser performance properties. Most of these QCL properties are captured in scientific text. This poses challenges in generating structured QCL properties data for exploration properties due to the specific nature of this domain. There is therefore a need for efficient methodologies that can be utilized to extract QCL properties from text and generate a semantically enriched and interlinked platform where the properties can be analyzed. There is also the need to maintain provenance and reference information on which these properties are based. Semantic Web technologies such as Ontologies and Knowledge Graphs (KGs) have proven capability in providing interlinked data platforms for knowledge representation in various domains. In this paper, we propose an approach for generating a QCL properties Knowledge Graph (KG) from text. The approach is based on the QCL ontology and a Retrieval Augmented Generation (RAG) enabled information extraction pipeline based on GPT 4-Turbo language model. The properties of interest include: working temperature, laser design type, lasing frequency, laser optical power and the heterostructure. The experimental results demonstrate the feasibility and effectiveness of this approach for efficiently extracting QCL properties from unstructured text and generating a QCL properties Knowledge Graph, which has potential applications in semantic enrichment and analysis of QCL data.

Keywords

Information Extraction, Knowledge Graphs, Linked Data, Ontologies, Retrieval Augmented Generation, Quantum Cascade Lasers, Semantic Web

1. Introduction

Quantum Cascade Lasers (QCL) are semiconductor laser devices which consist of a nanometric stack of different semiconductor materials and whose spectral emission is restricted to the frequency range from about 100 GHz to 10 THz. The stacks of different materials are referred to as heterostructures. Interactions between the various layers of materials result in various emission behaviours of the QCL laser device [1]. The nature of the radiations emitted by these laser devices have enabled various applications ranging from analysis of chemicals, high-resolution spectroscopy in astronomy, detection of organic compounds in drugs etc [2, 3, 4, 5, 6, 7].

QCL properties can be broadly classified into two categories: Design properties and the Optoelectronic/ Working properties. The design features consists of the laser design characteristics such as the laser design types, the material combinations used and the layer sequencing. The Optoelectronic properties on the other hand refers to the performance behaviour of a QCL device with particular design characteristics. Examples of working properties include Power, Temperature, Frequency etc.

LLM-TEXT2KG 2025: 4th International Workshop on LLM-Integrated Knowledge Graph Generation from Text (Text2KG) Co-located with the Extended Semantic Web Conference (ESWC 2025), June 1 - June 5, 2025, Portoroz, Slovenia.
*Corresponding author.
dkerre@strathmore.edu;deperias.kerre@lirmm.fr (D. Kerre); anne.laurent@umontpellier.fr (A. Laurent); Kenneth.Maussang@umontpellier.fr (K. Maussang); dowuor@strathmore.edu (D. Owuor)
0000–0002–7437–6735 (D. Kerre); 0000–0003–3708–6429 (A. Laurent); 0000–0002–8086–8461 (K. Maussang); 0000–0002–0968–5742 (D. Owuor)

The working properties of a QCL device are dependent on the design features. This implies that there is a relationship between the laser design and working properties. Understanding the relationships between the QCL design and working properties plays an important role in the fabrication of a QCL semiconductor device with target properties, for instance, the Working Temperature.

The QCL properties are described in text where several device designs and their corresponding working properties are proposed. Some of the laser properties mentioned in the text are based on references in other text. Some of the QCL properties such as temperature may have several values in text hence its important to extract the correct value of interest. A sample text description of QCL properties is given: *"GaAs/AlGaAs quantum cascade lasers based on four quantum well structures operating at 4.7 THz are reported. A large current density dynamic range is observed, leading to a maximum operation temperature of 150 K for the double metal waveguide device and a high peak output power more than 200 mW for the single surface plasmon waveguide device"* [8]. This text description contains several QCL properties such as lasing frequency (4.7 THz), power (200 mW), working temperature (200 K) and the heterostructure materials (GaAs/AlGaAs). The data on QCL properties therefore exist in heterogeneous text sources in an unstructured format and therefore require a lot of effort to collect data, structure, and explore the relationships between the various QCL properties.

Methods for generation of KGs from text that involve extraction of triples from text have been proposed. These methods are not suitable for the QCL domain, as the relations between entities are not explicitly mentioned in the text but deduced from expert knowledge. Generation of KGs for the domain can therefore be achieved by population of structured QCL data onto an expert-defined KG schema. This requires generation of structured QCL data from text and a foundational ontology to provide the KG schema semantics. There have been attempts to extract QCL properties from text using rule-based approaches [9], indicating a possibility to generate QCL data from text. A formal representation of the QCL properties in form of an ontology model has also been proposed [10].

The existing Knowledge KGs in the materials science domain don't capture properties for the QCL domain and cannot therefore be readily utilized to answer queries on QCL design features and the corresponding performance characteristics. There is therefore the need for a semantically enriched platform that captures the QCL properties in text, their provenance information together with links to the references for the properties. This will enable exploration of the relationships between the various properties in the form of queries. The insights derived from these information can be used in the fabrication of laser devices with target properties. This will also provide an interlinked platform where both machines and humans can explore and query the QCL properties data in a FAIR (Findable, Accessible, Interoperable, and Reusable) manner [11].

In this paper, we present an approach for generating a QCL properties KG from text. The main contributions of this paper are therefore as follows:

- We propose a Retrieval Augmented Generation (RAG) based approach for QCL property extraction from scientific text to generate structured QCL properties data.
- We present an experimental analysis of the RAG-based approach on various Large Language Models.
- Based on the QCL Ontology and other vocabularies in the materials science domain, we implement a mapping process to generate a KG for the QCL properties.
- We evaluate the ability of the generated Knowledge Graph in capturing domain knowledge using sample test cases.

The remaining sections of this paper is organized as follows: we give an overview of related works in information extraction and knowledge representation in the materials science domain in section 2, the Knowledge Graph generation workflow in section 3, the experimental results in section 4 and lastly conclude in section 5.

2. Related Work

2.1. Information Extraction in Materials Science Domain

Extraction of materials properties from text has gained a lot of interest recently as occasioned by the need for accelerated materials discovery. The methodologies developed can be broadly classified in to rule-based methods and machine learning based methods. Rule-based methods constitutes rules, grammars and other expert-defined structures for identification of properties in specific domains. The machine learning based approaches entails training of learning algorithms on labelled data to enable them to learn how to identify specific materials properties in text.

Examples of rule based toolkits adopted for properties extraction in materials science include chemDataExtractor [12], LeadMine [13], ChemicalTagger[14], tmChem [15] and ChemSpot [16]. The chemDataExtractor toolkit has also been widely adopted for materials properties extraction in other specific use cases which includes: thermo-electric materials [17], semiconductor bandgaps [18], refractive indices and dielectric constants [19], an auto-populated ontology of materials science [20], battery materials [21], transition temperatures of magnetic materials [22] and quantum cascade laser properties [9].

Machine learning methods have also been adopted in the extraction of materials science properties. Examples include generation of datasets of gold nano-particle synthesis procedures, morphologies and size entities [23], and materials synthesis recipes [24]. Another work is on the use of the combination of deep convolutional and recurrent neural networks for named entity recognition [25]. BERT (Bidirectional Encoder Representations from Transformers) models have also been proposed for the analysis of optical materials [26] and extraction of battery materials from scientific text [27].

The emergence of generative large language models have also opened opportunities in the extraction of materials properties from text. The models have been harnessed for text parsing in solid-state synthesis internary chalcogenides [28] .The in-context learning method is also applied to assess the ability of LLMs in processing materials data [29] and extracting materials data from research papers [30]. Lastly, LLMs have also been utilised in the construction of functional materials Knowledge Graph in multidisciplinary materials science [31].

Despite the great achievements in the adoption of Information Extraction methods for extracting materials science data from scientific text, there still exist open challenges that need to be addressed in order to develop efficient methodologies in generating structured data for QCL properties from scientific text. The rule-based approaches for materials properties extraction from text are domain specific and are limited in cases where there is slight change in the text structure. The machine learning models need retraining in order to be utilized for QCL properties extraction from text. The LLM based methodologies present a promising direction in the extraction of materials science data from text. The models however, require quality training data and massive computing resources in order to be fine-tuned for specific domain properties extraction from text. For the in-context learning approaches for LLMs, there is need to efficiently generate the best quality training examples to be used for prompting during the materials properties data extraction.

There is therefore, the need for efficient methodologies for extracting QCL properties from text to generate structured QCL properties data. This will provide a foundation for the generation of knowledge representation platforms for analyzing the relationship between the various QCL properties captured in various heterogeneous textual data sources.

2.2. Knowledge Graphs in Materials Science Domain

Knowledge Graphs have been proposed in representing knowledge for properties in the materials science domain. The KGs are based on several foundational and specific domain ontologies developed for this domain. The motivation behind these KGs is to provide an accelerated analysis of materials science properties for various reasons, including materials discovery.

Nanomine, a KG for nanocomposite materials is proposed in [32]. This KG provides a unified platform for generating visualizations and analyzing the relationship in nanocomposite materials. The

KG generation pipeline for Nanomine involves manual extraction of data from papers, structured it into files such as excel and uploading it to the ontology schema. Propnet is a KG proposed for a wide range of materials science properties [33]. The KG provides a computational framework that helps scientists to automatically calculate additional information from their datasets such as the Materials Project database.

Other KGs are also proposed for the general materials science data [34, 35, 36]. The generation pipelines entail use of advanced NLP techniques such as Named Entity Recognition and Relationship Extraction [34, 35] and deep-learning approaches [36]. Lastly, a materials terminology KG has been proposed in [37] and the materials experiment KG proposed in [38]. The materials terminology KG comprises of 8660 materials terms and their explanations automatically generated from text corpus via NLP techniques. The materials experiment KG captures provenance information of each material sample together with associated data and metadata.

Despite adoption of KGs in the representation and exploration of materials science data, the existing KGs cannot be readily used to represent the knowledge in QCL properties data. Nanomine KG captures knowledge on domain-specific nanocompoiste particles hence not suitable for the QCL domain. Propnet and the other general materials science KGs captures wide materials science concepts which do not capture knowledge on the QCL heterostructure design properties. The relationships to other working properties and working modes are also not captured. The materials terminology and materials experiment KGs do not capture the relationships suitable for the QCL properties. None of the KGs can therefore be adopted as a unified platform for representing and analyzing QCL properties data together with its provenance information.

The KG generation techniques from text need to improved in order to be adopted to the task of generating the QCL properties KG from text. For instance, for the QCL domain, the relationships between properties are not directly captured in text and can only be inferred from expert knowledge. This implies that relation extraction techniques are limited in QCL KG generation from text.

There is therefore need for efficient methodologies for extracting QCL properties from text to generate structured data and utilize this data to generate a Knowledge Graph for representing QCL properties, relationships among them and the provenance information. To the best of our knowledge, this work presents the initial steps for implementing the task of QCL properties extraction from text and KG generation for property exploration.

3. Methodology

The Knowledge Graph generation approach is composed of the following parts: Information Extraction pipeline and the KG modeling and data enrichment part.

3.1. Information Extraction Pipeline

In this module, we explore the use of large language models in developing the pipeline. Large language models such as GPT models are trained on general knowledge data and cannot be efficiently used on specific domain tasks without adaptation. Fine-tuning and in-context learning have been proposed as methods to adapt LLMs to domain specific tasks. Fine-tuning requires a lot of quality training data and requires a large amount of computational resources. For in-context learning, the quality of the output is dependent on the quality of the prompt and the examples it contains. In most cases especially for few shot learning, the examples have to be manually added in the prompt.

In this paper, we hypothesize that exposing the model to a labeled instruction data consisting of sample text describing QCL properties and the corresponding extracted properties improves the model's performance on this task of property extraction from text. This also aligns the model's output to the expected format and minimizes irrelevant responses. This also eliminates the need for fine-tuning and static prompt generation.

We propose a hybrid few shot learning strategy where the few shot examples consist of the best examples automatically generated from the instruction dataset by a RAG pipeline as opposed to normal

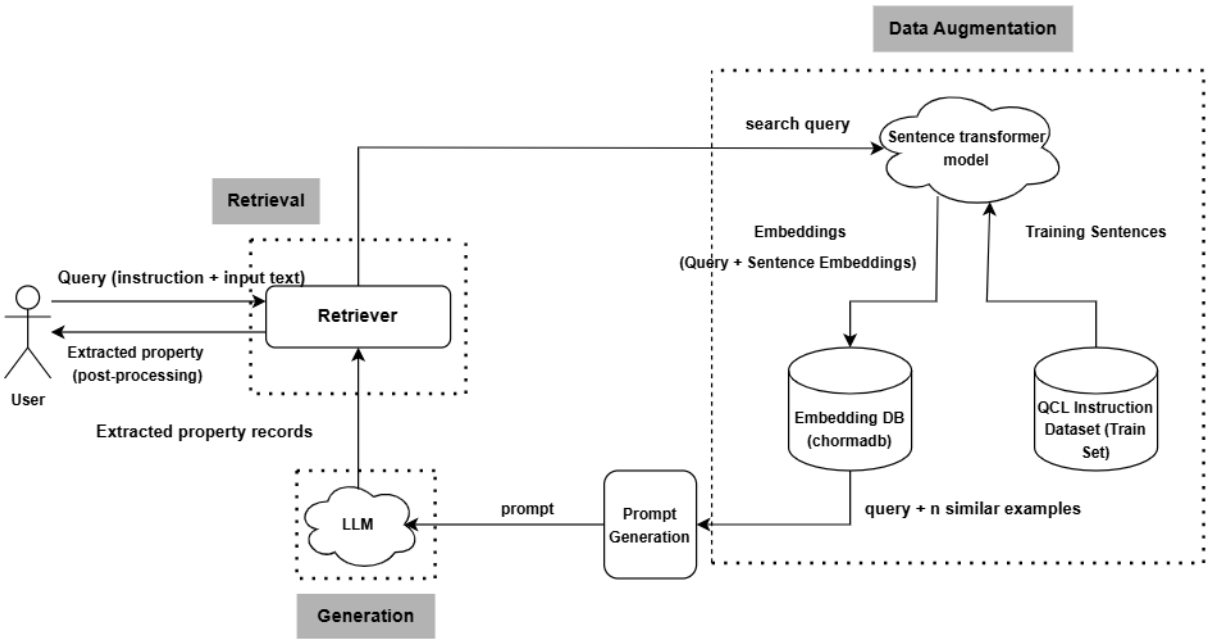

Figure 1: The Advanced RAG Pipeline

few shot learning where static examples have to be specified in every prompt. We adopt GPT-4 Turbo model in our approach. This is owed to GPT-4 Turbo improved efficiency in generating responses and the larger context window [39]. Our method is based on the Naive RAG approach [40]. As illustrated in Figure 1, the module has three sections i.e Retrieval, Data Augmentation and Data Generation. The rest of this subsection describes the pipeline modules.

3.1.1. Retrieval

An input sentence (query) containing QCL property of interest and an instruction to the large language model is submitted by the user to the retriever. This is then forwarded by the retriever to the data augmentation module for retrieval of similar responses based on the query.

3.1.2. Data Augmentation

This module consists of train data embeddings stored in an embedding database (DB). This provides the context to the model during information extraction. In our case, the context is provided by an original QCL properties instruction dataset [41]. The dataset description is given elsewhere[42]. We sample 80 % of this dataset for training and 10 % for testing. The dataset comprises of 1040 sample sentences containing QCL properties, an instruction to the model for information extraction together with the corresponding properties extracted.

Sentence embeddings are computed using a based pre-trained sentence transformer model[43]. We adopt the all-mpnet-base-v2[1] version of the sentence transformer model in our approach. The query embeddings are also computed by the sentence transformer model in order to allow for comparison with the embeddings in the training data.

Similarity scores between the query embeddings and the train sentence embeddings in the embedding DB are computed using the cosine similarity metric. The examples in the embedding DB that are more similar are retrieved. The examples and the user query are both passed to a prompt generator which prepares a prompt based on a defined prompt template. We define a prompt template that allows parsing a user query together with the relevant examples showing sample instructions with text containing

[1]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Problem Definition: Extraction of quantum cascade laser properties from text entails extracting properties from a given text description. This should be done without providing any other additional information or explanations. The output format should correspond to the one in the example sentences. Example sentences containing an instruction, input text and the extracted properties are given below:

Example Sentences: {Example Sentences Containing QCL properties, LLM Instruction and the Corresponding Extracted Properties}

Instruction: {An instruction to the LLM for extracting QCL property and the input text containing the properties. }

Figure 2: The Prompt Template

properties and the extracted properties. The generated prompt is then passed to the Generation phase. Figure 2 shows the prompt template and Figure 6 shows a sample regenerated prompt.

3.1.3. Generation

In the generation module, a response is generated by the language model in a zero-shot manner. This stage is flexible as any language model can be used for response generation. The model responses are then processed to remove any incomplete records or any irrelevant responses.

Formally, the general process of extraction of QCL properties from text based on the detailed RAG approach is carried out as follows: we frame the QCL property extraction from text task in the form of a conversation as follows: Given a set of input sentences S, where $S = \{s_1, s_2, s_3 ... s_n\}$, we design a prompt P that contains an instruction I, the Sentences S and contextual examples $C = \{C_1, C_2, C_3 ..., C_k\}$, where K depends on the number of examples desired for the context. This prompt is passed to the model in order to extract a particular QCL property record R, where $R = \{p_1, p_2, p_3 ..., p_n\}$ and n implies the number of properties in the record. It is worthwhile to note that each P is passed to M for each property record. Algorithm 1 shows the steps followed to extract the data.

Algorithm 1: The QCL Properties Extraction Steps

Input: Set of input sentences S, Prompt P, Instruction I, base model M, an embedded instruction dataset to provide the context C.

Output: QCL property record R.

1 Set the number of examples to be retrieved (k) to a specific value. `/* This will determine the number of examples to be retrieved from the context documents based on the user query. */`
2 Enter an instruction I and an input sentence S.
3 Set the query Q= Instruction (I) + Input Sentences (S).
4 Convert the queries to a vector embedding Eq.
5 Pass query Eq to the retriever.
6 Fetch k examples from C based on Eq.
7 Set Eq = Eq + k examples.
8 Set Q= Eq (decoded to plain text).
9 Update the prompt P with Q.
10 Pass P to the base model M.
11 **return** R
12 **Repeat** steps 2-9 until all the property records of interest are extracted.
13 Post-process R. `/* Entails removal of incomplete records and any unnecessary responses. */`

For proof of concept, we use the pipeline to generate sample structured QCL properties data for 36

QCL devices from 36 abstracts documenting QCL devices and their properties [8, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78]. The data is post-processed to have a clean file of all the properties extracted. We also include the metadata (DOI and URL) for provenance information and references for referencing the various mentioned properties during the post-processing phase. The missing values are also included in the post-processing stage. The final data constitutes a well-structured csv file containing QCL properties data for every device together with the associated provenance information and the references.

3.2. Knowledge Graph Modeling and Data Enrichment

In this section, there are two processes that we carry out: first we define the KG model to organize the data and secondly we map the data to enrich it. We detail them in the following subsections:

3.2.1. Ontologies and Knowledge Graph Modeling

A Knowledge Graph represents a semantic network of interlinked entities. Entities refer to real-world concepts or ideas that can be identified by a unique identifier on the web. A Knowledge Graph is defined in the form of triples, that consist of two entities and a relation (predicate) linking them. Formally, a Knowledge graph KG can be defined as $KG = \{t_1, t_2, t_3..., t_n\}$ and t refers to the various triples in the KG and n the number of triples in the KG. A triple $t = \{s, p, o\}$ where s is the subject, p the predicate,, and o the object. s, p and o are denoted by Resource Identifiers in the form of Uniform Resource Identifiers (URIs) or Literals (e.g., strings, numbers, dates). The semantics of a KG are provided by ontologies or standard vocabularies.

In our case, the entities are the various QCL properties, the various relationships among them, and the provenance information for these properties. In order to define a KG schema to represent the knowledge for QCL properties, we adopt concepts from several ontologies that capture the terms of interest for the QCL properties and the relationships between them. The semantics for the QCL properties and the relationships between them are provided by the QCL ontology[2] [10]. This is an ontology for properties in the QCL domain hence suitable for capturing knowledge on QCL terms.

The other vocabularies adopted in the schema include concepts from BIBO[3] and schema.org[4]. Formal definitions for working properties are adopted from the Materials Design Ontology (MDO) [79]. The provenance ontology is used to model the provenance information of the QCL properties. Figure 3 shows the KG schema used to organize the data (with the instances in shaded boxes) and Table 1 shows the prefixes and URIs for the namespaces used in the KG schema.

The KG schema covers the following categories of information that we enrich: laser heterostructure (heterostructure materials), laser working properties (power, temperature and the lasing frequency), laser design types, provenance information and citation tracks. The heterostructure captures the materials stacking properties of the laser and the related design. It is defined by the concept QpOnto:Laser-Heterostructure. A heterostructure contains heterostructure materials (QpOnto:HeterostructureMaterials) and the material combination has a formula (QpOnto:matFormula) which indicates the materials used and the ratio of combination in a string. A heterostructure also has a design type. Examples of design types includes the resonant phonon and the LO phonon design types.

The optoelectronic properties captures the QCL performance behaviour as a result of injection of current. These includes the working temperature, power and frequency. These properties are related to particular design information i.e design types and heterostructure materials. The working temperature also depends on the laser working mode i.e whether the emission is in continuous or pulse mode.

[2] https://github.com/DeperiasKerre/qcl_Onto/blob/main/qclontology/version-1.0/qclonto.owl
[3] https://dcmi.github.io/bibo/
[4] https://schema.org/

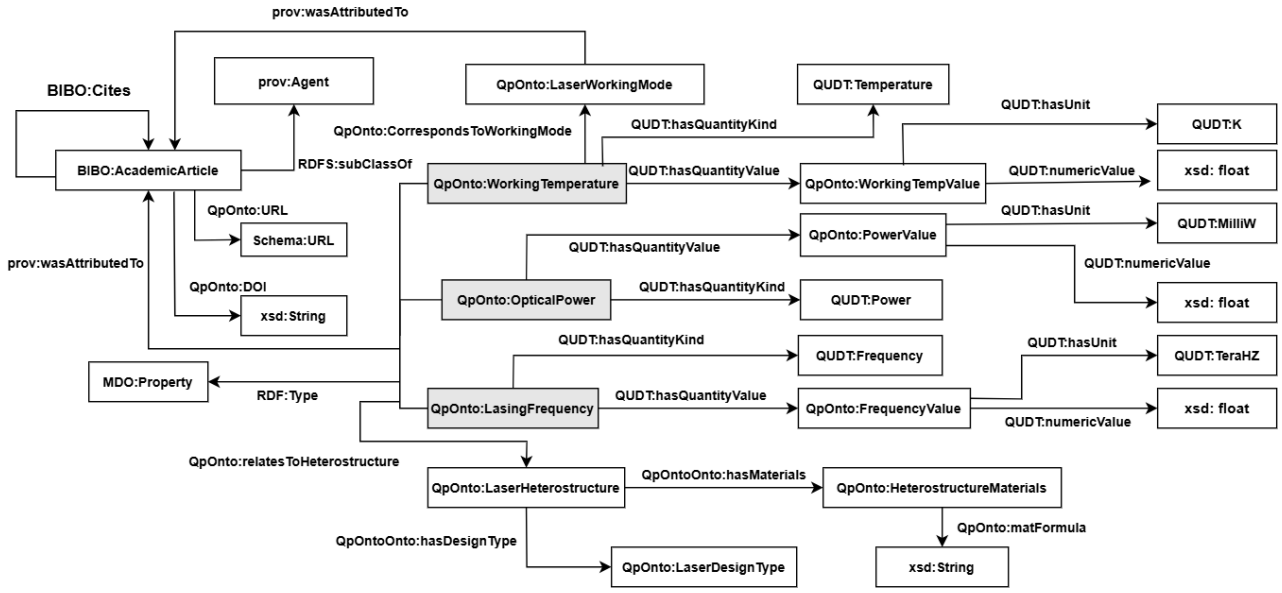

Figure 3: The Knowledge Graph Schema

Table 1
Namespace URIs

Prefix	Namespace URI
QpOnto	https://github.com/DeperiasKerre/qcl_Onto/blob/main/qclontology/version-1.0/qclonto.owl#
MDO	https://w3id.org/mdo/core/
BIBO	https://dcmi.github.io/bibo/#:
prov	http://www.w3.org/ns/prov#
RDFS	http://www.w3.org/2000/01/rdf-schema#
QUDT_Properties	https://qudt.org/schema/qudt/
QUDT_Units	https://qudt.org/vocab/unit/
QUDT_Quantity Kinds	https://qudt.org/vocab/quantitykind/

The QCL properties provenance information is also captured in the schema. This is implemented by capturing the metadata (DOI and URL) of the articles documenting the various laser properties. We also provide links to references in those articles as some of the properties proposed in a given design are based on another design in a referenced article. This is done using the concept BIBO:cite from the BIBO ontology. The semantics of the units, quantity kinds, numerical values and their associated relationships for the QCL working properties are modeled by re-using the terms in the QUDT ontology.

3.2.2. Data Enrichment

In this phase, several steps are carried out in order to populate the KG schema with the structured QCL properties data generated in section 3.1 in order to generate the KG. The mapping process entail mapping the data literal values to the respective entities via data properties and linking them through the object properties as defined by the KG schema. We also specify the data types for all the data instances. This is implemented using the rdflib library [5]. The units, design types and the relevant working modes descriptions are also enriched with the relevant URIs. The redundant triples are also examined and eliminated from the generated Knowledge Graph.

The final RDF file is then serialized into the Turtle and the RDF/XML formats in order to generate

[5]https://github.com/RDFLib/rdflib

the Knowledge Graph for querying and exploration. The generated Knowledge Graph contains a total of 2979 triples containing the QCL properties and their associated provenance information. A visualization of a sample instance of a QCL heterostructure (capturing the design type and material combination information) and its provenance information is shown in Figure 4 with the data values and the provenance information adopted from [70] .

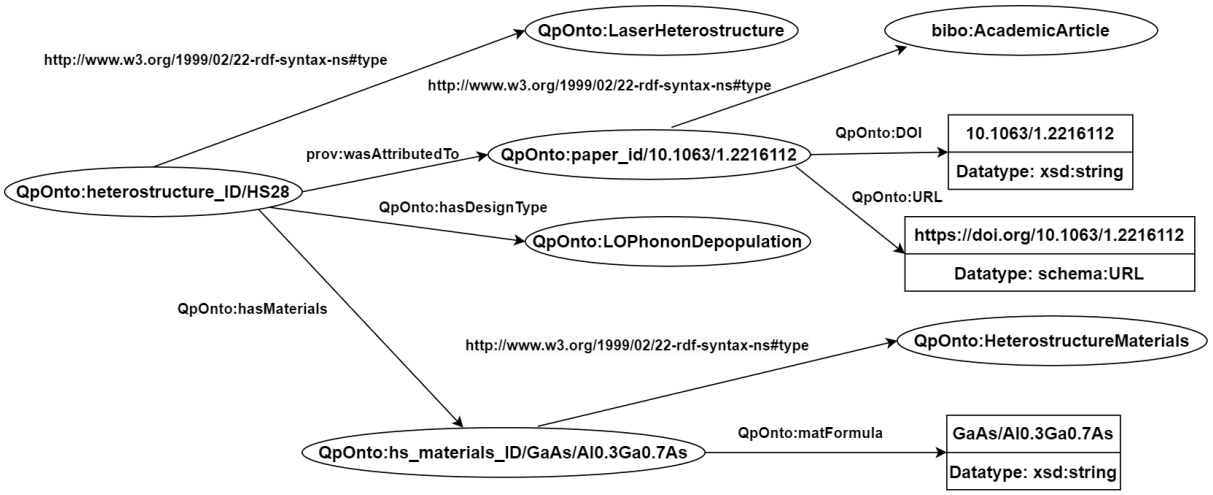

Figure 4: Sample Instance of the QCL Heterostructure Visualization from the Generated Knowledge Graph [70].

4. Experiments and Results

In this section, we carry out experiments to evaluate the KG generation approach. This is done is based on three strategies: the performance of the approach in QCL property extraction from text, the consistency and the correctness of the generated KG in terms conformance to the QCL properties domain requirements i.e the ability to capture the intended knowledge correctly. We also provide an analysis of the completeness of the proposed KG.

4.1. Property Extraction From Text

4.1.1. Evaluation Metrics

For the evaluation of the information extraction module, we adopt the expert validation approach. This entails comparison of the model's output with the expert annotated ground truth label in the evaluation dataset. We use the precision and recall in order to evaluate the performance of the approach on QCL property extraction from text[80]. Precision is the fraction of correct (relevant) records among all extracted records and the recall is the fraction of successfully extracted records among all correct (relevant) records in the dataset. The metrics are determined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

TP refers to the true positive count (the number of correct records extracted), FP is the false positive count (the number incorrect records extracted), and FN corresponds to the false negative count (the number of correct records that are not extracted). We define the terms correct and incorrect in our context as follows:

Definition 1: The word "correct" in this context implies that the property value extracted

can be validated by a human expert when reading the corresponding sentence containing the property. It should also match with the ground truth label value in the evaluation dataset. Property values with units are only considered correct if they are extracted together with the units.

Definition 2: An "incorrect" (false) record suggests that the value extracted does not correspond to the correct/expected value as compared to the ground truth value in the evaluation dataset.

4.1.2. Evaluation Dataset and Baselines

We evaluate the information extraction pipeline on a test dataset obtained from the instruction dataset described in section 3.1. This contains a total of 130 sentences containing the different QCL properties. Table 2 gives the summary statistics of the test data per QCL property.

Table 2
Statistics of the test data.

Quantum Cascade Laser Property	Number of Sentences
Laser Optical Power	26
Working Temperature	34
Laser Design Type	27
Laser Frequency	29
Laser Heterostructure	14
Total	**130**

For the baseline, we compare the hybrid few-shot learning approach (RAG-enhanced approach) with direct prompting where no contextual information is provided. We evaluate our approach on GPT 4-Turbo and the Mistral-7b-Instruct model. We don't perform evaluations on GPT-4 due to the context window limitations. We analyze the performance per QCL property as different properties have varying levels of difficulty during the extraction process.

4.1.3. Evaluation Results

We present the results for QCL property extraction from text analyzed per QCL property as the properties have a varying level of difficulty during the extraction process. The results are as shown in Table 3 for the laser power, Table 4 for working temperature, Table 5 for laser heterostructure materials, Table 6 for lasing frequency, Table 7 for the laser design type and Table 8 for the average performance of the models for all the properties.

Table 3
Performance Metrics for Laser Output Power.

Model	Precision	Recall
gpt 4-turbo + RAG	100 %	100 %
gpt 4-turbo (direct prompting)	100 %	100 %
mistral-7b-instructv0.3 + RAG	100 %	100 %
mistral-7b-instructv0.3 (direct prompting)	100 %	96.15 %

Table 4
Performance Metrics for the Laser Working Temperature.

Model	Precision	Recall
gpt 4-turbo + RAG	100 %	100%
gpt 4-turbo (direct prompting)	90.91 %	96.77 %
mistral-7b-instructv0.3 + RAG	85.30%	100 %
mistral-7b-instructv0.3 (direct prompting)	85.29 %	100 %

Table 5
Performance Metrics for the Laser Heterostructure.

Model	Precision	Recall
gpt 4-turbo + RAG	85.71 %	100 %
gpt 4-turbo (direct prompting)	91.67 %	84.62 %
mistral-7b-instructv0.3 + RAG	85.71%	85.71 %
mistral-7b-instructv0.3 (direct prompting)	92.86 %	100 %

Table 6
Performance Metrics for the Laser Frequency.

Model	Precision	Recall
gpt 4-turbo + RAG	100%	100%
gpt 4-turbo (direct prompting)	92.86 %	96.30 %
mistral-7b-instructv0.3 + RAG	96.43 %	96.43 %
mistral-7b-instructv0.3 (direct prompting)	92.59 %	92.59 %

Table 7
Performance Metrics for the Laser Design Type.

Model	Precision	Recall
gpt 4-turbo + RAG	100 %	33.33 %
gpt 4-turbo (direct prompting)	92.31 %	46.15 %
mistral-7b-instructv0.3 + RAG	100%	92.60 %
mistral-7b-instructv0.3 (direct prompting)	33.33 %	29.41 %

Table 8
Average Performance of the Models for all the Properties

Model	Precision	Recall
gpt 4-turbo + RAG	**97.14 %**	**86.67 %**
gpt 4-turbo (direct prompting)	93.55 %	83.59 %
mistral-7b-instructv0.3 + RAG	93.49 %	92.01 %
mistral-7b-instructv0.3 (direct prompting)	80.82 %	80.69 %

The laser power, working temperature and frequency exhibit higher precision in the extraction process in both direct prompting and the proposed approach. This is attributed to the fact that these properties are generally available in the general knowledge of the models. However, there is lower recall for the simple queries for these properties. This is in cases where more than one value of these properties are mentioned and the models struggle to identify the required value hence failing to extract these properties. This is also the case when the values are given in ranges, for instance, the lasing frequency. Provision of examples capturing such scenarios improves the model performance on these properties. This is indicated by the best performance exhibited by the GPT 4-Turbo RAG based approach as shown in Tables 3, 4 and 6. It is also noted that there is a significant improvement in performance for both the precision and recall for the lasing frequency and the working temperature with the RAG based approach(Tables 4 and 6). This is due to the ability of the model to learn how to identify the properties of interest based on the provided examples.

The laser design type property entails a QCL domain specific property. The models exhibits higher precision in cases where the keyword “design type” is explicitly used in the text description but fails totally to recognize and extract the property in cases where the key word is not used. For both models, there is an improvement in generalization (in terms of precision) as the models utilize the labeled examples to recognize this domain specific terms. There is however a decrease in recall for the GPT-Turbo model. This is attributed to cases where the model does not completely extract terms with more variations form the context documents. In this case, the Mistral-instruct-7b model exhibits the best performance with the proposed RAG approach as shown in Table 7.

The QCL heterostructure materials property is also a QCL domain specific property. With this property, there is no significant improvement with the proposed approach for all the models. Despite the heterostructure being a domain specific concept, its description is characterized by the terms“‘structure”, “heterostructure” or “materials” which enables the models to identify the properties at relatively higher

precision even with direct prompting. The best performance is exhibited by the mistral 7b-instruct model in a direct prompting format (Table 5).

In summary, exposure of large language models to quality labeled data improves their ability in recognizing and extracting the relevant QCL properties. Averagely, an improvement of performance is achieved as follows: precision +3.59 %, recall +3.08 % for GPT-4 Turbo and precision +12.67 %, recall +11.32 % for Mistral-7b-instruct model. The proposed approach enables the update of the models context without full fine-tuning that is computationally expensive. With the proposed approach, it's even possible to train the model to adopt a certain output format for the extracted properties to avoid any unnecessary responses or undesired output formats. This approach can be extended for the other QCL properties as it enables the model to learn how to identify the domain specific properties with the provided examples with less resources. With this approach, the model's performance is however dependent on data quality and the properties covered. The model's performance therefore increases with more diverse examples in the model's context.

The performance of the model (in terms of the recall) is still low for some properties such as the design type due to the domain specific nature of this property and the varying styles in which it is expressed in text as compared to other properties. More training strategies can be explored to improve the model's performance on such properties. This can vary from training methods to developing diverse datasets for the task of extracting such properties.

4.2. Knowledge Graph Evaluation

The generated Knowledge Graph is evaluated based on two metrics: the consistency in the KG triples, the correctness of the KG in capturing the intended knowledge in the QCL domain properties and the completeness of the KG. The consistency ensures the logical soundness of the defined triples in the generated KG. The correctness of the KG in terms of the domain requirements aims to evaluate the generated KG‘s ability in capturing the domain knowledge of interest and being able to provide answers to questions regarding the various QCL properties.

4.2.1. Logical Consistency

The Knowledge Graph consistency is validated by lack of inconsistencies/contradictions in the generated triples. This is implemented using logical reasoners. We the validate the KG consistency using the pellet reasoner [81].

4.2.2. Knowledge Graph Correctness

Knowledge Graph correctness refers to the accuracy/relevance of the facts stored in the KG. Under this we evaluation metric, we adopt the expert validation approach to assess the ability of the proposed KG in capturing the domain requirements correctly. With the help of QCL domain experts, we define a set of test cases in form of competency questions (CQs). The CQs comprise of set of inquiries that can be utilized by experts to query and explore the various QCL properties and the relationship between them. The responses obtained from the queries can be used for insights regarding the design of optimal QCL devices with target properties. This enables a quicker exploration of QCL properties from heterogeneous data sources.

The competency questions capture the whole QCL properties of interest i.e the design, working properties, the laser working modes and the provenance information. Table 9 shows the classes of the CQs. The X,Y and Z are place holders for any particular properties of interest to be used in the queries. For every class of queries in Table 9, we design and run several possibilities of the queries. We compare the queries output with the expected output in order to determine the precision in question answering by the Knowledge Graph.

We run a total of 20 queries in order to validate the suitability of the generated KG in capturing the QCL properties, the relationships between them and their provenance information. Table 10 shows

Table 9
List of Competence Questions

Question ID	Question Text
CQ1	What are the possible material compositions of a QCL laser heterostructure with a design type X ?
CQ2	What is the working property X of a QCL laser working in mode Y ?
CQ3	What is the performance property X of a QCL laser having a heterostructure with material composition Y?
CQ4	For a particular performance property X, what are the corresponding laser heterostructure designs?
CQ5	For a particular performance property X, what are the corresponding heterostructure material compositions?
CQ6	What are the the DOIs and/ the URLs of the scientific articles documenting a laser with performance property W or with heterostructure materials X or working mode Y or design type Z.
CQ7	What are the DOIs and URLs of the articles being referenced by a QCL device with property W or with heterostructure materials X or working mode Y or design type Z?

the specific queries run for each class of queries specified in Table 9. The queries range from simple to complex queries regarding the QCL properties and their provenance information.

Table 10
The Specific Queries Implemented for each Query Class

Query Number	Query Text
1.1	What are the possible material compositions of a QCL laser heterostructure with an LO Phonon Design Type ?
1.2	What are the possible material compositions of a QCL laser heterostructure with a Resonant Phonon Design Type?
1.3	What are the possible material compositions of a QCL laser heterostructure with a Bound to Continuum design type?
2.1	What are the working temperatures for a QCL laser operating in the continuous wave mode?
2.2	What are the working temperatures for a QCL laser operating in the pulsed mode?
3.1	What are the possible power values for a QCL laser with a heterostructure with material composition GaAs/Al0.15Ga0.85As?
3.2	What are the possible frequency values for a QCL laser with a heterostructure with material composition In0.53Ga0.47As/GaAs0.51Sb0.49?
3.3	Query 3.3: What are the possible working temperature values for a QCL laser with a heterostructure with material composition GaAs/Al0.3Ga0.7As?
4.1	Query 4.1: What are the possible heterostructure designs for a QCL device with a working temperature greater than 100 K in Pulsed Mode?
4.2	What are the possible heterostructure designs for a QCL device with an optical power less than 50 mW?
5.1	What are the possible heterostructure material compositions for a QCL device with a working temperature less than 85 K in the contionous wave mode?
5.2	What are the possible heterostructure material compositions for a QCL device with a lasing frequency greater than 1.5 THz?
6.1	What are the DOIs and URLS of scientific articles documenting QCL laser devices with an optical power greater than 10mW?
6.2	What are the DOIs and URLS of scientific articles documenting QCL laser devices with a working temperature greater than 100 K in pulse mode?
6.3	What are the DOIs and URLs of scientific articles documenting a QCL laser with a material composition of GaAs/Al0.25Ga0.75As?
6.4	What are the DOIs and URLs of scientific articles documenting QCL lasers with bound to continuum design type?
6.5	What are the DOIs and URLs of articles documenting QCL lasers with a heterostructure of material composition GaAs/Al0.15Ga0.85As, LO phonon design type and working temperatures greater than 70 K in pulse mode operation?
7.1	What are the DOIs and URLs of the articles being referenced by a QCL device with a working temperature greater than 225 K in the continuous wave mode?
7.2	What are the DOIs and URLs of the articles being referenced by a QCL device with an optical power less than 1 mW?
7.3	What are the DOIs and URLs of the articles being referenced by a QCL device with a lasing frequency greater than 2.5 THz and an LO Phonon design type?

We present an example of a scenario where an expert requires specific information regarding a QCL property and its relation to another property in order to make decisions regarding an optimal QCL design. We illustrate how this information can be retrieved via the KG using queries in Table 10.

Example 4.1: *Consider a scenario where a QCL expert is interested in the possible heterostructure materials composition of a QCL device with a certain working property, for instance, a lasing frequency value greater than 1.5 THz.*

The question in example 4.1 can be captured by a query for CQ 5.2 in Table 10. A corresponding SPARQL query and the retrieved results for query 5.2 are shown in Figure 5. All the queries are

```
PREFIX QpOnto:
<https://github.com/DeperiasKerre/qcl_Onto/blob/main/qclontology/version
-1.0/qclonto.owl#>
    PREFIX qudt:<https://qudt.org/schema/qudt/>
    PREFIX qudt_qv: <https://qudt.org/schema/qudt/#>
    SELECT DISTINCT ?material_composition
WHERE
{
 ?lf qudt_qv:hasQuantityValue ?fv;
 QpOnto:relatesToHeterostructure ?HS.
 ?fv qudt:numericValue ?value;
 qudt:hasUnit ?unit.
 ?HS QpOnto:hasMaterials ?HM.
 ?HM QpOnto:matFormula ?material_composition.
 FILTER(?unit=<https://qudt.org/vocab/unit/TeraHZ>&&?value>1.5).
}
```

material_composition
"GaAs/Al0.15Ga0.85As"
"InAs/AlAs0.16Sb0.84"
"GaAs/AlAs"
"Al0.03Ga0.97As"
"AlInGaAs"
"GaAs/Al0.14Ga0.86As"
"GaAs/AlGaAs"
"In0.53Ga0.47As/GaAs0.51Sb0.49"
"GaAs/Al0.25Ga0.75As"
"GaAs/Al0.1Ga0.9As"

Figure 5: A SPARQL Query and the Results for Query 5.2

successfully answered by the generated KG and the complete results are available in the GitHub repository [6].

The ability of the generated KG in successfully answering the competency questions indicates its capability in capturing QCL properties information from the various textual sources. This provides a unified platform that allows exploration of QCL properties in a structured manner to be able to derive insights on the relationship between the properties as opposed to manually exploring the textual documents for these properties. This is useful in scenarios where there is need for a quicker comparison of the various QCL properties for instance, the working properties for a particular laser design.

The ability to capture the provenance information for the QCL properties also makes it possible to track the sources of this information via permanent identifiers such as the DOI and the URL. With the generated KG, it is also possible to have a linked access to the references for the various QCL properties as some of the properties are based on other properties mentioned in the references. This therefore provides an efficient way of accessing this information for quicker analysis.

4.2.3. Knowledge Graph Completeness

Knowledge Graph completeness estimates the proportion of information contained in the KG in relation to the required information [82]. This metric is important in checking the possibility crucial information being left out in the KG. In this work, we assess the completeness of the KG in two dimensions: Schema completeness and Property Completeness. Schema completeness refers to the degree to which classes and properties are presented in a schema. On the other hand, property completeness refers to the extent of the missing property values of a specific kind of property [83].

Schema Completeness: In order to assess the schema completeness, we establish the mandatory properties for a class in order to determine the missing facts in a class [84]. A mandatory property for a given class instance refers to a relation that every instance of the class should be involved in, for instance in our case, every QCL optolectronic property should have a unit.

[6] https://github.com/DeperiasKerre/qKG

In order to determine the mandatory properties, we consider the following conditions: (i) QCL properties class instances (design/working properties) should be linked to the relevant provenance information (should contain links to papers describing them), (ii) The working temperatures should have a corresponding working mode, (iii) The optoelectronic properties should be linked to units, have quantity kind, quantity value and lastly (iv) The optoelectronic characteristics should be related to corresponding design features. The mandatory relations are therefore *wasAttributedTo, correspondsToWorkingMode, hasQuantityValue, hasQuantityKind, hasUnit* and *relatesToHeterostructure*. For each of these mandatory attributes, we determine the the ratio of instances that actually have the properties in the data captured by the KG. The results are presented in Table 11.

Table 11
Results for Schema Completeness Analysis

Mandatory Relation	Number of Instances in the KG (Expected to have Relation)	Instances with the Relation	Ratio
wasAttributedTo	100	100	1
correspondsToWorkingMode	36	36	1
hasQuantityValue	79	79	1
hasQuantityKind	79	79	1
hasUnit	69	69	1

As illustrated in Table 11, all the mandatory relations are adequately captured in the KG schema with no information missing for the KG classes. This implies that all the class instances are adequately associated with other instances, hence validating the KG schema quality.

Property Completeness: In order to assess the property completeness, we consider the data properties and check whether there is any missing information for these properties. We assert that all articles in the KG should have DOIs, all heterostructure materials should have a material formula, and all quantity values for the optoelectronic properties should have numeric values. We determine the ratio of the values as compared to the data properties for the instances in the KG. We present the results in Table 12.

Table 12
Results for Property Completeness Analysis

Property	Number of Instances in the KG (Expected to have Data)	Number of Data Values in KG	Ratio
DOI	474	474	1
matFormula	14	14	1
numericValue	69	69	1

The results in Table 12 indicate that all data values for property instances are captured in the KG. This implies that all articles in the KG have DOIs, all heterostructure materials have a corresponding material formula and all the quantity values in the KG have a corresponding numerical values. This validates the property completeness of the generated KG.

5. Conclusion and Future Work

In this paper, we address the issue of semantic enrichment of QCL properties in text by presenting an approach for generating a Knowledge Graph for QCL properties from text. The approach is composed of an information extraction pipeline for extracting QCL properties from text based on an LLM enabled RAG approach and the data enrichment part where all the data is mapped and the relationships interlinked. We evaluate the performance of the approach in QCL property extraction from text and the correctness of the generated Knowledge Graph in modeling the knowledge in the QCL properties domain.

The proposed information extraction approach presents competitive results indicating the model's ability to learn how to identify domain specific properties with the help of curated examples. The generated Knowledge Graph indicates its ability in modeling the knowledge in the QCL properties and their provenance information hence providing a semantically enriched, unified platform for quicker analysis and insights regarding the fabrication of QCL laser devices with target properties. Our work represents an important step towards the development of automated methods for extracting and representing complex scientific knowledge regarding QCL properties from text. This can be extended to other domains in order to develop methods for KG generation from text especially for domain specific KG generation from text. We believe that this approach has the potential to transform the way that researchers interact with scientific literature, and open up new avenues for discovery of facts in scientific literature.

The generated Knowledge Graph is however based on a limited number of articles and properties. This can be extended with more articles and other QCL properties such as the layer sequences, thickness, the current density among others in an incremental way using the same pipeline in an incremental way. There is also the need for more analysis of the proposed approach on other QCL properties with more diverse datasets. The concepts in the KG as represented by the QCL ontology can be extended using various AI techniques for instance, the use of LLMs in learning and extracting the new entities/concepts in the scientific literature for ontology population. This will enable timely update of the concepts in the ontology and the KG in general. Future works may include extending the KG with more data, concepts and proposing learning methods for the QCL laser working properties prediction based on design features. An example of learning methods for understanding the relationship between the design features that can be explored entails the use of KG embeddings for predicting the relationship between the QCL properties.

Availability of Materials

The source code and the materials used for the production of this work are publicly available at our GitHub repository: https://github.com/DeperiasKerre/qKG.

Acknowledgement

This work was funded by the French Embassy in Kenya (Scientific and Academic Cooperation Department) and the CNRS (under the framework "Dispositif de Soutien aux Collaborations avec l'Afrique sub-saharienne"). The authors would also like to thank Strathmore University, School of Computing and Engineering Sciences, and the Doctoral Academy for creating an opportunity for this work to be produced.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Faist, J., Capasso, F., Sivco, D. L., Hutchinson, A. L., Sirtori, C., & Cho, A. Y. (1995). Quantum cascade laser: a new optical source in the mid-infrared. Infrared Physics & Technology, 36(1), 99-103. https://doi.org/10.1016/1350-4495(94)00058-S

[2] Mittleman, D. M., Jacobsen, R. H., Neelamani, R., Baraniuk, R. G., & Nuss, M. C. (1998). Gas sensing using terahertz time-domain spectroscopy. Applied Physics B: Lasers & Optics, 67(3). https://doi.org/10.1007/s003400050520

[3] Federici, J. F., Schulkin, B., Huang, F., Gary, D., Barat, R., Oliveira, F., & Zimdars, D. (2005). THz imaging and sensing for security applications—explosives, weapons and drugs. Semiconductor science and technology, 20(7), S266. https://doi.org/10.1088/0268-1242/20/7/018

[4] Jepsen, P. U., Cooke, D. G., & Koch, M. (2011). Terahertz spectroscopy and imaging–Modern techniques and applications. Laser & Photonics Reviews, 5(1), 124-166. https://doi.org/10.1002/lpor.201000011

[5] Wubs, J.R., Macherius, U., Weltmann, K.D., Lü, X., Röben, B., Biermann, K., Schrottke, L., Grahn, H.T. & Van Helden, J.H.(2023). Terahertz absorption spectroscopy for measuring atomic oxygen densities in plasmas. Plasma Sources Science and Technology, 32(2), 025006. https://doi.org/10.1088/1361-6595/acb815

[6] Richter, H., Wienold, M., Schrottke, L., Biermann, K., Grahn, H. T., & Hübers, H. W. (2015). 4.7-THz local oscillator for the GREAT heterodyne spectrometer on SOFIA. IEEE Transactions on Terahertz Science and Technology, 5(4), 539-545. https://doi.org/10.1109/TTHZ.2015.2442155

[7] Shur, M., & Liu, X. (2022, March). Biomedical applications of terahertz technology. In Advances in Terahertz Biomedical Imaging and Spectroscopy (Vol. 11975, p. 1197502). SPIE. https://doi.org/10.1117/12.2604800

[8] Ohtani, K., Turčinková, D., Bonzon, C., Benea-Chelmus, I.C., Beck, M., Faist, J., Justen, M., Graf, U.U., Mertens, M. & Stutzki, J. (2016). High performance 4.7 THz GaAs quantum cascade lasers based on four quantum wells. New Journal of Physics, 18(12), 123004. https://doi.org/10.1088/1367-2630/18/12/123004

[9] Kerre, D., Laurent, A., Maussang, K., & Owuor, D. (2023, August). A text mining pipeline for mining the quantum cascade laser properties. In European Conference on Advances in Databases and Information Systems (pp. 393-406). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42941-5_34

[10] Kerre, D., Laurent, A., Maussang, K., & Owuor, D.(2025). A Concise Ontological Model of the Design and Optoelectronic Properties in the Quantum Cascade Laser Domain. Semantic Web, 16 (4), 1-17. https://doi.org/10.1177/22104968251359870

[11] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. & Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9. https://doi.org/10.1038/sdata.2016.18

[12] Swain, M. C., Cole, J. M. (2016). ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. Journal of chemical information and modeling, 56(10), 1894-1904.https://doi.org/10.1021/acs.jcim.6b00207

[13] Lowe, D. M., Sayle, R. A. (2015). LeadMine: a grammar and dictionary driven approach to entity recognition. Journal of cheminformatics, 7(1), 1-9.https://doi.org/10.1186/1758-2946-7-S1-S5

[14] Hawizy, L., Jessop, D. M., Adams, N., Murray-Rust, P. (2011). ChemicalTagger: A tool for semantic text-mining in chemistry. Journal of cheminformatics, 3, 1-13. https://doi.org/10.1186/1758-2946-3-17

[15] Leaman, R., Wei, C. H., Lu, Z. (2015). tmChem: a high performance approach for chemical named entity recognition and normalization. Journal of cheminformatics, 7(1), 1-10. https://doi.org/10.1186/1758-2946-7-S1-S3

[16] Rockt¨aschel, T., Weidlich, M., Leser, U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics, 28(12), 1633-1640. https://doi.org/10.1093/bioinformatics/bts183

[17] Sierepeklis, O., Cole, J. M. (2022). A thermoelectric materials database auto-generated from the scientific literature using ChemDataExtractor. Scientific Data, 9(1), 648. https://doi.org/10.1038/s41597-022-01752-1

[18] Dong, Q., Cole, J. M. (2022). Auto-generated database of semiconductor band gaps using chemdataextractor. Scientific Data, 9(1), 193. https://doi.org/10.1038/s41597-022-01294-6

[19] Zhao, J., Cole, J. M. (2022). A database of refractive indices and dielectric constants auto-generated using chemdataextractor. Scientific data, 9(1), 192. https://doi.org/10.1038/s41597-022-01295-5

[20] Mavracic, J., Court, C. J., Isazawa, T., Elliott, S. R., Cole, J. M. (2021). ChemDataExtractor 2.0:

Autopopulated ontologies for materials science. Journal of Chemical Information and Modeling, 61(9), 4280 4289. https://doi.org/10.1021/acs.jcim.1c00446

[21] Huang, S., Cole, J. M. (2020). A database of battery materials auto-generated using ChemDataExtractor. Scientific Data, 7(1), 260. https://doi.org/10.1038/s41597-020-00602-2

[22] Court, C. J., Cole, J. M. (2018). Auto-generated materials database of Curie and N´eel temperatures via semi-supervised relationship extraction. Scientific data, 5(1), 1-12. https://doi.org/10.1038/sdata.2018.111

[23] Cruse, K., Trewartha, A., Lee, S., Wang, Z., Huo, H., He, T., Kononova, O., Jain, A. & Ceder, G. (2022). Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities. Scientific Data, 9(1), 234. https://doi.org/10.1038/s41597-022-01321-6

[24] Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., Tshitoyan, V. & Ceder, G. (2019). Textmined dataset of inorganic materials synthesis recipes. Scientific data, 6(1), 203. https://doi.org/10.1038/s41597-019-0224-1

[25] Korvigo, I., Holmatov, M., Zaikovskii, A., Skoblov, M. (2018). Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules. Journal of cheminformatics, 10(1), 1-10. https://doi.org/10.1186/s13321-018-0280-0

[26] Zhao, J., Huang, S., Cole, J. M. (2023). OpticalBERT and OpticalTable-SQA: Text-and Table-Based Language Models for the Optical-Materials Domain. Journal of Chemical Information and Modeling. https://doi.org/10.1021/acs.jcim.2c01259

[27] Huang, S., Cole, J. M. (2022). BatteryBERT: A Pretrained Language Model for Battery Database Enhancement. Journal of Chemical Information and Modeling, 62(24), 6365-6377. https://doi.org/10.1021/acs.jcim.2c00035

[28] Thway, M., Low, A. K., Khetan, S., Dai, H., Recatala-Gomez, J., Chen, A. P., & Hippalgaonkar, K. (2024). Harnessing GPT-3.5 for text parsing in solid-state synthesis–case study of ternary chalcogenides. Digital Discovery, 3(2), 328-336. https://doi.org/10.1039/D3DD00202K

[29] Choi, J., & Lee, B. (2024). Accelerating materials language processing with large language models. Communications Materials, 5(1), 13. https://doi.org/10.1038/s43246-024-00449-9

[30] Polak, M. P., & Morgan, D. (2024). Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nature Communications, 15(1), 1569. https://doi.org/10.1038/s41467-024-45914-8

[31] Ye, Y., Ren, J., Wang, S., Wan, Y., Razzak, I., Xie, T., & Zhang, W. (2024). Construction of Functional Materials Knowledge Graph in Multidisciplinary Materials Science via Large Language Model. arXiv preprint arXiv:2404.03080. https://doi.org/10.48550/arXiv.2404.03080

[32] McCusker, J. P., Keshan, N., Rashid, S., Deagen, M., Brinson, C., & McGuinness, D. L. (2020, November). Nanomine: A knowledge graph for nanocomposite materials science. In International Semantic Web Conference (pp. 144-159). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-62466-8_10

[33] Mrdjenovich, D., Horton, M.K., Montoya, J.H., Legaspi, C.M., Dwaraknath, S., Tshitoyan, V., Jain, A. & Persson, K.A. (2020). Propnet: a knowledge graph for materials science. Matter, 2(2), 464-480. https://doi.org/10.1016/j.matt.2019.11.013

[34] Nie, Z., Liu, Y., Yang, L., Li, S., & Pan, F. (2021). Construction and application of materials knowledge graph based on author disambiguation: revisiting the evolution of LiFePO4. Advanced Energy Materials, 11(16), 2003580. https://doi.org/10.1002/aenm.202003580

[35] Venugopal, V., & Olivetti, E. (2024). MatKG: An autonomously generated knowledge graph in Material Science. Scientific Data, 11(1), 217. https://doi.org/10.1038/s41597-024-03039-z

[36] Zhao, X., Greenberg, J., McClellan, S., Hu, Y.J., Lopez, S., Saikin, S.K., Hu, X. & An, Y. (2021, December). Knowledge graph-empowered materials discovery. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 4628-4632). IEEE. https://doi.org/10.1109/BigData52589.2021.9671503

[37] Zhang, Y., Chen, F., Liu, Z., Ju, Y., Cui, D., Zhu, J., Jiang, X., Guo, X., He, J., Zhang, L. & Zhang, X. (2024). A materials terminology knowledge graph automatically constructed from text corpus. Scientific Data, 11(1), 600. https://doi.org/10.1038/s41597-024-03448-0

[38] Statt, M. J., Rohr, B. A., Guevarra, D., Suram, S. K., & Gregoire, J. M. (2023). The materials experiment

knowledge graph. Digital Discovery, 2(4), 909-914. https://doi.org/10.1039/D3DD00067B

[39] OpenAI. 2023. New Models and Developer Products Announced at DevDay. https://openai.com/index/new-models-and-developer-products-announced-at-devday/. Accessed: 2024-09-28.

[40] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H. & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997. https://doi.org/10.48550/arXiv.2312.10997

[41] Kerre, D., Laurent, A., Maussang, K., & Owuor, D., (2024). An Instruction Dataset for Extracting Quantum Cascade Laser Properties from Scientific Text. Recherche Data Gouv: DOI: 10.57745/U3U7XR. https://doi.org/10.57745/U3U7XR

[42] Kerre, D., Laurent, A., Maussang, K., & Owuor, D. (2025). An instruction dataset for extracting quantum cascade laser properties from scientific text. Data in Brief, 58, 111255. https://doi.org/10.1016/j.dib.2024.111255

[43] Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). Mpnet: Masked and permuted pretraining for language understanding. Advances in neural information processing systems, 33, 16857-16867. https://doi.org/10.5555/3495724.3497138

[44] Kumar, S., Williams, B. S., Hu, Q., & Reno, J. L. (2006). 1.9 THz quantum-cascade lasers with one-well injector. Applied Physics Letters, 88(12). https://doi.org/10.1063/1.2189671

[45] Williams, B. S., Callebaut, H., Kumar, S., Hu, Q., & Reno, J. L. (2003). 3.4-THz quantum cascade laser based on longitudinal-optical-phonon scattering for depopulation. Applied Physics Letters, 82(7), 1015-1017. https://doi.org/10.1063/1.1554479

[46] Kumar, S., Hu, Q., & Reno, J. L. (2009). 186 K operation of terahertz quantum-cascade lasers based on a diagonal design. Applied Physics Letters, 94(13).https://doi.org/10.1063/1.3114418

[47] Razavipour, S.G., Dupont, E., Chan, C.W.I., Xu, C., Wasilewski, Z.R., Laframboise, S.R., Hu, Q. & Ban, D. (2014). A high carrier injection terahertz quantum cascade laser based on indirectly pumped scheme. Applied Physics Letters, 104(4). https://doi.org/10.1063/1.4862177

[48] Dupont, E., Fathololoumi, S., Wasilewski, Z.R., Aers, G., Laframboise, S.R., Lindskog, M., Razavipour, S.G., Wacker, A., Ban, D. & Liu, H.C. (2012). A phonon scattering assisted injection and extraction based terahertz quantum cascade laser. Journal of Applied Physics, 111(7).https://doi.org/10.1063/1.3702571

[49] Hempel, M., Röben, B., Niehle, M., Schrottke, L., Trampert, A., & Grahn, H. T. (2017). Continuous tuning of two-section, single-mode terahertz quantum-cascade lasers by fiber-coupled, near-infrared illumination. AIP Advances, 7(5). https://doi.org/10.1063/1.4983030

[50] Wang, T., Liu, J. Q., Chen, J. Y., Liu, Y. H., Liu, F. Q., Wang, L. J., & Wang, Z. G. (2013). Continuous-wave operation of terahertz quantum cascade lasers at 3.2 THz. Chinese Physics Letters, 30(6), 064201. https://doi.org/10.1088/0256-307X/30/6/064201

[51] Lü, X., Röben, B., Schrottke, L., Biermann, K., & Grahn, H. T. (2021). Correlation between frequency and location on the wafer for terahertz quantum-cascade lasers. Semiconductor Science and Technology, 36(3), 035012. https://doi.org/10.1088/1361-6641/abdd4b

[52] Khabibullin, R.A., Shchavruk, N.V., Pavlov, A.Y., Klochkov, A.N., Glinskiy, I.A., Tomosh, K.N., Ponomarev, D.S., Cirlin, G.E. & Zhukov, A.E. (2019). Design and fabrication of terahertz quantum cascade laser with double metal waveguide based on multilayer GaAs/AlGaAs heterostructures. In IOP Conference Series: Materials Science and Engineering (Vol. 475, No. 1, p. 012020). IOP Publishing.https://doi.org/10.1088/1757-899X/475/1/012020

[53] Ushakov, D. V., Afonenko, A. A., Afonenko, A. A., Khabibullin, R. A., Fadeev, M. A., Gavrilenko, V. I., & Dubinov, A. A. (2024). Feasibility of GaAs/AlGaAs quantum cascade laser operating above 6 THz. Journal of Applied Physics, 135(13). https://doi.org/10.1063/5.0198236

[54] Deutsch, C., Krall, M., Brandstetter, M., Detz, H., Andrews, A.M., Klang, P., Schrenk, W., Strasser, G. & Unterrainer, K. (2012). High performance InGaAs/GaAsSb terahertz quantum cascade lasers operating up to 142 K. Applied Physics Letters, 101(21). https://doi.org/10.1063/1.4766915

[55] Brandstetter, M., Deutsch, C., Krall, M., Detz, H., MacFarland, D.C., Zederbauer, T., Andrews, A.M., Schrenk, W., Strasser, G. & Unterrainer, K. (2013). High power terahertz quantum cascade lasers with symmetric wafer bonded active regions. Applied Physics Letters, 103(17).https://doi.org/10.

1063/1.4826943

[56] Li, Y.Y., Liu, J.Q., Liu, F.Q., Zhang, J.C., Zhai, S.Q., Zhuo, N., Wang, L.J., Liu, S.M. & Wang, Z.G. (2016). High power-efficiency terahertz quantum cascade laser. Chinese Physics B, 25(8), 084206. https://doi.org/10.1088/1674-1056/25/8/084206

[57] Wang, X., Shen, C., Jiang, T., Zhan, Z., Deng, Q., Li, W., Wu, W., Yang, N., Chu, W. & Duan, S. (2016). High-power terahertz quantum cascade lasers with $\sim 0.23W$ in continuous wave mode. Aip Advances, 6(7).https://doi.org/10.1063/1.4959195

[58] Brandstetter, M., Kainz, M.A., Zederbauer, T., Krall, M., Schönhuber, S., Detz, H., Schrenk, W., Andrews, A.M., Strasser, G. & Unterrainer, K. (2016). InAs based terahertz quantum cascade lasers. Applied Physics Letters, 108(1).https://doi.org/10.1063/1.4939551

[59] Valmorra, F., Scalari, G., Ohtani, K., Beck, M., & Faist, J. (2015). InGaAs/AlInGaAs THz quantum cascade lasers operating up to 195 K in strong magnetic field. New Journal of Physics, 17(2), 023050.https://doi.org/10.1088/1367-2630/17/2/023050

[60] Walther, C., Scalari, G., Faist, J., Beere, H., & Ritchie, D. (2006). Low frequency terahertz quantum cascade laser operating from 1.6 to1. 8THz. Applied Physics Letters, 89(23). https://doi.org/10.1063/1.2404598

[61] Walther, C., Fischer, M., Scalari, G., Terazzi, R., Hoyler, N., & Faist, J. (2007). Quantum cascade lasers operating from 1.2 to1. 6THz. Applied Physics Letters, 91(13).https://doi.org/10.1063/1.2793177

[62] Hu, Q., Williams, B. S., Kumar, S., Callebaut, H., Kohen, S., & Reno, J. L. (2005). Resonant-phonon-assisted THz quantum-cascade lasers with metal–metal waveguides. Semiconductor science and technology, 20(7), S228.https://doi.org/10.1088/0268-1242/20/7/013

[63] Olariu, T., Senica, U., & Faist, J. (2024). Single-mode, surface-emitting quantum cascade laser at 26μm. Applied Physics Letters, 124(4).https://doi.org/10.1063/5.0176281

[64] Scalari, G., Amanti, M. I., Fischer, M., Terazzi, R., Walther, C., Beck, M., & Faist, J. (2009). Step well quantum cascade laser emitting at 3 THz. Applied physics letters, 94(4).https://doi.org/10.1063/1.3068496

[65] Schrottke, L., Lü, X., Rozas, G., Biermann, K., & Grahn, H. T. (2016). Terahertz GaAs/AlAs quantum-cascade lasers. Applied Physics Letters, 108(10).https://doi.org/10.1063/1.4943657

[66] Fasching, G., Benz, A., Unterrainer, K., Zobl, R., Andrews, A.M., Roch, T., Schrenk, W. & Strasser, G. (2005). Terahertz microcavity quantum-cascade lasers. Applied Physics Letters, 87(21).https://doi.org/10.1063/1.2136222

[67] Ohtani, K., Beck, M., Scalari, G., & Faist, J. (2013). Terahertz quantum cascade lasers based on quaternary AlInGaAs barriers. Applied Physics Letters, 103(4).https://doi.org/10.1063/1.4816352

[68] Deutsch, C., Benz, A., Detz, H., Klang, P., Nobile, M., Andrews, A.M., Schrenk, W., Kubis, T., Vogl, P., Strasser, G. & Unterrainer, K. (2010). Terahertz quantum cascade lasers based on type II InGaAs/GaAsSb/InP. Applied Physics Letters, 97(26).https://doi.org/10.1063/1.3532106

[69] Li, L., Chen, L., Zhu, J., Freeman, J., Dean, P., Valavanis, A., Davies, A.G. & Linfield, E.H. (2014). Terahertz quantum cascade lasers with >1 W output powers. Electronics letters, 50(4), 309-311.https://doi.org/10.1049/el.2013.4035

[70] Williams, B. S., Kumar, S., Qin, Q., Hu, Q., & Reno, J. L. (2006). Terahertz quantum cascade lasers with double-resonant-phonon depopulation. Applied physics letters, 88(26).https://doi.org/10.1063/1.2216112

[71] Adams, R.W., Vijayraghavan, K., Wang, Q.J., Fan, J., Capasso, F., Khanna, S.P., Davies, A.G., Linfield, E.H. & Belkin, M.A. (2010). GaAs/Al0. 15Ga0. 85As terahertz quantum cascade lasers with double-phonon resonant depopulation operating up to 172 K. Applied Physics Letters, 97(13).https://doi.org/10.1063/1.3496035

[72] Williams, B. S., Kumar, S., Callebaut, H., Hu, Q., & Reno, J. L. (2003). Terahertz quantum-cascade laser at $\lambda \approx 100$m using metal waveguide for mode confinement. Applied Physics Letters, 83(11), 2124-2126.https://doi.org/10.1063/1.1611642

[73] Luo, H., Laframboise, S. R., Wasilewski, Z. R., Aers, G. C., Liu, H. C., & Cao, J. C. (2007). Terahertz quantum-cascade lasers based on a three-well active module. Applied physics letters, 90(4). https://doi.org/10.1063/1.2437071

[74] Adams, R.W., Vizbaras, A., Jang, M., Grasse, C., Katz, S., Boehm, G., Amann, M.C. & Belkin, M.A. (2011). Terahertz sources based on intracavity frequency mixing in mid-infrared quantum cascade lasers with passive nonlinear sections. Applied Physics Letters, 98(15).https://doi.org/10.1063/1.3579260

[75] Bosco, L., Franckié, M., Scalari, G., Beck, M., Wacker, A., & Faist, J. (2019). Thermoelectrically cooled THz quantum cascade laser operating up to 210 K. Applied Physics Letters, 115(1).https://doi.org/10.1063/1.5110305

[76]

[77] Lander Gower, N., Levy, S., Piperno, S., Addamane, S. J., Reno, J. L., & Albo, A. (2023). Two-well injector direct-phonon terahertz quantum cascade lasers. Applied Physics Letters, 123(6). https://doi.org/10.1063/5.0155250

[78] Vitiello, M. S., Scamarcio, G., Spagnolo, V., Dhillon, S. S., & Sirtori, C. (2007). Terahertz quantum cascade lasers with large wall-plug efficiency. Applied Physics Letters, 90(19).https://doi.org/10.1063/1.2737129

[79] Li, H., Armiento, R., & Lambrix, P. (2020, November). An ontology for the materials design domain. In International Semantic Web Conference (pp. 212-227). Cham: Springer International Publishing.

[80] Schilling-Wilhelmi, M., Ríos-García, M., Shabih, S., Gil, M.V., Miret, S., Koch, C.T., Márquez, J.A. & Jablonka, K.M. (2025). From Text to Insight: Large Language Models for Materials Science Data Extraction. Chem. Soc. Rev. 54, 1125–1150. https://doi.org/10.1039/D4CS00913D

[81] Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical owl-dl reasoner. Journal of Web Semantics, 5(2), 51-53. https://doi.org/10.1016/j.websem.2007.03.004

[82] Issa, S., Adekunle, O., Hamdi, F., Cherfi, S. S. S., Dumontier, M., & Zaveri, A. (2021). Knowledge graph completeness: A systematic literature review. IEEE Access, 9, 31322-31339.

[83] Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211-218.

[84] Lajus, J., & Suchanek, F. M. (2018, April). Are all people married? Determining obligatory attributes in knowledge bases. In Proceedings of the 2018 World Wide Web Conference (pp. 1115-1124).

A. Appendix: Sample Generated Prompt

```
[INST]
Problem Definition: Extraction of quantum cascade laser properties from text entails extracting properties from a give
n text description. This should be done without providing any other additional information or explanations. The output
format should correspond to the one in the example sentences. Example sentences containing an instruction, input text
and the extracted properties are given below:

Example Sentences: ['From the following sentence, please extract the design type of the Quantum Cascade Semiconductor
laser device. Print none if the value does not exist in the input text. Input Text: Given the above-normal operating t
emperatures and duty cycles, we assert that utilizing direct LO-phonon-based depopulation proves to be a robust method
in the realization of long-wavelength THz quantum cascade lasers.\t\t\n\nExtracted Properties:\n\nLO-phonon', 'From th
e following sentence, please extract the design type of the Quantum Cascade Semiconductor laser device. Print none if
the value does not exist in the input text. Input Text: Considering the relatively high temperatures at which they ope
rate and the frequency of their cycles, we propose that leveraging direct LO-phonon-based depopulation is a dependable
means of obtaining quantum cascade lasers that emit long-wavelength THz frequencies.\t\t\n\nExtracted Properties:\n\nL
O-phonon', 'From the following sentence, please extract the design type of the Quantum Cascade Semiconductor laser dev
ice. Print none if the value does not exist in the input text. Input Text: Our experimental results show that by incre
asing the operating temperature to 100 degrees Celsius and doubling the duty cycle to 0.8, direct LO-phonon-based depo
pulation becomes an effective technique for generating quantum cascade lasers operating in the long-wavelength THz ran
ge.\n\nExtracted Properties:\n\nLO-phonon']

Instruction: Please extract the design type of the quantum cascade laser device. Please print none if the value is not
in text and do not give any explanations. In the output, just include only the extracted property. Input Text: We repo
rt the development of a quantum cascade laser, at l587.2 mm, corresponding to 3.44 THz or 14.2 meV photon energy. The
GaAs/Al0.15Ga0.85As laser structure utilizes longitudinal-optical LO-phonon scattering for electron depopulation. Lase
r action is obtained in pulsed mode at temperatures up to 65 K, and at 50% duty cycle up to 29 K. Operating at 5 K in
pulsed mode, the threshold current density is 840 A/cm2, and the peak power is approximately 2.5 mW. Based on the rela
tively high operating temperatures and duty cycles, we propose that direct LO-phonon-based depopulation is a robust me
thod for achieving quantum cascade lasers at long-wavelength THz frequencies.

[/INST]
```

Figure 6: Sample Regenerated Prompt