FAIRify Energy Data for Knowledge Graph Exploration

Md.Siam

Student ID: 19701075

Session: 2018-2019

Supervisor: Dr. Rudra Pratap Deb Nath,
Associate Professor

Department of Computer Science and Engineering
UNIVERSITY OF CHITTAGONG

This dissertation is submitted for the degree of
Bachelor of Science in Engineering (B.Sc. Engg.) in
Computer Science and Engineering

CSECU September 2024

Report Code:

University of Chittagong

Department of Computer Science and Engineering

8th Semester B.Sc. Engineering Examination 2022

Course No.: CSE 800

Title: FAIRify Energy Data for Knowledge Graph Exploration

Report Code:

University of Chittagong

Department of Computer Science and Engineering

8th Semester B.Sc. Engineering Examination 2022

Course No.: CSE 800

Title: FAIRify Energy Data for Knowledge Graph Exploration

Student Name: Md.Siam
Student ID: 19701075
Session: 2018-2019
Hall: Shaheed Abdur Rab

Signature of Student:

Submission Date: 22 September 2024

APPROVAL FOR SUBMISSION

This thesis entitled "*FAIRify energy data for knowledge graph exploration*", by *Md.Siam*, Student ID: 19701075, Session: 2018-2019, has been approved for submission to the Department of Computer Science and Engineering, University of Chittagong, in partial fulfillment of the requirements for the Bachelor of Science in Engineeering (B.Sc. Engg.) in Computer Science and Engineering.

Dr. Rudra Pratap Deb Nath
Associate Professor
Department of Computer Science and Engineering
University of Chittagong
Chattogram-4331, Bangladesh.
Email: rudra@cu.ac.bd

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Md.Siam
Student ID:19701075
Session:2018-2019
Department of Computer Science and Engineering
University of Chittagong
Chattogram-4331, Bangladesh.
Email: siammtalha8506@gmail.com

I dedicate this thesis to Almighty Allah, whose divine guidance has enlightened my way and given me strength in all difficulties. I would like to dedicate this work to my parents, who have stood by me through thick and thin, with great knowledge, wisdom, and words of encouragement to give me the bedrock in this journey. Their belief in me has never wavered and is the source of motivation and inspiration throughout. I would also like to thank all those who have assisted in my growth and success: mentors, colleagues, and friends. Their support and encouragement during this undertaking are beyond value and truly appreciated.

. . . .

Acknowledgements

I take this opportunity to express my deep and sincere gratitude to my supervisor, Dr. Rudra Pratap Deb Nath, for his inestimable guidance and support regarding this research. His great ability mingled with learned advice and constant encouragement shaped this venture of mine. Dr. Nath's quest for excellence and apt constructive feedback significantly contributed to refining this work further and adding value to it. I am also indebted to Safkat Raihan, my assistant supervisor, whose help and advice cannot be worthy of return as those have been highly influencing the advancement of this research. Expert advice by Safkat was complemented by his commitment and encouragement, thus forming a cornerstone for this study. I also wish to thank Tonmoy Chandra Das, Mohaimen Sarkar, and Tithi Rani Das for their support and cooperation that have enriched the research process. Their many and varied contributions, along with dedication to worthwhile debate, have been crucial in moving this work forward. I am grateful, finally, to everyone who kept me company on the way: friends, family, and colleagues alike. Because of all their encouragement and support, this research journey was so much better with them aboard.

Abstract

The Bangladesh Integrated Energy and Power Master Plan (IEPMP) outlines key strategies to support the country's development. IEPMP ensures energy security with net zero carbon emissions while Bangladesh makes its journey to become a prosperous, developed nation. Improving the economy leads to increasing the electricity demand, which falls under one of the priorities. In supporting energy security the plan aims at two main focus areas, and these are 1) utilization of renewable sources of energy at a low cost and 2) improvement in the utilization of data in the energy sector. Bangladesh Power Development Board (BPDB) shares the information about electricity checked through the National Load Dispatch Center (NLDC) in a PDF format. This format has a host of issues, including 1) inaccessible for machine interpretation, 2) unable to gain meaningful insights, 3) Failure to comply with Findable, Accessible, Interoperable, and Reusable (FAIR) principles, and 4) not externally linked with other datasets. To address these issues, the paper advocates for the creation of the Electricity Knowledge Graph of Bangladesh (ElecKG), which will be based on the FAIRification process to provide a more effective approach to exploring electricity data in Bangladesh. More concretely, data FAIRification makes them more machine-readable, effectively reusable, and aligned with the Semantic Web (SW) technologies. In such a context, we presented a Multidimensional (MD) framework showing ElecKG and evaluated it concerning FAIR principles, interlinking of external knowledge graphs, and compatibility with Online Analytical Processing (OLAP) operations. The questions generated from ElecKG reveal some details that are not captured in the current data. This research bridges the gap between energy data analysis and helps make informed decisions toward the energy sector.

Keywords: FAIR Data, Knowledge Graph, Linked Data, Multidimensional Analysis, Energy

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modern society has evolved greatly, and the role played by the energy sector has shaped many aspects of development that are associated with economic growth for many nations. Today's energy sector is characterized by rising complexity and a growing need to handle data effectively. By using structured data collection, integration, publication, and analysis, stakeholders may optimize decision-making processes within the energy sector. This data-driven approach [5], hence, provides greater efficiency in operations and, along with the sustainable utilization of resources, establishes sound policies for global energy security and environmental sustainability.

Bangladesh aims to achieve high-income country status by its 70th anniversary under Vision 2041, with plans [24] for a significant economic expansion that will drive increased energy demand. In response to global climate concerns, Bangladesh committed at the 2021 COP26 [27] to achieve up to 40% clean energy in its power generation mix. This strategy aims to ensure a secure, affordable, and sustainable energy supply, aligning national development goals with international commitments. Thus the energy sector 1) forms a cornerstone of economic foundation and development, and 2) significantly contributes to meeting the United Nations Sustainable Development Goal (SDG) 7: Affordable and Clean Energy [20]. The Bangladeshi government and the concerned organizations emphasize that this industry needs to be blended with energy management using advanced technology and data-driven insights to build a sustainable and resilient future [21]. Regarding this, BPDB [4] updates and publishes energy-related information, such as Energy Generation, Maximum Substation Loads, and Demand for Electricity, on its website. The National Load Dispatch Center (NLDC) [15] is responsible for monitoring the national grid round the clock, ensuring a reliable supply of electricity to the consumers by balancing demand and supply optimizing dispatch of electricity from different sources of power plants, and handling load shedding during peak periods or emergencies. Due to the intrinsic limitations, the available dataset of BPDB, in PDF format, has several challenges. It does not meet the requirements of Findable, Accessible, Interoperable, and Reusable (FAIR) data principles [18] because it is in an inactive format. In particular, the dataset's PDF format makes it harder to find and access makes it harder to integrate with other datasets, and doesn't offer machine-readability or the semantic structure needed for sophisticated data manipulation.

To address these issues and transform the dataset into a more user-friendly format, the ElecKG project employs a strict FAIRification [GO FAIR] procedure. The first step in this process is the annotation of the data using multidimensional (MD) modeling techniques, which add a rich semantic framework to the dataset. The data is then meticulously transformed into a semantic format appropriate for contemporary data systems through an Extract, Transform, Load (ETL) procedure [8]. After being altered, the data is put into a semantic repository and stored there in Resource Description Framework (RDF) [10] representation. The QB4OLAP [12] is the extension of the RDF Data Cube, QB vocabulary, it further provides the multidimensional semantics for business intelligence over RDF data and analytical queries using OLAP operations. ElecKG project uses this vocabulary to extend the dataset. With this SPARQL queries can conduct OLAP analysis and facilitate deep data exploration and insightful analysis. ElecKG not only fills in the gaps inherent in the original BPDB dataset, but it also converts this into an interoperable semantically rich format that can be integrated with other datasets, enhancing its usefulness and impact immensely.

1.1 Background

Today's world is focusing on sustainable energy systems [26] as it is beneficial for both the environment and the economy of a nation. Energy generation from renewable energy sources minimizes the effect on greenhouse effect by producing less greenhouse gas. Electricity generation from renewable energy sources is practiced by different nations to lower the environmental risk and boost the economy. The sustainable electrical power system is the main priority for notable growth in the economy. Constructing a sustainable electrical power system [26] requires several components and depends on the historical data analysis of the electricity data warehouse. Analysis of past data betterment the decision-making process to ensure efficient electricity generation, transmission, and consumption. Data reusability is a crucial part of the analysis. FAIRification [GO FAIR] of data comes in this scenario. FAIR data follows the FAIR principle and enlarges the data reusability. FAIRifying electricity data guarantees data is easily obtainable and interoperable with other data. Conversion of electricity data into RDF triples makes electricity data interoperable with other data as RDF triple [16] is formed by presenting data into subject, predicate, and object relationships. A set of electricity data triples is linked with each other and allows other datasets to be connected with it. Linked electricity data [6] adhere to the Semantic Web (SW) principles. The SW [19] is the network of interlinked data making the web more insightful and faster in interpretation by making data both human and machine-readable. Electricity data can be explored according to multidimensional aspects to get meaningful insights from it. Multidimensional electricity data from the electricity data cube consists of facts which are electricity data points and levels are collections of data attributes organized hierarchically into dimensions. This multi-dimensional data, commonly referred to as the Electricity Data Warehouse (EDH), can be used for OLAP queries. Multidimensional RDF data can be implemented using QB4OLAP vocabulary [12]. QB4OLAP vocabulary enables RDF triples that are compatible with OLAP operations: Roll-up, Drill-down, Slice, and Dice using SPARQL queries [25]. SPARQL queries were applied to get meaningful insights from the multidimensional electricity data cube which plays a key role in decision-making.

1.2 Problem Statement

The detailed workings of the power grid system are illustrated in Figure 1.1. It highlights the transmission and distribution lines linking power plants, substations, and consumers. The figure portrays the centralized monitoring system at the NLDC, which is responsible for overseeing electricity generation, distribution, and consumption data within the power grid. The challenge depicted in this figure lies in the existing data format available on the web, mostly in documents that cannot be read by machines. This format hinders efficient data comprehension by automated systems. To address this, the conversion of electricity data into a 5-star deployment scheme for Linked Data (LD) [31], adhering to Linked Open Government Data (LOGD) [6] standards, becomes essential.

This conversion process, following the FAIR principle, necessitates a series of transformations to elevate the data from a 1-star rating to a 5-star rating. Utilizing RDF data models, the structured representation of electricity data into RDF triples forms the foundation for this conversion. By assigning Universal Resource Identifiers (URIs) to these resources, the data becomes accessible and linkable in a semantic manner, thereby supporting the Semantic Web (SW) principles. Semantic energy data is necessary to improve system efficiency and speed up decision-making in the power grid architecture.

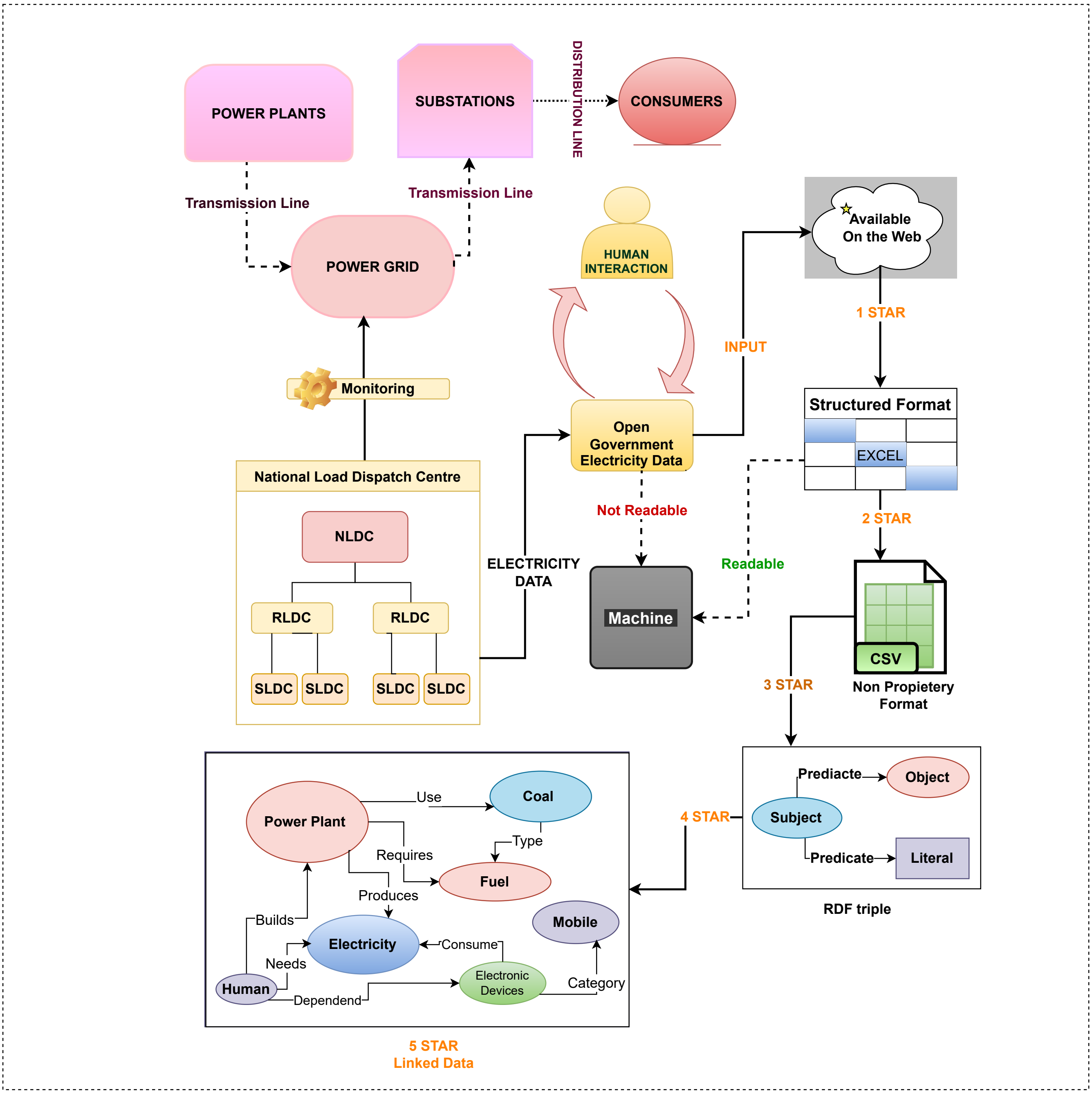


Fig. 1.1 Semantic conversion of BPDB electricity data

1.3 Research Challenges

Bangladesh electricity data is not aligned with the Semantic Web standard, which hampers its efficiency and effectiveness for extensive analysis of energy data. Machine-readable data is more efficient for the analysis of business intelligence. Converting PDF data into linked data that supports business intelligence is the main goal of this study. This section identifies several significant challenges that are faced in the current BPDB electricity dataset given below:
Research challenge 1. Turning non-FAIR electricity data into FAIR.
Research challenge 2. Represents unstructured data into multidimensional schema.
Research challenge 3. Converting electricity data into a knowledge graph, ElecKG.
Research challenge 4. Interlinked ElecKG with external knowledge graph.
Research challenge 5. Enabling OLAP operation on non-OLAP complement electricity data.

1.4 Summary of Contribution

The objective of this thesis was to propose a knowledge graph, ElecKG, which 1) follows the FAIR principles, 2) is modeled through multidimensional semantics which enables business intelligence to the energy field of the country, 3) is Externally linked with other datasets, and 4) is compatible with OLAP queries. A summary of the contribution of this paper is given below.

- Fairification of ElecKG. BPDB publishes non-FAIR electricity data on the web. Our research converts that data into a set of RDF triples, ElecKG, which is findable through Uniform Resource Identifier (URI) or Internationalized Resource Identifier (IRI), accessible on the web, interoperable with external knowledge graphs, and reuses the ElecKG datasets as it holds well-described metadata with clear and accessible data usage license. Which addresses our research challenge 1.

- Multidimensionality of ElecKG. ElecKG follows the multidimensional semantics using QB4OLAP vocabulary. The current electricity data of BPDB is not able to be analyzed according to multiple dimensions and fails to get insights from the data. Our research organizes PDF electricity data into multiple dimensions for better analysis, which addresses research challenge 2.

- Generating ElecKG. After representing electricity data in a multidimensional construct electricity PDF data follows the knowledge graph generation by semantic extract, transform, and load process. Electricity PDF data is extracted into CSV format according to the multidimensional model. The transformation process follows several steps: defining target TBox, generating source TBox, source-to-target mapping, and finally level and fact entry generation. The transformation process converts electricity CSV data into set of RDF triples which ends the ElecKG generation process addressing the research challenge 3.

- Linkablity of ElecKG. Data linking helps to get more meaningful information which makes a dataset valuable. The electricity data of BPDB is organized into the national grid areas which can be linked with Bangladesh divisions. Our research externally linked grid areas with the external knowledge graph which addresses the research challenge 4.
- Compatibility of ElecKG with OLAP queries. Multidimensional data warehouse supporting OLAP queries to get insights from it. Multidimensional data of the ElecKG is defined with QB4OLAP constructs enabling OLAP operations which addresses the research challenge 5.

1.5 Thesis Organization

The Thesis is organized into five chapters to present all the key aspects of this research. Chapter 1 provides a comprehensive introduction, covering the background, problem statement, and research challenges, laying the groundwork for the study. Chapter 2 reviews related work, examining previous research on knowledge graphs, multidimensional modeling, and electricity datasets to position this thesis in context with existing studies.

Chapter 3 presents the proposed methodology, starting with key concepts such as knowledge graphs and the QB4OLAP vocabulary. It then describes the source datasets and target TBox modeling, followed by the process of generating the Electricity Knowledge Graph (ElecKG) through extraction, transformation, and loading (ETL). This chapter is pivotal as it outlines the entire knowledge graph creation process.

Chapter 4 focuses on the experiments and evaluation of the ElecKG. It offers a structural overview of the dimensions and facts within the graph and assesses the availability, ETL performance, and quality through FAIRification, completeness, timeliness, and granularity metrics. The insights generated from ElecKG are also discussed.

Finally, Chapter 5 concludes the thesis by summarizing the key findings and suggesting future research directions, emphasizing areas where further exploration and improvements can be made.

Chapter 2

Related Work

In this section, we present how ElecKG is related to prior research and methodologies. The relational data model [7], proposed by Codd in 1970 and also the basis of the modern database industry, is mainly for Online Transactional Processing (OLTP). The model organizes data into tables with rows and columns. Meanwhile, for the purpose of analytics, the multidimensional data model [19], established in the 1990s, has large advantages. Where they differ, however, is that these multidimensional models represent data across a huge number of dimensions in a cube for complex analysis. ElecKG follows the multidimensional approach to offer comprehensive analysis about electricity data from several perspectives and improve the understanding of the domain of electricity [19]. Traditional relational databases, optimized for OLTP, capture current transactional data but fail to analyze historical trends or predict future scenarios. To remedy this situation, data warehousing techniques store historical data that are then analyzed using Online Analytical Processing, commonly referred to as OLAP. ElecKG utilizes this approach by keeping three years of electricity data, thus allowing for in-depth analysis and predictions of trends that might shape the future of Bangladesh's electricity sector [28]. The process of integrating and analyzing this data involves the ETL (Extract, Transform, Load) methodology. In the ElecKG framework, BPDB electricity data is extracted from PDF archives and converted into CSV format. The transformation phase includes semantic conversion steps such as Target TBox and source TBox generation using R2RML [2], Source to Target Mapping generation, and ABox creation utilizing semantic integration tools like SETLBI [9]. The final step involves loading ElecKG into a triple store like Virtuoso for efficient querying and analysis [8]. ElecKG is designed as a knowledge graph that embeds both the TBox and ABox semantically. Knowledge graphs represent data in such a way as to connect pieces of information by applying both human and machine-readable methods based on the principles of the Semantic Web, designed by Tim Berners-Lee. Translation of the data in BPDB into RDF format generates the graph structure behind the knowledge graph. RDF is the enabling factor of QB4OLAP [11], a representation method that models OLAP cubes by defining their dimensions, measures, and hierarchies using OWL [14]. SPARQL [25] is used to query OLAP data to take advantage of the semantics provided by OWL and RDF [16]. Pursuing the latest efforts on linking open government data, ElecKG is being published as linked open data, adhering to established principles and standards [22]. In addition, ElecKG follows the FAIR principles [33][18], that is, Findable, Accessible, Interoperable, and Reusable, which enhances usability. The FAIRification here involves access to non-FAIR electricity data from BPDB, analysis of the structure and relationships, and semantic modeling using OWL/RDF. It precisely models entities and their relations in the domain of electricity, hence allowing interoperability and reusability [GO FAIR]. FAIRified ElecKG publishes metadata and licensing information, hence making sure its accessibility and discoverability feature into the broader linked open data ecosystem.

Chapter 3

Proposed Methodology

We build the foundation of our investigation in this section by beginning with key concepts and terms. We subsequently go over the conceptual model intended for integration as well as the source datasets. Lastly, we go into depth about the construction process of ElecKG.

3.1 Preliminaries

The basic ideas and terminology that are the focus of this paper are introduced in this section. These include the fundamental concepts of RDF structures, multidimensional modeling, knowledge graphs, and the QB4OLAP vocabulary, all of which are necessary to comprehend the models and procedures employed in ElecKG.

3.1.1 Knowledge Graph

A knowledge graph is a linked structure of data that semantically represents knowledge. It is a directed graph. The nodes represent entities, while the arcs represent relations that connect these entities. In general, a KG consists of two main parts, namely the Terminological Box (TBox) and the Assertional Box (ABox). TBox defines the ontology of the data and how datasets are structured into concepts (classes) and properties (relationships between the classes). TBox is used to define the vocabulary of the ElecKG like dimensions, levels, level attributes, hierarchies, and measures of electricity data. ABox contains all the observations of the ElecKG that follow the TBox semantics. In this paper, we will present the KG using RDF graphs.

In RDF, graphs consist of sets of triples, denoted as $G \subseteq (U \cup B) \times U \times (U \cup B \cup L)$, where U refers to the set of Uniform Resource Identifiers (URIs), B is the set of blank nodes, and L represents the set of literals. Each RDF triple $t = (s, p, o)$ comprises a subject s, a predicate p, and an object o. Here, $s \in (U \cup B)$ indicates that subjects are either URIs or blank nodes, while $p \in U$ implies that predicates are always URIs. Objects $o \in (U \cup B \cup L)$ can be URIs, blank nodes, or literals.[3]

Moreover, literals L are split into two mutually exclusive subsets: plain literals L_p and typed literals L_t, such that $L = L_\mathrm{p} \cup L_\mathrm{t}$ and $L_\mathrm{p} \cap L_\mathrm{t} = \emptyset$. The relationship $U \cap B \cap L_\mathrm{p} \cap L_\mathrm{t} = \emptyset$ further ensures that URIs, blank nodes, and literals are non-overlapping sets, highlighting the distinct nature of these components in RDF.[3]

URIs (e.g., http://purl.org/linked-data/cube#Observation) serve as global identifiers for resources on the web. In RDF data representation, literals represent lexical values enclosed in quotation marks which are categorized into plain literals (e.g., "Eastern Zone") and typed literals (e.g., 20000000^^xsd:float, which denotes the total electricity demand in megawatts for the eastern zone) in Turtle syntax [32]. Blank nodes (e.g., _:bnode1) are denoted with an underscore prefix to represent anonymous entities without specific URIs. This framework is designed to cover all kinds of resource references within RDF graphs.

Prefixes:

```
@prefix xsd<http://www.w3.org/2001/XMLSchema#>.
@prefix mdAttribute:<http://bike-csecu.com/datasets/energy/abox/
    mdAttribute#>.
@prefix qb4olap<http://purl.org/qb4olap/cubes#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
```

Triple:

```
# Triple 1:
# Subject: <http://www.w3.org/2001/XMLSchema#string#5585>
# Predicate: <http://bike-csecu.com/datasets/energy/abox/
    mdAttribute#AreaName>
# Object: "Rangpur" (literal value)
<http://www.w3.org/2001/XMLSchema#string#5585>
        <http://bike-csecu.com/datasets/energy/abox/mdAttribute#
           AreaName>"Rangpur" .
```

3.1.2 Multidimensional modeling and the QB4OLAP vocabulary

Electricity data management concerns effective structuring and querying of data where multidimensional electricity data plays a pivotal role. It is specially laid out, especially in data warehousing and OLAP systems, in a format that allows for complex analysis and reporting. It provides a comprehensive view on data by structuring data from various

perspectives or dimensions. Multidimensional modeling of ElecKG helps represent various aspects of electricity data in the electricity domain (e.g., energy generation, load, and demand) in a structured and queryable format. It employs dimensions (e.g., TimePeriod, NationalGrid, PowerPlant) and measures (e.g., energy generated, max load, load shed) to create an in-depth view of the data. We use QB4OLAP [12] vocabulary to represent multidimensional electricity data to support OLAP features. This vocabulary is essential for capturing the complexity of multidimensional data, including hierarchies and rollup relationships, which are vital for analyzing data across various levels of granularity. For example, the `NationalGrid` dimension includes a hierarchy called `Grid` (`Area` → `Zone` → `CountryGrid`) enable users to explore and (dis)aggregate the energy generation, electricity demand, load, and load shed information of Bangladesh at various administrative levels of detail (see in Figure 3.1).
Table 3.1 lists the various constructs used in the Electricity Knowledge Graph (ElecKG) to express multidimensional data using the QB4OLAP vocabulary.

MD Concept	**Related QB4OLAP Element**
Dimension	`qb:DimensionProperty`
Hierarchy	`qb4o:Hierarchy`
Link between Dimensions and Hierarchies	`qb4o:inDimension`, `qb4o:hasHierarchy`
Fact Entity	`qb:Observation`
Level Definition	`qb4o:LevelProperty`
Level Member	`qb4o:LevelMember`
Level Connection	`qb4o:HierarchyStep`
Attributes	`qb4o:LevelAttribute`
Measures	`qb:MeasureProperty`
Aggregation Method	`qb4o:AggregateFunction`
Data Cube	`qb:DataStructureDefinition`
Dataset Structure	`qb:DataSet`

Table 3.1 Mapping of MD concepts to QB4OLAP constructs

3.2 Dataset Descriptions and Target TBox Design

In the following section, we outline the source datasets and introduce the target TBox that was utilized to integrate and interpret the knowledge in those datasets.

3.2.1 Source Dataset Descriptions

We use the electricity dataset from BPDB [4]. The NLDC monitors the country's power system and releases the information to the website of BPDB. The Daily Generation Archive subsection of the website of the BPDB publishes everyday electricity data on energy generation in the power plant, electricity demand and load shed at various grid locations, and maximum load at various substations. We have segregated the dataset into three key segments: Energy Generation, Max Load, and Electricity Demand & Load Shed.

Energy Generation Dataset: In the energy generation dataset, we consider data selectively as measures that can be effectively analyzed based on three key dimensions: PowerPlant, TimePeriod, and NationalGrid. Focusing our attention on the measures of energy generation, data is extracted from 1 May 2022 to 30 April 2024 for various powerplants situated in nine different areas of the national grid: Barishal, Chattogram, Dhaka, Khulna, Rajshahi, Rangpur, Mymensingh, Comilla, and Sylhet. The power plants, according to the fuel source, are divided into ten distinct categories such as gas, wind, solar, coal, etc., which further fall into renewable and non-renewable energy sources.

Max Load Dataset: The Max Load dataset does this with an intensive examination where the data is featured in three critical dimensions: GridSubstation, TimePeriod, and NationalGrid. Records encompass the maximum electrical load that substations have dealt with in order to handle the power distribution of a national grid. The analysis has consisted of extracting maximum load data from these substations, organized into several grid areas, from 1st May 2022 to 30th April 2024. We are therefore looking in three dimensions to get a thorough perspective of the performance and capacity of substations across the different regions in the national grid. This enlightens us on major trends, peak loads, and possible areas of concern within the infrastructure of the grid. Therefore, the detailed assessment of the parameters significantly contributes to the overall management and optimization of electrical load distribution across the national grid for operational efficiency and stability.

Demand and Load Shed Dataset: We collect holistic data on electricity demand and load shedding from 1st May 2022 to 30th April 2024 within the grid areas dimensioned by two very important dimensions: TimePeriod and NationalGrid. This was done by collecting data through detailed recording and analysis of cases of demand and load shedding in different grid areas. Thus, the integration of these two dimensions will enable us to understand how electricity demand varies in time and the incidence of load-shedding events in space across different parts of the national grid. It further enables analysis of the trend, assessment of the impact resulting from fluctuations in demand, and verification of whether any strategy to manage the load during that period is effective and very useful for drawing valuable inferences related to grid performance optimization in light of reliable electricity supply.

Dataset Segment	**Original Format**	**Time Period**	**Instances (Days)**	**Key Dimension**
Load Shed & Demand	PDF (Page 1)	1 May 2022 - 30 April 2024	730	NationalGrid, TimePeriod
Energy Generation	PDF (Page 2)			NationalGrid, TimePeriod, PowerPlant
Max Load	PDF (Page 3)			NationalGrid, TimePeriod, GridSubstation
Total	**PDF**	**2022-2024**	**730**	**Multidimensional**

Table 3.2 An overview of BPDB datasets

3.2.2 Target TBox definition

We define the TBox of ElecKG to integrate the electricity datasets. The TBox of ElecKG is shown in Figure 3.1. Data are structured in MD constructs similar to cubes, where dimensions span their axes. Instead of having single cube data, we use three distinct cuboids known as Load, Energy Generation, and Demand Loadshed for composing fact constellation schema where each level is in normalized form. The figure uses cubic boxes for fact table cubes and lists for levels, while dimensions are shown as blue rectangles and hierarchies as round rectangles. Crow's feet notation is used to depict cardinality, or how many instances of one entity could be associated with instances of another.

The TBox has four dimensions: GridSubstation, PowerPlant, TimePeriod, and NationalGrid. Four dimensions are used to analyze the electricity data from multiple perspectives, enabling OLAP operations. To describe the data being analyzed at various levels of abstraction, hierarchies are essential components of analytical applications. The `NationalGrid` dimension includes a hierarchy called `Grid` (`Area` → `Zone` → `CountryGrid`) that enables users to explore and (dis)aggregate the energy generation, electricity demand, load, and load shed information of Bangladesh at various administrative levels of detail. The TimePeriod dimension features a single hierarchy, Date, with the aggregation path `Day` → `Month` → `Year`. This hierarchy allows the analysis of energy data over different periods, facilitating trend analysis. The GridSubstation dimension represents various substations that manage electricity load at different levels. It features a hierarchy known as SubstationVoltageLevel with the aggregation path `Substation` → `VoltageLevel`. Energy distribution, load management, and operational details of substations are explored and analyzed in this hierarchy. It

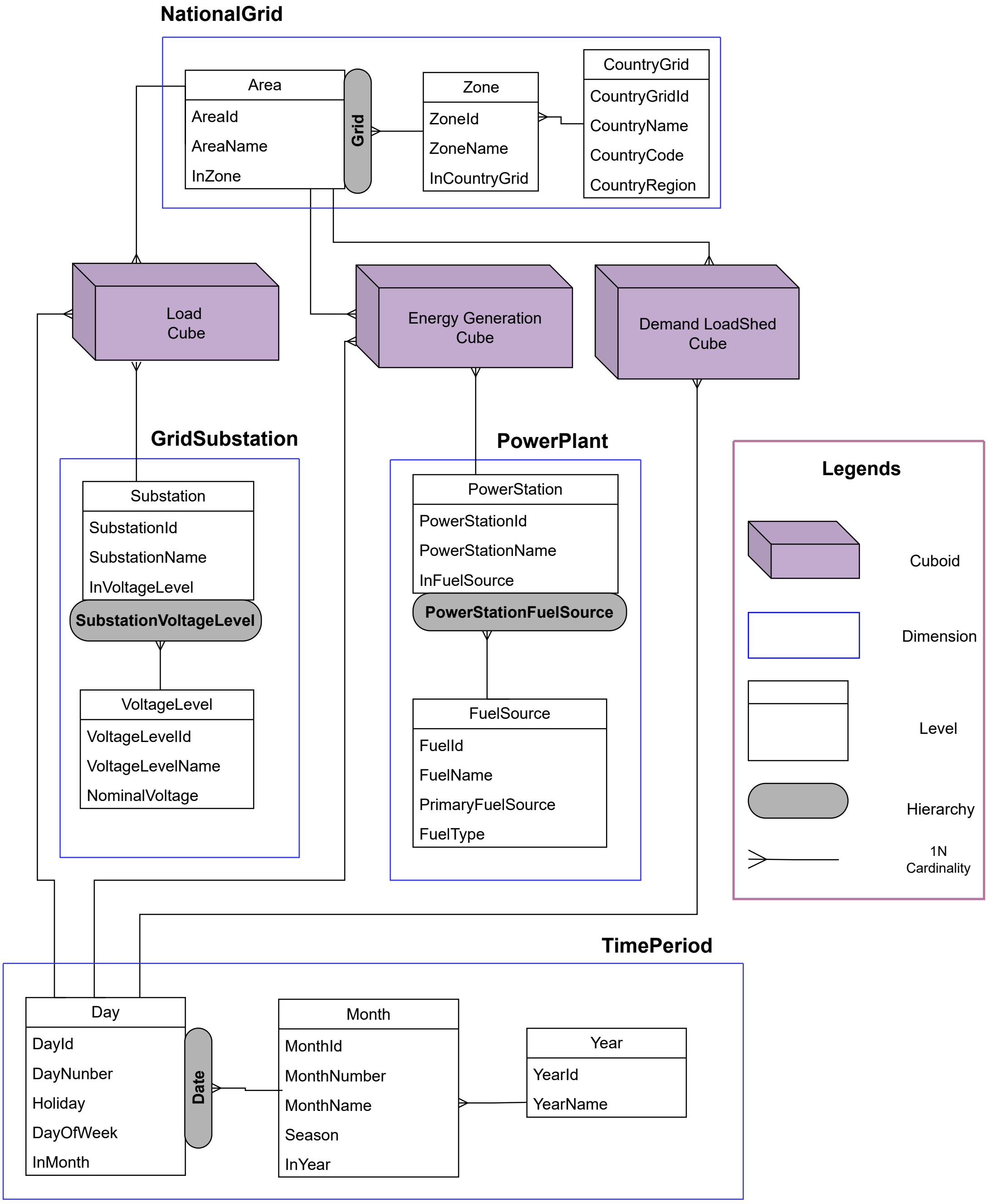


Fig. 3.1 Multidimensional data model diagram of ElecKG

also provides insights into how electricity is handled and distributed across different voltage levels.

The PowerPlant dimension categories PowerStations according to the FuelSources reflecting the energy generation of the power stations using different fuels. It includes a hierarchy named PowerStationFuelSource, with an aggregation path `PowerStation` → `FuelSource`. This structure enables detailed analysis of how different fuels contribute to energy production, offering a comprehensive view of the operational dynamics and fuel efficiency at each power station. Levels in a hierarchy are the specific layer or tiers that shows the broadness of the data and provides zoomed-in or zoom-out view on data.

In general, the attributes of an entity at a given level describe and define features or details about the entities at that level. Area level represents the different individual areas that may exist in any given zone. It has attributes like AreaId, unique for each area; AreaName, the name of the area; and InZone, which relates the area to its respective zone. A detailed view of energy distribution in certain areas of the grid is provided at this level. Zone level provides for the country grid-specific zones. The attributes are ZoneId, a unique identifier of the zone; ZoneName, the zone's name; and InCountryGrid, the link attribute from the zone to the country grid. The levels help in controlling the energy distribution in various zones. CountryGrid level is the national grid at the country scale. It contains attributes like CountryGridId, which is a unique identification for the grid; CountryName, defining the name of the country; CountryCode, which is the ISO for the country; and CountryRegion, defining the region within the country. Put altogether, these attributes summarize the energy distribution network at a national overview. Substation level: This identifies specific substations using such attributes as SubstationId, which is unique in nature; SubstationName; and InVoltageLevel, which links it with the associated voltage level. VoltageLevel level: This describes voltage levels such attributes as VoltageLevelId, unique identifier; VoltageLevelName; and NominalVoltage, the standard voltage of the level. PowerStation level: This models individual power stations using such attributes as PowerStationId, which is unique in nature; PowerStationName; and InFuelSource, linking to the fuel source applied. FuelSource level details the fuel types used, with attributes including FuelId (unique ID), FuelName, PrimaryFuelSource (indicating if it's the main source), and FuelType (the category of fuel). Day level represents individual days with attributes like DayId (unique ID), DayNumber, Holiday (indicating if it's a holiday), DayOfWeek, and InMonth (linking to the month). The Month level consists of months of a year. The characteristics of this level are MonthId, MonthNumber, MonthName, Season, and InYear that links to the year. The Year level is established by a year, and it is characterized by the following attributes: YearId, YearName.

3.3 Methodology of Generating Electricity Knowledge Graph (ElecKG)

Figure 3.2 illustrates the step-by-step ETL process for generating ElecKG from the BPDB Daily Generation Archive. Each step of this process is aimed at different sections of data extraction, transformation, and loading to make it a knowledge graph accessible for analysis. Each component depicted here is explained in detail as follows:

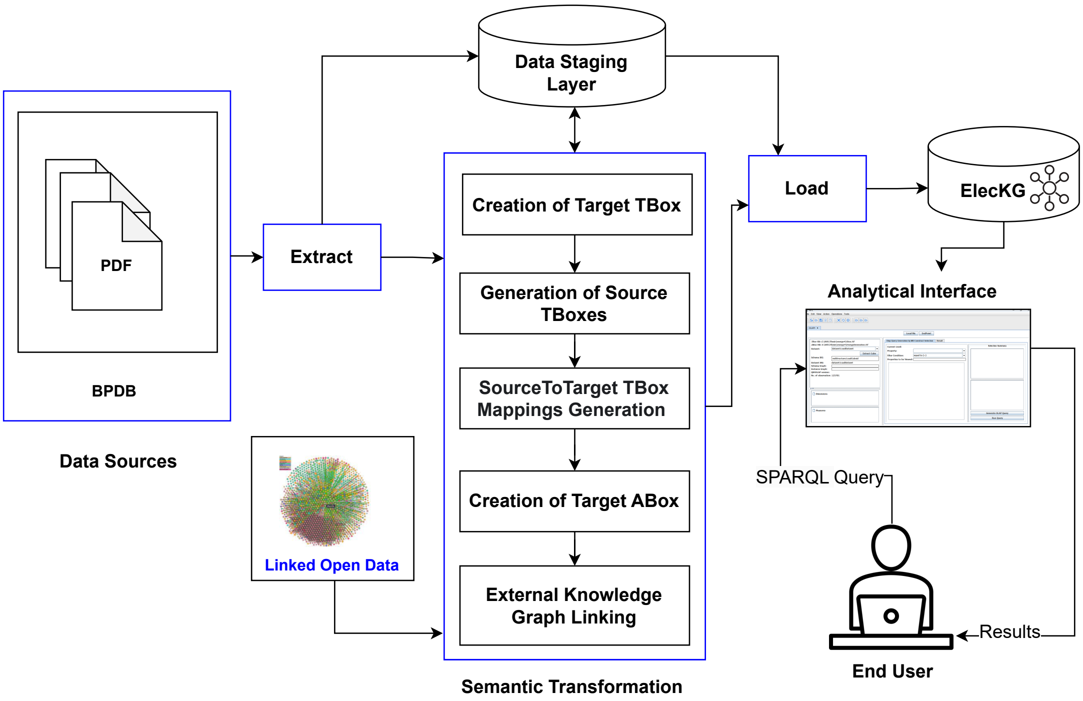


Fig. 3.2 End-to-end ETL process for ElecKG

Steps in the ETL Process

The process starts with the BPDB Daily Generation Archive, raw data in PDF format, which includes the source of energy generation, maximum load, electricity demand, and load shedding. After that comes the extraction process, where the PDF files are converted into CSV format so that they are more manipulable and analyzable with the data organized into a tabular form. The target TBox of ElecKG defines the schema by specifying classes, properties, and relationships to model the data about electricity.

Once the target TBox is specified, the raw data structure is reflected by creating source TBoxes from the extracted CSV data. This ensures that the data is represented correctly in their pre-transformation state. Mappings between the source TBoxes and the target TBox are created to make the data flow in conformity with the predefined schema of the target TBox. Therefore, the flow of data coming from various sources can be harmonized and in order.

After that, the semantic transformation step includes applying the required transformations on the data according to the mappings and definitions of the target TBox, so that the raw data will conform to the semantics specified in the target TBox. The ABox will be generated, based on the results of the transformation, with assertions corresponding to actual data points and relationships defined by the target TBox.

This is further enriched by linking the resources both in the target ABox and TBox to external knowledge graphs available in the Linked Open Data (LOD) cloud, which provides additional contextual information. The last step is loading a completed knowledge graph into a triple store, where it gets stored and is queried or accessed efficiently. A knowledge graph analytical interactive interface is provided for exploring and analyzing knowledge graph data. Complex queries can also be carried out using SPARQL, enabling retrieval and analysis of detailed data from ElecKG.

3.3.1 Extraction

Electricity data is publicly available on the BPDB website under the BPDB Daily Generation Archive, accessible at https://misc.bpdb.gov.bd/daily-generation-archive. These data are provided as PDFs and contain electricity generation, demand, and load shedding. This daily BPDB electricity data is divided over three pages per PDF: Page 1 covers information on the load shedding in the various grid areas; Page 2: energy generation at various power stations; Page 3: maximum load recorded at substation level and the electricity demand at grid areas. First, the daily PDFs for each month are downloaded from the BPDB website. Since the data is divided across three pages, these pages are merged for each month using the PDF processing tool PDFtk, available at https://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/, which allows the consolidation of data into a single PDF file per month. The merged PDF files are then converted into CSV format using the Online2PDF tool, accessible at https://online2pdf.com/, with irrelevant data filtered out during the conversion to produce a streamlined CSV output. Finally, the CSV data is cleansed to ensure that it is organized and formatted correctly, making it ready for further analysis.

3.3.2 Transformation

Following the extraction and preparation of electrical data, there are several crucial phases involved in transforming the data into a semantic format for the semantic web. Initially, the ontology and constructs required for the intended semantic model are defined in a Target TBox. The ontologies and schemas of the original data are then represented by Source TBoxes. Next, to ensure seamless integration, there is the Generation of Source-to-Target Mappings, which translates concepts and relationships from the source TBoxes to the target TBox. Following that, the Target ABox is built by adding data and instances that match the target TBox, enhancing the semantic model. To improve its overall usefulness and incorporate more context, the augmented ElecKG is finally linked to pertinent External Knowledge Graphs.

Target TBox Creation. This step aims at specifying the target TBox according to Section 3.2.2, by using RDFS, OWL, and QB4OLAP constructs at the RDF level defined in Section 3.1.2. The QB4OLAP constructs extend basic OWL classes and RDF properties with MD semantics. It is possible to write a TBox with MD semantics directly by hand, but it is more practical to use tools like Protege and SETLBI. A sample of such a process is given by Listing 3.1, which depicts a fragment of our TBox annotated with QB4OLAP constructs.

```
@prefix owl: <http://www.w3.org/2002/07/owl#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix qb: <http://purl.org/linked-data/cube#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix qb4o: <http://purl.org/qb4olap/cubes#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix mdStructure: <http://bike-csecu.com/datasets/energy/abox/mdStructure#>.
@prefix mdAttribute: <http://bike-csecu.com/datasets/energy/abox/mdAttribute#>.
@prefix mdProperty: <http://bike-csecu.com/datasets/energy/abox/mdProperty#>.
@prefix dataset: <http://bike-csecu.com/datasets/energy/abox/data#>.
@prefix onto: <http://bike-csecu.com/datasets/energy/abox/>.

#DIMENSIONS
mdProperty:PowerPlant a qb:DimensionProperty;
  rdfs:label "Power Plant represents the power stations that produce energy using fuel"@en;
  qb4o:hasHierarchy mdStructure:PowerStationFuelSource;
  rdfs:range xsd:string.
```

```

#HIERARCHIES

mdStructure:PowerStationFuelSource a qb4o:Hierarchy;
  rdfs:label "Power Station Fuel Source Hierarchy"@en;
  qb4o:inDimension mdProperty:PowerPlant;
  qb4o:hasLevel mdProperty:FuelSource, mdProperty:PowerStation.

#LEVELS
mdProperty:FuelSource a qb4o:LevelProperty;
  rdfs:range onto:FuelSource;
  qb4o:hasAttribute mdAttribute:PrimaryFuelSource,
      mdAttribute:FuelType, mdAttribute:FuelName,
      mdAttribute:FuelId;
  rdfs:label "Fuel used by power station to generate electric
      energy"@en.

mdProperty:PowerStation a qb4o:LevelProperty;
  rdfs:range onto:PowerStation;
  qb4o:hasAttribute mdAttribute:PowerStationName,
      mdAttribute:PowerStationId, mdAttribute:InFuelSource;
  rdfs:label "Power station generating electric energy"@en.

#Hierarchy steps
_:hs6 a qb4o:HierarchyStep;
  qb4o:inHierarchy mdStructure:PowerStationFuelSource;
  qb4o:childLevel mdProperty:PowerStation;
  qb4o:parentLevel mdProperty:FuelSource;
  qb4o:pcCardinality qb4o:OneToMany;
  qb4o:rollup mdAttribute:InFuelSource.

#ATTRIBUTES

mdAttribute:PowerStationName a qb4o:LevelAttribute;
  rdfs:range xsd:string;
  rdfs:label "Each unique power station is identified by a name"
      @en.

mdAttribute:PrimaryFuelSource a qb4o:LevelAttribute;
  rdfs:range xsd:string;
  rdfs:label "Each unique fuel source is identified by whether it
      is a primary source or not"@en.

#MEASURES
```

```
56 mdProperty:EnergyGenerated a qb:MeasureProperty;
57   rdfs:label "Energy generated by power station"@en;
58   rdfs:range xsd:float.
59
60 #CUBES
61 mdStructure:EnergyGenerationCube a qb:DataStructureDefinition;
62   dct:conformsTo <http://purl.org/qb4olap/cubes>;
63   qb:component [ qb4o:dimension mdProperty:TimePeriod;
       qb4o:cardinality qb4o:OneToMany];
64   qb:component [ qb4o:dimension mdProperty:NationalGrid;
       qb4o:cardinality qb4o:OneToMany];
65   qb:component [ qb4o:dimension sdw:PowerPlant; qb4o:cardinality
       qb4o:OneToMany];
66   qb:component [ qb:measure mdProperty:EnergyGenerated;
       qb4o:aggregateFunction qb4o:avg, qb4o:count, qb4o:min,
       qb4o:max, qb4o:sum].
67
68 #CUBOIDS
69
70 mdStructure:EnergyGenerationCuboid a qb:DataStructureDefinition;
71   qb4o:isCuboidOf mdStructure:EnergyGenerationCube;
72   dct:conformsTo <http://purl.org/qb4olap/cubes>;
73   qb:component [ qb:measure mdProperty:EnergyGenerated;
       qb4o:aggregateFunction qb4o:avg, qb4o:count, qb4o:min,
       qb4o:max, qb4o:sum];
74   qb:component [ qb4o:level mdProperty:PowerStation;
       qb4o:cardinality qb4o:OneToMany];
75   qb:component [ qb4o:level mdProperty:Area; qb4o:cardinality
       qb4o:OneToMany];
76   qb:component [ qb4o:level mdProperty:Day; qb4o:cardinality
       qb4o:OneToMany].
77
78 #DATASETS
79 dataset:EnergyGenerationDataset a qb:DataSet;
80   qb:structure mdStructure:EnergyGenerationCuboid.
```

Listing 3.1 TBox representation of ElecKG using QB4OLAP vocabulary

In QB4OLAP, dimensions, hierarchies, and levels are structured using the elements qb4o:DimensionProperty, qb4o:Hierarchy, and qb4o:LevelProperty (lines 13-30). Each dimension can include one or more hierarchies; for example, mdProperty:PowerPlant contains a single hierarchy. The relationships between a dimension and its hierarchies are established using qb4o:inDimension and its inverse, qb4o:hasHierarchy (lines 16 and

23). A hierarchy is composed of levels organized in a particular sequence, which is defined by `qb4o:HierarchyStep`. For instance, `mdStructure:PowerStationFuelSource` consists of two levels (line 24). The hierarchy step demonstrates that power stations are grouped by fuel sources (lines 37-44), meaning `mdProperty:PowerStation` is aggregated into `mdProperty:FuelSource` via the roll-up property `mdAttribute:InFuelSource` (line 43). Levels have attributes defined by `qb4o:LevelAttribute` (line 47), which can be either object properties (linking instances) or datatype properties (connecting instances to literal values). For example, the `mdAttribute:InFuelSource` attribute links `mdProperty:PowerStation` to `mdProperty:FuelSource` instances. In QB4OLAP, the structure of a cube is outlined through dimensions and measures (lines 60-67), while a cuboid is described by levels and measures (lines 68-77). Both structures are defined using `qb:DataStructureDefinition`. Measures, that capture numerical data for analysis (lines 55-59), are specified using `qb:MeasureProperty`. Finally, a dataset is created with `qb:Dataset` and linked to the specified cube or cuboid structure (lines 78-80).

Source TBox Creation: Target and source construct mapping is required to populate ElecKG at the TBox level. This entails taking TBoxes from pre-existing sources and improving them using RDFS and OWL components. Data is extracted, cleaned up, and stored in tabular form during the extraction step. Next, by considering table names as OWL classes and attribute names as OWL properties, each table's schema is transformed into a source TBox. Direct mapping vocabulary or RDB to RDF Mapping Language (R2RML) can also be used by users in this procedure. Listing 3.2 shows that the RDF data defines <http://bike-csecu.com/datasets/energy/abox/PowerStation> as an OWL class (lines 6-7), as indicated by the statement a <http://www.w3.org/2002/07/owl#Class>. This class represents the concept of a PowerStation within the ontology. Additionally, <http://bike-csecu.com/datasets/energy/abox/InFuelSource> is defined as an OWL datatype property (line 9). This property is associated with the `PowerStation` class and has a range of `xsd:string`, meaning its values are expected to be strings.

```
1 @prefix owl: <http://www.w3.org/2002/07/owl#>.
2 @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
3 @prefix onto: <http://bike-csecu.com/datasets/energy/abox/>.
4 @prefix dataset: <http://bike-csecu.com/datasets/energy/abox/data#
     >.
5
6 <http://bike-csecu.com/datasets/energy/abox/PowerStation>
7     a <http://www.w3.org/2002/07/owl#Class>.
8
9 <http://bike-csecu.com/datasets/energy/abox/InFuelSource>
```

```
    a <http://www.w3.org/2002/07/owl#DatatypeProperty>;
    <http://www.w3.org/2000/01/rdf-schema#domain>
        <http://bike-csecu.com/datasets/energy/abox/PowerStation>;
    <http://www.w3.org/2000/01/rdf-schema#range>
        <http://www.w3.org/2001/XMLSchema#string>.

<http://bike-csecu.com/datasets/energy/abox/PowerStationName>
    a <http://www.w3.org/2002/07/owl#DatatypeProperty>;
    <http://www.w3.org/2000/01/rdf-schema#domain>
        <http://bike-csecu.com/datasets/energy/abox/PowerStation>;
    <http://www.w3.org/2000/01/rdf-schema#range>
        <http://www.w3.org/2001/XMLSchema#string>.

<http://bike-csecu.com/datasets/energy/abox/PowerStationId>
    a <http://www.w3.org/2002/07/owl#DatatypeProperty>;
    <http://www.w3.org/2000/01/rdf-schema#domain>
        <http://bike-csecu.com/datasets/energy/abox/PowerStation>;
    <http://www.w3.org/2000/01/rdf-schema#range>
        <http://www.w3.org/2001/XMLSchema#string>.
```

Listing 3.2 Source TBox of PowerStation of the PowerStation dataset

Source To Target TBox Mapping. Source-to-target mapping (S2TMAP) is a structured methodology for aligning elements between different ontologies, ensuring data integration and interoperability. In the provided RDF Turtle syntax example, this approach is illustrated through three key components: the Dataset Mapper, Concept Mapper, and Property Mapper. As shown in lines 6-8, `map:PowerPlantDataset` is defined as a dataset using `map:Dataset`, linking the source TBox (<http://bike-csecu.com/datasets/energy/abox/tbox#>) to the target TBox (<http://www.onto.org/schema#energyTbox>). The Concept Mapper, defined on lines 11-18 as `map:PowerStation_PowerStation`, associates with the dataset `map:PowerPlantDataset` and maps the concept `PowerStation` from the source (<http://bike-csecu.com/datasets/energy/abox/PowerStation>) to the target concept `mdProperty:PowerStation`, using `PowerStationId` as the common property. Lastly, the Property Mapper, starting at line 21 with`map:PropertyMapper_01_PowerStationId_PowerStationId`, details the mapping of the `PowerStationId` property from the source (<http://bike-csecu.com/datasets/energy/abox/PowerStationId>) to the target (`mdAttribute:PowerStationId`), and is linked to the Concept Mapper defined earlier (line 22). This well-organized approach ensures precise and consistent mapping between different ontologies, facilitating effective data integration across systems.

```
1 @prefix map:<http://www.map.org/example#>.
```

```
@prefix dataset:<http://bike-csecu.com/datasets/energy/abox/data#
    >.

# Dataset mapper
map:PowerPlantDataset a map:Dataset;
map:sourceTBox <http://bike-csecu.com/datasets/energy/abox/tbox#>;
map:targetTBox <http://www.onto.org/schema#energyTbox>.

# Concept mapper
map:PowerStation_PowerStation a map:ConceptMapper;
map:dataset map:PowerPlantDataset;
map:iriValue <http://bike-csecu.com/datasets/energy/abox/
    PowerStationId>;
map:iriValueType map:SourceAttribute;
map:matchedInstances "All";
map:sourceConcept <http://bike-csecu.com/datasets/energy/abox/
    PowerStation>;
map:targetCommonProperty mdAttribute:PowerStationId;
map:targetConcept mdProperty:PowerStation.

# Property mapper
map:PropertyMapper_01_PowerStationId_PowerStationId a
    map:PropertyMapper;
map:ConceptMapper map:PowerStation_PowerStation;
map:sourceProperty <http://bike-csecu.com/datasets/energy/abox/
    PowerStationId>;
map:sourcePropertyType map:SourceProperty;
map:targetProperty mdAttribute:PowerStationId.
```

Listing 3.3 SourceToTarget mapping definitions between the target TBox mdProperty: and the source TBox dataset:PowerStation

Target ABox Generation. The Target ABox Generation process transforms cleansed datasets into an ABox for ElecKG, following the ElecKG TBox semantics and mappings. It starts by collecting the target TBox, cleansed datasets, and mapping definitions. These mappings align source data with the target TBox structure. The ABox is built by instantiating data according to the target TBox's class and property definitions. Each individual identifies with the respective concepts and attributes. In QB4OLAP, each member of a level has an IRI and proper attributes for deep analysis or querying of the data. In this way, it is ensured that the ABox is a proper image of the semantics and structure of the target TBox for effective integration and exploitation of data. For example, Listing 3.4 shows a PowerStation defined as a `qb4o:LevelMember`. It is enhanced with attributes such as its PowerStationId, PowerStationName, and FuelSource. Additionally, it is linked to `mdProperty:FuelSource`,a level member, with the roll-up property InFuelSource.

```
@prefix dataset: <http://bike-csecu.com/datasets/energy/abox/data#>.
@prefix onto: <http://bike-csecu.com/datasets/energy/abox/>.

<http://bike-csecu.com/datasets/energy/abox/PowerStation#sylhet50MWPP-EPL->
    a       <http://purl.org/qb4olap/cubes#LevelMember>;
    <http://bike-csecu.com/datasets/energy/abox/mdAttribute#InFuelSource>
            <http://bike-csecu.com/datasets/energy/abox/FuelSource#150001>;
    <http://bike-csecu.com/datasets/energy/abox/mdAttribute#PowerStationId>
            "sylhet50MWPP(EPL)" ;
    <http://bike-csecu.com/datasets/energy/abox/mdAttribute#PowerStationName>
            "Sylhet 50 MW PP (EPL)" ;
    <http://purl.org/qb4olap/cubes#memberOf>
            <http://bike-csecu.com/datasets/energy/abox/mdProperty#PowerStation>.
```

Listing 3.4 A level member of mdProperty:PowerStation

```
@prefix dataset: <http://bike-csecu.com/datasets/energy/abox/data#>.
@prefix onto: <http://bike-csecu.com/datasets/energy/abox/>.

<http://bike-csecu.com/datasets/energy/abox/data/EnergyGenerationDataset#28-May-22_bhola225MWCCPP_1006>
        a       <http://purl.org/linked-data/cube#Observation>;
        <http://bike-csecu.com/datasets/energy/abox/mdProperty#Area>
                <http://bike-csecu.com/datasets/energy/abox/Area#1006>;
        <http://bike-csecu.com/datasets/energy/abox/mdProperty#Day>
                <http://bike-csecu.com/datasets/energy/abox/Day#28-May-22>;
        <http://bike-csecu.com/datasets/energy/abox/mdProperty#EnergyGenerated>
                "1323000" ;
        <http://bike-csecu.com/datasets/energy/abox/mdProperty#PowerStation>
                <http://bike-csecu.com/datasets/energy/abox/PowerStation#bhola225MWCCPP>;
        <http://purl.org/linked-data/cube#dataSet>
                <http://bike-csecu.com/datasets/energy/abox/data#EnergyGenerationDataset>.
```

Listing 3.5 An observation of the mdStructure:EnergyGenerationCuboid cuboid

To represent a fact, RDF uses an `Observation` (an instance of `qb:Observation`) as shown in Listing 3.5. Each observation is uniquely identified by an IRI and provides specific data related to energy generation. In this case, the fact describes energy generation on a specific date (`28-May-22`) at a particular power station (`bhola225MWCCPP`) within a given area (`1006`). The `EnergyGenerated` property quantifies the energy produced, which is `"1323000"` (likely in MWh). The observation is part of the `EnergyGenerationDataset`, which organizes data into a cuboid structure that includes levels such as `Area`, `Day`, and `PowerStation`, with a measure of `EnergyGenerated`. This structure enables a detailed and precise representation of energy generation data, facilitating both querying and analysis within the dataset.

Linking ElecKG with external knowledge graphs. Linking ElecKG with the external knowledge graphs enhances interoperability with other datasets and increases its reusability. Linking the target TBox concepts involves connecting the conceptual structure such as dimensions, hierarchies, and properties with external knowledge graph concepts. Specific instances of the target ABox level are linked with the external knowledge graph. ElecKG is externally linked with Wikidata [30] and Geonames [1] using OWL property `owl:sameAs`. For example, the triple <onto:Area1006 owl:sameAs geo:1337229, wiki:Q459723.> indicates onto:Area1006 is connected to geo:1337229, and wiki:Q459723. We used OpenRefine [29] to connect ElecKG with external resources.

3.3.3 Load

Following the data's semantic transformation into RDF format, the ElecKG load procedure gets underway. This stage entails putting the RDF data that has been translated into a specified storage system. The data can also be directly imported into a triple store, such as Virtuoso Universal Server, a database, or a data warehouse. For triple stores, the RDF data are loaded using system-specific utilities or processes in the case of Virtuoso, DB.DBA.TTLP(). Alternatively, the RDF data can be persisted locally as a dump file in a format like Turtle, RDF/XML, or N-Triples for safekeeping and access later. This is the phase that makes sure ElecKG has been integrated into a target storage system for query and analysis purposes.

Chapter 4

Experiments and Evaluation

In this regard, the performance of ElecKG will be reviewed on the mentioned specifications: CPU-1.60GHz, up to 1.80 GHz Intel Core i5-8250U; installed RAM: 12.0 GB. In the first place, let's give an overview of the main components of ElecKG and its organizational structure by briefly summarizing it. Then, the performance of ETL in constructing ElecKG regarding efficiency and speed of data transformation and processing will be measured. The compatibility of ElecKG with analytics programs will be examined concerning OLAP query processing and insightful data generation.

4.1 Structural Analysis of the ElecKG Knowledge Graph

This section provides a structural overview of the ElecKG, detailing its dimensions and factual components.

4.1.1 Dimension overview

These include dimensions, levels within each dimension, and related metrics, which are described in detail in Table 4.1 below for a detailed summary. This dataset includes 10 dimensions, 35 level attributes, and 1141 instances resulting in a total of 4996 RDF triples. More significantly, there are 8 integrated external links that enable semantic interlinking of data. This structured representation facilitates seamless data integration into the knowledge network, enabling comprehensive semantic analysis and multidimensional data exploration.

Dimensions	**Levels**	**Quantity of attributes**	**Quantity of instances**	**Quantity of external links**	**Quantity of RDF triples**
mdProperty:NationalGrid	mdProperty:Area	3	9	8	35
	mdProperty:Zone	3	2	0	6
	mdProperty:CountryGrid	4	1	0	4
mdProperty:TimePeriod	mdProperty:Day	5	730	0	3650
	mdProperty:Month	5	24	0	120
	mdProperty:Year	2	3	0	6
mdProperty:PowerPlant	mdProperty:PowerStation	3	179	0	537
	mdProperty:FuelSource	4	9	0	36
mdProperty:GridSubstation	mdProperty:Substation	3	178	0	534
	mdProperty:VoltageLevel	3	6	0	18
Total	**10**	**35**	**1,141**	**766**	**4996**

Table 4.1 Table showing the number of attributes, instances, external links, and RDF triples for each level within the dimensions

4.1.2 Fact overview

The ElecKG facts consist of three main cuboids, each representing distinct aspects of energy data: the EnergyGenerationCuboid, LoadCuboid, and DemandLoadShedCuboid, as shown in Table 4.2. The EnergyGenerationCuboid contains 125717 observations and 754302 RDF triples, occupying 107 MB. The LoadCuboid, significantly smaller at 97.2 MB, features 123,694 observations and 742173 RDF triples. The smallest, the DemandLoadShedCuboid, is 4.72 MB with 6,418 observations and 38,508 RDF triples. Collectively, these cuboids form an ABox of 208.92 MB, with 255,829 observations and 1,534,983 RDF triples, illustrating the dataset's overall size and complexity.

Cuboid	**ABox Volume**	**Quantity of observations**	**Quantity of RDF Triples**
mdStructure:EnergyGenerationCuboid	107 MB	125717	754302
mdStructure:LoadCuboid	97.2 MB	123,694	742173
mdStructure:DemandLoadShedCuboid	4.72 MB	6,418	38,508
Total	208.92 MB	255829	1,534,983

Table 4.2 Summary of the ElecKG cuboid statistics

4.2 Availability

To adhere to the FAIR principles, the ElecKG, an energy data knowledge graph, has been structured to ensure it is both findable and accessible through a variety of interfaces once it has been identified by its unique identifiers. Remote access to the ElecKG is provided by means of a SPARQL endpoint, which is available under http://bike-csecu.com:8896/sparql/. This allows users to conduct far more advanced and flexible queries for knowledge graphs with features of detailed analysis and data retrieval. It is ensured that through remote access, people can make full use of SPARQL in interacting with the data, no matter how far away from where this is kept. For the ones that prefer a more intuitive interface, it is also possible to query the ElecKG using the Virtuoso-powered query interface located at https://bikecsecu.com/datasets/ElecKG/. In that sense, it offers a much more friendly way of querying the knowledge graph, where less technical requirements in terms of SPARQL are needed.In addition to these remote access options, users can interact with ElecKG through the OLAP interface SETBI, which is available at https://github.com/bi-setl/SETL. SETBI is a powerful analytical environment where, based on the energy data, users can perform multi-dimensional analysis and gain insight in a more structured and interactive way. Put together, these resources guarantee that ElecKG will not only be findable through its identifiers but also accessible via several platforms, which meet various user preferences and technical skills. Whether through the powerful SPARQL endpoint, the easy-to-use Virtuoso query interface, or the capabilities for analysis available in SETBI, ElecKG is conceived to support extensive and detailed exploration and analytics of energy data.

4.3 Efficiency of the ETL process

This section focuses on the time required for different stages of the ETL process in creating the ElecKG. The duration and percentage of the total time utilized in each step of the ETL operations are depicted in Table 4.3.

- **Extraction Phase.** The extraction phase entails downloading, merging, and conversion of the PDFs into CSV format and takes up the lion's share of the overall time, 86.96%. First, this phase will take more time because the available electricity data from BPDB is in a read-only format; therefore, it involves much preprocessing, cleansing, and formatting to extract attribute information and get it into tabular format.

- **Transformation Phase.** We use the SETLBI tool for generating TBoxes, source-to-target mapping definitions, and ABoxes. The overall transformation time is around 1.5 hours (8.69%). Notice that TBoxes and the mapping definitions creation require user intervention; hence the time taken can vary depending on the expertise of a particular end.

- **Linking Phase.** We employ OpenRefine to create links between internal and external resources. Similar to previous steps, this too takes longer due to user involvement. The duration of the external linking stage is about 0.5 hours (2.88

- **Loading Phase.** The final phase involves loading the processed data into the Virtuoso triple store. This phase, although brief, is critical for making the data accessible for query and analysis. The loading phase takes approximately 0.25 hours (1.47%).

Phase	Extraction	Transformation	External Linking	Loading	Total
Time (hours)	15.00h	1.50h	0.50h	0.25h	17.25h
Percentage (%)	86.96%	8.69%	2.88%	1.47%	100%

Table 4.3 ETL Process statistics for ElecKG

4.4 FAIRification of ElecKG

The FAIRification process consists of several key steps to enhance the quality and usability of ElecKG, ensuring it adheres to the principles of Findability, Accessibility, Interoperability, and Reusability.

Retrieve Non-FAIR Data: The first step involves obtaining the raw data for ElecKG, which amounts to 1.82 GB of data spread across approximately 2,190 files from the BPDB website available at https://misc.bpdb.gov.bd/daily-generation-archive. Initially in PDF format, this data must be converted and processed for further use.

Examine the Retrieved Data: Upon retrieval, electricity data is analyzed to understand its content and structure. This includes examining the represented concepts, data organization, and relationships between data elements. Understanding parameters such as time, grid, energy type, voltage levels, electricity demand, and load-shedding data is crucial for accurate data representation. How electricity generation, transmission, distribution, and consumption processes work in the national grid is crucial for modeling electricity data.

Describe the Semantic Model: A semantic model is created to define the meaning of data elements and their relationships in ElecKG. This model employs a multidimensional RDF triple structure and integrates standard vocabularies like RDF, RDFS, OWL, and QB4OLAP, ensuring meaningful and actionable data representation.

Make Data Linkable: The electricity data is transformed into a linkable format using the semantic model. Converting electricity data into RDF triples generates the ElecKG and links it to Wikidata and Geonames ontologies.

Assign License: An explicit license is applied to the dataset to make usage terms explicit and hence to stimulate reuse. The detailed licensing information accompanying ElecKG stipulates conditions for accessing the data and its usage while making it openly available without imposing restrictions on data usage.

Describe the Dataset's Metadata: Metadata of this dataset is carefully designed to be FAIR: Findable, Accessible, Interoperable, and Reusable. Metadata describes the semantic structure of the dataset, giving a definition of classes, properties, and relationships defined in the schema TBox. The main information comprised in the ElecKG datasets includes cube, dimension, measure, and hierarchy descriptions. Provenance data record source information, transformations, and mappings taken to ontologies. Usage guides have established the context of use in various fields and make it possible for interoperability to work among a range of platforms. They will help make better analyses of data and hence lead to proper decision-making.

Deploy FAIR Data Resource: The final step involves publishing the FAIRified data along with its license and metadata. ElecKG is accessible through centralized endpoints to ensure easy retrieval. The dataset can be accessed at the following locations:

- **TBox:** https://bike-csecu.com/datasets/ElecKG/tbox.ttl
- **ABox:** https://bike-csecu.com/datasets/ElecKG/abox.ttl

ElecKG is designed in a FAIR way, which is to say it should be easily findable, accessible, interoperable, and reusable. It is hosted in a Virtuoso triple store and queried at the SPARQL endpoint http://bike-csecu.com:8896/. ABox and TBox available at https://bike-csecu.com/datasets/ElecKG/. Standard vocabularies like RDF, RDFS, OWL and QB4OLAP are used to guarantee interoperability and the ability of integration with other datasets. The data set is findable by using unique IRIs and dedicated URLs, while its detailed metadata and usage guidelines contribute to its reusability for further research and applications.

4.5 Quality of ElecKG

The quality of ElecKG, the generated knowledge graph, will be covered in this section. It will address things like granularity, timeliness, and completeness. It will also determine whether ElecKG can be used with OLAP queries and rate its overall accuracy.

4.5.1 Completeness

Completeness, in the context of data and knowledge graphs, measures how much relevant and essential information is contained and correctly represented in a dataset or model. It evaluates how well real-world entities it sets out to model are captured by the dataset. Completeness of Property quantifies the completeness of every single data property without missing values;

Population completeness examines how much every entity, relationship, and constraint is captured in the data schema; while Dataset completeness covers how much each relevant real-world object or category is assessed. Each of these dimensions ensures that the knowledge graph or dataset represents the target domain thoroughly and precisely.

Schema Completeness. Schema completeness for ElecKG addresses the extent to which the data model captures all the information necessary, such as attributes, entities, relationships, and constraints, which are necessary to present the domain of electricity. In ElecKG, TBox is designed as a multidimensional cube with three cuboids: Load, Energy Generation, and Demand Loadshed. The schema includes four dimensions: GridSubstation, PowerPlant, TimePeriod, and NationalGrid. Hierarchies in these dimensions will enable detailed analytics in various dimensions. For example, the GridSubstation dimension includes the SubstationVoltageLevel hierarchy, while the PowerPlant dimension features the PowerStationFuelSource hierarchy. This schema ensures that all relevant aspects of electricity generation, distribution, and demand are comprehensively represented and analyzable.

Property Completeness. ElecKG measures the degree of missing values for a given property by looking at how complete the property is [17]. ElecKG, for example, has issues with property completeness, especially when data about energy generation, demand for electricity, load shedding, and maximum load are lacking. For example:

- The property `mdProperty:Demand`, representing “Electricity demand of the consumer,” has a property completeness score of 99.30%.
- The property `mdProperty:EnergyGenerated`, referring to “Energy generated by power stations,” has a property completeness score of 98.64%.
- The property `mdProperty:LoadShed`, detailing “Mitigates blackout risk via controlled reduction,” also has a property completeness score of 99.30%.
- The property `mdProperty:MaxLoad`, representing “Maximum load served by different substations,” has a property completeness score of 95.96%.

This deficiency, akin to column completeness, affects the integrity of the dataset [23]. The property completeness scores for the specific properties in ElecKG reflect the proportion of data present versus what is missing, impacting the overall completeness of the knowledge graph.The property completeness P for each property is calculated using the following formula 4.1 :

$$P = \left[1 - \frac{\text{Number of incomplete items}}{\text{Total number of items}}\right] \times 100 \tag{4.1}$$

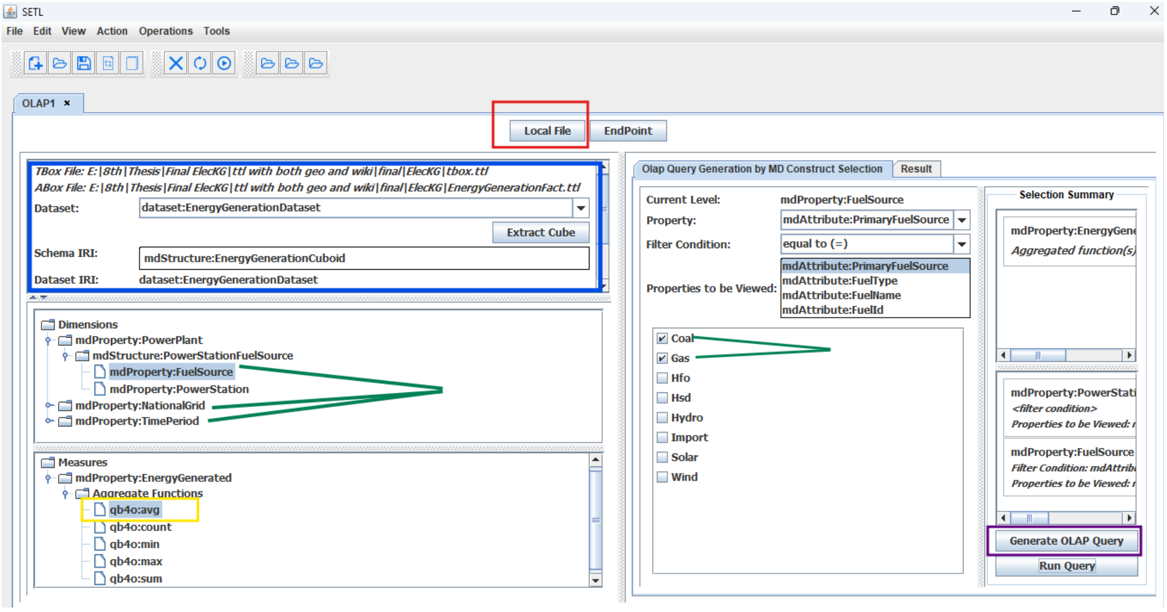
4.5.2 Timeliness and Granularity

The most recent electrical data, spanning the years 2022 to 2024, has been added to the ElecKG. All substation and power plant details are included in this collection, which is arranged geographically. The data's granularity is carefully organized to offer a comprehensive picture of the infrastructure and consumption of electricity.

The dataset offers area-specific information on power demand and load shedding, with an emphasis on regional granularity. Curiously, higher-level aggregations such as the zone and country grid data levels are not included in this data. This is important to ensure the dataset represents accurate localized information, crucial for effective regional analysis and decision-making.

4.6 Ensuring OLAP Compatibility for ElecKG

The Multidimensional TBox of the ElecKG using QB4OLAP is implemented at the SETLBI Definition Layer. ABox of the ElecKG is generated at the ETL Layer of the SETBI KG generation process. This section ensures that the ElecKG is compatible with OLAP queries by loading it in the SETLBI OLAP Layer. The successful execution without errors confirms the knowledge graph's suitability for business analytics.



TBox File: E:|8th|Thesis|Final ElecKG|ttl with both geo and wiki|final|ElecKG|tbox.ttl
ABox File: E:|8th|Thesis|Final ElecKG|ttl with both geo and wiki|final|ElecKG|EnergyGenerationFact.ttl
Dataset: dataset:EnergyGenerationDataset
Schema IRI: mdStructure:EnergyGenerationCuboid
Dataset IRI: dataset:EnergyGenerationDataset

Current Level: mdProperty:FuelSource
Property: mdAttribute:PrimaryFuelSource
Filter Condition: equal to (=)
Properties to be Viewed: mdAttribute:PrimaryFuelSource
mdAttribute:FuelType
mdAttribute:FuelName
mdAttribute:FuelId

☑ Coal
☑ Gas
☐ Hfo
☐ Hsd
☐ Hydro
☐ Import
☐ Solar
☐ Wind

mdProperty:EnergyGene
Aggregated function(s)

mdProperty:PowerStati
<filter condition>
Properties to be Viewed: r

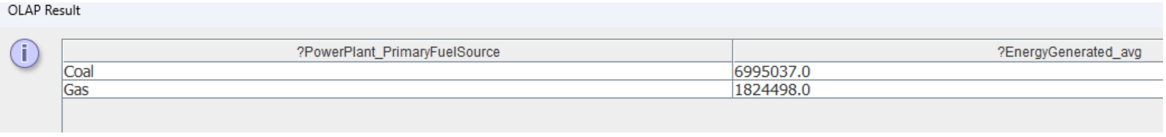
mdProperty:FuelSource
Filter Condition: mdAttribu
Properties to be Viewed: r

Fig. 4.1 Enabling SETLBI's OLAP layer on ElecKG

```
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX qb4o: <http://purl.org/qb4olap/cubes#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT ?PowerPlant_PowerStationName ?PowerPlant_PrimaryFuelSource
       (AVG(xsd:float(?m1)) AS ?EnergyGenerated_avg)
WHERE {
  ?o a qb:Observation .
  ?o qb:dataSet <http://bike-csecu.com/datasets/energy/abox/data#
     EnergyGenerationDataset>.
  ?o <http://bike-csecu.com/datasets/energy/abox/mdProperty#
     EnergyGenerated>?m1 .
  ?o <http://bike-csecu.com/datasets/energy/abox/mdProperty#
     PowerStation>?PowerPlant_PowerStation .
  ?PowerPlant_PowerStation qb4o:memberOf
        <http://bike-csecu.com/datasets/energy/abox/mdProperty#
           PowerStation>.
  ?PowerPlant_PowerStation <http://bike-csecu.com/datasets/energy/
     abox/mdAttribute#PowerStationName>?
     PowerPlant_PowerStationName .
  ?PowerPlant_PowerStation <http://bike-csecu.com/datasets/energy/
     abox/mdAttribute#InFuelSource>?PowerPlant_FuelSource .
  ?PowerPlant_FuelSource qb4o:memberOf
        <http://bike-csecu.com/datasets/energy/abox/mdProperty#
           FuelSource>.
  ?PowerPlant_FuelSource <http://bike-csecu.com/datasets/energy/
     abox/mdAttribute#PrimaryFuelSource>?
     PowerPlant_PrimaryFuelSource .
  FILTER (REGEX(?PowerPlant_PrimaryFuelSource, "Coal", "i")
      || REGEX(?PowerPlant_PrimaryFuelSource, "Gas", "i"))
}
GROUP BY ?PowerPlant_PowerStationName ?
    PowerPlant_PrimaryFuelSource
ORDER BY ?PowerPlant_PowerStationName ?
    PowerPlant_PrimaryFuelSource
```

Listing 4.1 SPARQL query on ElecKG



OLAP Result

?PowerPlant_PrimaryFuelSource	?EnergyGenerated_avg
Coal	6995037.0
Gas	1824498.0

Fig. 4.2 Query results

SETLBI OLAP Layer interface is presented in Figure 4.1 showing successful loading of the ElecKG TBox and ABox(highlighted in the upper-left corner with a blue rectangle). The datasets, dimensions, levels, measures, and aggregation functions can be specified within the visualization panel to formulate queries. By clicking the "Generate OLAP Query" buttons OLAP query of the subsequent selections is generated and by clicking the "Run Query" buttons (highlighted in violet rectangles) the user can execute queries and view the generated SPARQL code.

The generated SPARQL query, as shown in Listing 4.1, retrieves and aggregates data related to the 'mdSTructure:EnergyGenerationCuboid' by selecting 'mdProperty:FuelSource' from the 'mdProperty:PowerPlant' dimension. The result of this SPARQL query, displayed in Figure 4.2, demonstrates the aggregation of energy generation data filtered by fuel source. This confirms ElecKG's OLAP compliance and its capability for multidimensional business analytics.

4.6.1 Queries that ElecKG can answer

This section assesses ElecKG effectively addresses a wide range of queries available at http://surl.li/mwvohm shown in Table 4.4, showcasing its enhanced performance and superior data handling compared to the original dataset. These queries are further categorized into 1) intra-cube (Q1-Q10), 2) inter-cube (Q11-Q13), and 3) federated (Q14) queries. Intra-cube queries: We apply the following OLAP operations such as roll-up, slice, dice, and drill down on a single cuboid. The roll-up operation zooms out (Day → Month → Year) on any dimension of that cuboid in ElecKG. The Drill down does the zooming in (CountryGrid → Zone → Area) to show a finer level of detail. The Slice operation selects the specific level member of a dimension and leaves other dimensions intact. The Dice operation selects two or more level members of several dimensions. Table 4.4 summarizes these queries by showing their types and the accuracy of the answer obtained from ElecKG compared to the original BPDB dataset.

Among the aggregation queries, ElecKG is the best option for aggregation along multiple dimensions in the Roll-up category: Q1-Q3. For instance, Query Q1 tries to find the amount of energy produced by each power station during the defined time period, monthly or yearly. Indeed, the original data set does support this, but it will not be efficient, whereas ElecKG may allow for an optimized version of this. Likewise, Query Q2 explores the overall energy produced from different fuel types-renewable or not within a certain period. The knowledge is missing in the raw data but is easily provided by ElecKG. On the contrary, Q3 explores the overall energy produced by different fuel types, such as gas and coal, by each power station within a certain period, which is poorly covered in the raw data but is delivered precisely by ElecKG.

For Slice queries (Q4-Q6) targeting particular attributes of data, ElecKG performs better. Query Q4 investigates the maximum load measured for a given grid substation, like "Hathazari", over various years and seasons. Also, such information is not supplied by the original dataset but is provided by ElecKG. Query Q5 fetches the average demand and load shedding of a particular grid area, like 'Chittagong', for different seasons. While the original dataset lacks this detail, ElecKG addresses it very effectively. Query Q6 examines the average energy generated, by power plants using specific fuels such as hydro, dissected by plant and season. Again, such a granular analysis is not presented by the original dataset but was provided by ElecKG.

In the Dice category (Q7-Q9), ElecKG performs the best due to its multi-dimensional data analysis. Query Q7 queries the average of maximum load based on different substations for various voltage levels, areas, and seasons from a certain year. The analysis that this query requires cannot be done with the original dataset but is handled effectively by ElecKG. Query Q8: This query estimates the average energy generation produced by power stations based on certain fuels in selected grid areas and seasons. While this is not possible in the original dataset, ElecKG enables this type of information retrieval. Query Q9: It finds the average demand and load shedding electricity within a given year for a specific area-like Dhaka-over different days of the weeks and seasons. While this type of information retrieval is not enabled from the original dataset, ElecKG does so effectively.

For the Drill-down query (Q10), ElecKG has demonstrated excellent performance regarding the acquisition of fine-grained insight, which could not be realized even by the original dataset. This query explores the typical energy generation of the power stations, going deep into a multi-dimensional analysis: primary fuel source, geographic location at country, zone, area, and temporal aspects of year, month, and day. Such detail is beyond the original dataset but flows like water in ElecKG.

Inter-cube queries (Q11-Q13), which involve integration from more than one data cube, further extend the complex capabilities of ElecKG. Query Q11, for example, requests the total energy generated and total demand for each area in the national grid for the year. Such integration is impossible with the original dataset but is aptly handled by ElecKG. Query Q12 asks for the maximum electricity demand and the maximum load in each national grid zone during a specific year, which is not supported by the original dataset but is possible by using ElecKG. Finally, Query Q13 analyzes how the maximum values of energy generation and load change within different areas and times (months and years). This is a comparative study which could not have been possible with the original dataset but is handled quite nicely by ElecKG.

Federated query Q14 uses ElecKG and Wikidata to analyze electricity demand, load shedding, per capita demand, and load shedding-to-demand ratios across regions monthly. It reveals energy impacts not covered by the original dataset, as detailed in Section 4.6.2. The query results can be accessed by running them at the Virtuoso SPARQL endpoint: http://bike-csecu.com:8896/sparql/.

Eventually, ElecKG emerges as the best explorer to enhance the analytical capability of data complexity in its own BPDB dataset. Its ability to manage OLAP operations like Roll-up, Drill-down, Slice, and Dice in the multi-dimensional analysis of electricity data is unparalleled. Performance by ElecKG across intra-cube queries, such as Q1-Q10, shows the nature of structural definitions of the dataset for deeper insights in energy production, fuel types, and geographical dimensions, which were hard or impossible to gain from the dataset in its original form. Its power in handling inter-cube queries, Q11-Q13, further attests to its worth, enabling integration across diverse data cubes and thus facilitating comparative studies of energy demand and generation with load management across diverse regions and time frames.

The Federated queries, such as Q14, extend the functionality of ElecKG by allowing external sources such as Wikidata to be combined for an even larger study, including socio-economic factors such as per capita demand and regional population impact. These advanced queries can be executed by various stakeholders to assess disparities in energy distribution and management across regions concerning both temporal and spatial variations. Besides, ElecKG is capable of operating with big data, making it an extremely useful tool for policymakers and energy planners to optimize grid reliability and reduce the possibility of inefficiency.

ElecKG helps develop focused, evidence-based energy policies to bridge the regional disparity in demand and load shedding by availing more granular data insights. It thus allows better decision-making in resource allocation and infrastructure development toward a balanced and equative energy distribution system. The scalability of the platform, along with the ability to link data from different sources, provides further avenues for future extension and improvements. ElecKG stands as a key asset in achieving sustainable development goals and advanced improvements within Bangladesh's energy sector.

Query Number	Query Type	Target Cuboid	Query Description	Response from Original Dataset (BPDB)	Response from ElecKG	Is ElecKG Accurate?
Q1	Rollup	mdStructure:EnergyGenerationCuboid	What is the total energy generated by each power station over a time period (e.g., month or year)?	Available (Inefficient)	Can Provide Answer	Y
Q2	Rollup	mdStructure:EnergyGenerationCuboid	What is the total energy generated from different fuel sources (renewable or non-renewable) over a time period?	Not Available	Can Provide Answer	Y
Q3	Rollup	mdStructure:EnergyGenerationCuboid	What is the sum of energy generated by each fuel type (e.g., gas, coal) at power stations over a time period?	Available (Inefficient)	Can Provide Answer	Y
Q4	Slice	mdStructure:LoadCuboid	What is the maximum load recorded for a specific grid substation (e.g., "Hathazari") across different years and seasons?	Not Available	Can Provide Answer	Y
Q5	Slice	mdStructure:DemandLoadShedCuboid	What are the average demand and load shedding values for a specific grid area (e.g., 'Chittagong') for each season?	Not Available	Can Provide Answer	Y
Q6	Slice	mdStructure:EnergyGenerationCuboid	What is the average amount of energy generated by power plants using a specific fuel type (e.g., hydro) as their primary fuel source, broken down by power plant and season?	Not Available	Can Provide Answer	Y
Q7	Dice	mdStructure:LoadCuboid	What is the average maximum load at different substations in specific areas, across various voltage levels and seasons, during a particular year?	Not Available	Can Provide Answer	Y
Q8	Dice	mdStructure:EnergyGenerationCuboid	What is the average energy generation by power stations using specific fuels in selected grid areas and seasons?	Not Available	Can Provide Answer	Y
Q9	Dice	mdStructure:DemandLoadShedCuboid	What is the average electricity demand and load shedding for a specific area (e.g., Dhaka) across different days and seasons (e.g., Summer or Winter) in a specific year (e.g., 2023)?	Not Available	Can Provide Answer	Y
Q10	Drill-Down	mdStructure:EnergyGenerationCuboid	What is the average energy generation by power stations when analyzed in detail across different dimensions such as primary fuel source, geographic location (country, zone, area), and time (year, month, day)?	Not Available	Can Provide Answer	Y
Q11	Inter-cube	mdStructure:DemandLoadShedCuboid and mdStructure:EnergyGenerationCuboid	What are the total energy generated and total demand for each area in the national grid for the year?	Not Available	Can Provide Answer	Y
Q12	Inter-cube	mdStructure:LoadCuboid and mdStructure:DemandLoadShedCuboid	What are the maximum electricity demand and maximum load in each national grid zone for a specific year?	Not Available	Can Provide Answer	Y
Q13	Inter-cube	mdStructure:EnergyGenerationCuboid and mdStructure:LoadCuboid	How do the maximum values of energy generation and load vary across different areas and time periods (months and years)?	Not Available	Can Provide Answer	Y
Q14	Federated	mdStructure:DemandLoadShedCuboid and Wikidata	How do electricity demand and load shedding impact different areas and their populations monthly, including metrics like per capita demand and load shedding, and the ratio of load shedding to demand?	Not Available	See Section 4.6.2	Y

Table 4.4 Competency queries and their answers in ElecKG.

4.6.2 Electricity Demand and Load Shedding Trends Across Bangladesh's Divisions

This section will look in detail at the trend analysis of electricity demand and associated load-shedding across Bangladesh from May 2022 to April 2024, as per the federated query (Q14) in Section 4.6.1. The analysis is performed to identify the major trends, regional inequalities, and their implications for socio-economic development and infrastructure. Numerical differences also allow us to extract a lot about structural problems and performance concerning the national power grid. This contextual approach brings major issues to the fore and performance metrics, which in turn affect different divisions, therefore providing a better perspective on the factors affecting energy distribution and load management across the country.

Figure 4.3 presents the average monthly electricity demand and load shedding, showing very distinct seasonal patterns and implications. In the hot months of July and August, demand peaks to about 3,800 MW due to increased air conditioning usage; this is very closely related to energy consumption. The load shedding in peak demand months also goes up quite dramatically, thus averaging about 500 MW. This indicates the wide pressure on the grid during this time. There is a very good relationship between the high demand for electricity and increased load shedding; this may actually indicate that during these months, the grid infrastructure is often stretched beyond its capacity. This trend further brings to light cyclic demand and hence stress on the electricity supply system. These patterns indeed suggest that strategic improvements are eminently needed in grid capacity and management. Improved infrastructure and an active load management strategy during peak periods become necessary to curtail the deleterious effects of load shedding. Once these issues have been done, there will be a greater reliance on and resiliency of the electricity supply that spikes in demand do not equate to service disruptions. The analysis also showcased that peak load management could be enhanced via factors such as demand response programs, modernization of the grid, and energy efficiency initiatives that lessen the frequency and severity of the events of load shedding. The surmounting of these challenges will go a long way toward ensuring the reliability of electricity supplies, thus further helping to boost the overall system performance together with user satisfaction.

Figure 4.4 shows the scatter plot of demand per capita versus load shed per capita for different grid areas. The trend observed from this plot is that generally, the higher the per capita demand of an area, the higher the load shedding per capita. For example, the data for Dhaka shows that the demand per capita is around 450 kWh/year, while the load shed per capita is around 75 kWh/year. This is a significant figure compared with other regions. In contrast, other areas like Sylhet have a per capita demand of about 200 kWh/year, with a

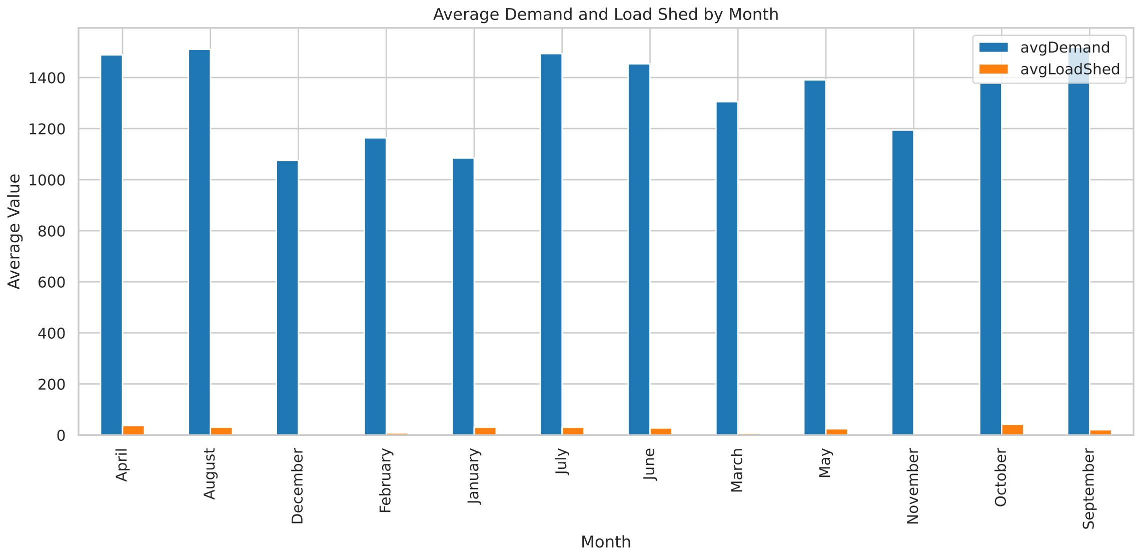


Fig. 4.3 Comparison of average demand and load shedding

corresponding shedding load per capita of only 25 kWh/year. Other areas, such as Chittagong, show intermediate values: the demand per capita is approximately 300 kWh/year, and the shed load is 50 kWh/year per capita. Above the scatter plot shows that high per capita demand is capable of recording increased load shedding per capita. High demands in Dhaka are fairly related to the high load shedding which may indicate that as per capita demand increases, the tendency of load shedding also goes up. This tends to point out the high-energy-use areas face numerous challenges in terms of demand management strategies and, at the same time, improved grid infrastructure. The observed trend underlines addressing energy demands in these high-consumption areas as key to minimizing load shedding. Hence, effective grid management and development are important in every aspect of ensuring a steady, reliable supply of electricity, especially in highly populated areas with high energy consumption per capita.

Figure 4.5 shows the demand-to-load shedding ratio in various grid areas reflecting the efficiency of demand management in each area. This is the total demand for electricity compared to the amount of load shedding that has taken place. The demand over load shedding ratio in Dhaka is about 5.8, which is a low value. A big share of demand is not met and thus shed. Such a low ratio indicates considerable problems with balancing high demand against available supply and may be interpreted as demand management strategies at Dhaka needing further improvement to cope with high electricity consumption and reduce the frequency of load shedding. On the contrary, Khulna has a higher ratio of approximately 8.5. It means that Khulna manages its demand more effectively compared to its load shedding,

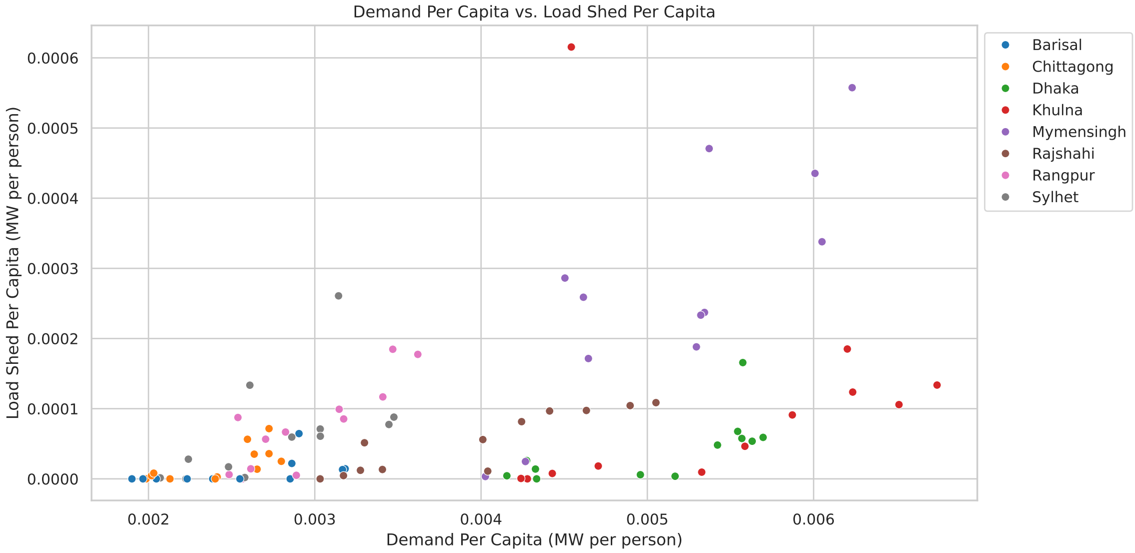


Fig. 4.4 Demand per capita vs. load shed per capita

which is an indicator of better grid performance and efficient management of electricity demand. The higher ratio in Khulna reflects better demand management strategies that could act as a benchmark in similar regions. Overall, this is the ratio that will act as an indicator based on performance for grid efficiency. The higher the ratio, the better, in general, the utility's performance in demand management is than by shedding loads. Strategies may be targeted to those whose ratios are low to enhance demand management at the local level to reduce the impact of shedding load on the supply of electricity.

Figure 4.6 shows in detail the percentage of load shedding concerning the total electricity demand of different grid areas in Bangladesh. The figure indicates that there is immense variation among different areas in experiencing the load shedding percentage. The capital and the premier economic city, Dhaka, faces one of the highest percentages of load shedding at about 17%. This shows that nearly one-fifth of the city's demand for electricity is not met. Again, this pinpoints chronic issues about energy supply in the city and, simultaneously, puts pressure on it to meet growing demands for electricity. This high percentage indicates that systemic problems exist within the distribution network of the capital and demonstrates the urgent need for upgrading its infrastructure or using alternatively effective means of energy management. By contrast, Khulna and Rajshahi have a regular 10% and 12%, respectively, in their load-shedding percentage, representing that these districts suffer from relatively low energy deficits. While this is still considerable load shedding in these areas, the lower percentage figures perhaps suggest that supply and demand factors there might be more balanced, or their power grids could be more efficient in sustaining fluctuations in power

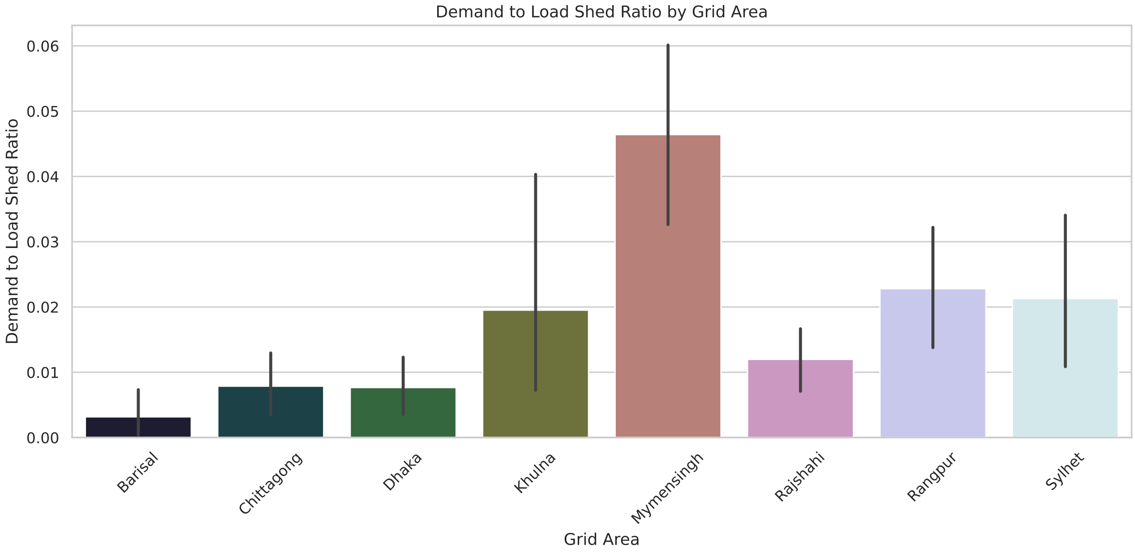


Fig. 4.5 Demand to load shed ratio by grid area

usage. These variations, in effect, raise the imperative need to assess the grid performance on a regional basis, considering that certain areas appear to perform better than others despite challenges in the overall national energy supply. The load shedding percentages in the figure prove the need for targeted interventions, especially in regions with higher demand like Dhaka, where the implications of load shedding are most enormous in daily life and economic activities. Any improvement in the grid performance in these high-demand areas can reduce the energy gap to some extent, thereby contributing to economic growth and an improved quality of life. The ones with lower percentages are not to be trifled with either, as it remains an uphill task to maintain this relatively better performance through constant vigil and pro-action. The above data underlines the all-important need for adopting a region-specific approach to energy management in Bangladesh, where infrastructure capacities are different in differing areas and hence contribute to variations in the load-shedding experiences.

Figure 4.7 presents the gross demand for electricity from various grid areas in Bangladesh starting from May 2022 to April 2024. Variations in consumption patterns are quite evident from the data, where the urban setup demands a quantity of electricity that is many folds higher compared to the rural areas. Dhaka represents the largest consumer, which reaches a peak demand of about 3500 MW. It reflects the role of being the capital and economic center of the country, high population density, and highly concentrated industries, commercial establishments, and residential areas. High demand shows how much stress infrastructure and the continuous need for energy are in keeping the growth in this city. The second consumer,

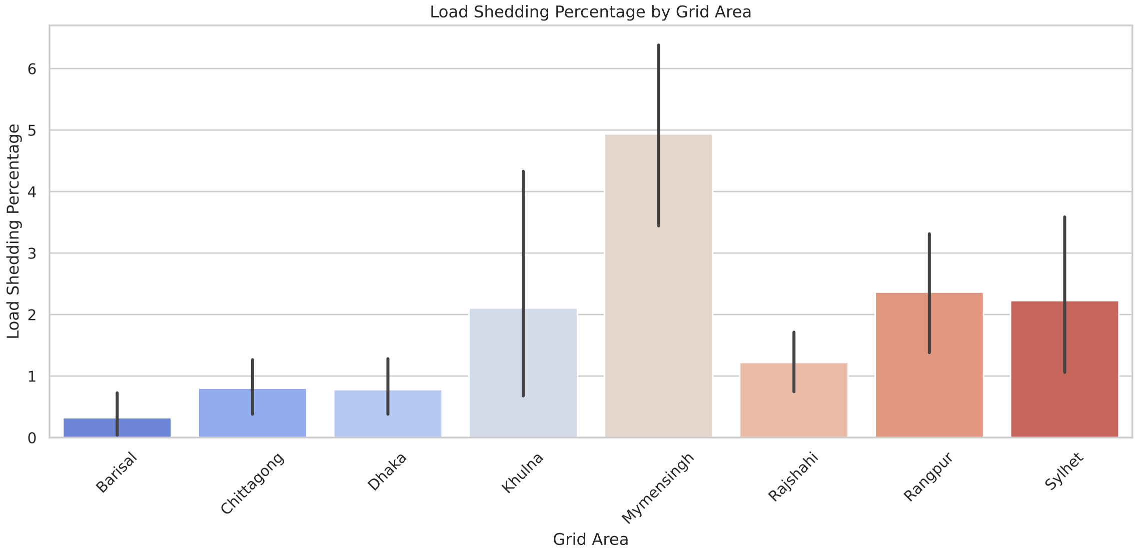


Fig. 4.6 Load shedding percentage by grid area

Chittagong, records a demand of about 1,800 MW. As a major port city and industrial center, the demand for energy here is dominated by manufacturing and shipping activities as well as urban construction. The above comparison of Dhaka and Chittagong highlights the focused utilization of energy in the main economic centers of the country. By contrast, demands for electricity are much lower in Sylhet and Barisal - around 350 MW and 400 MW respectively. This is because these areas are more rural and less industrialized, reflecting a reduced energy consumption. This also contrasts high-demand urban areas with the lower demand in the rural zones and thus infers a clear urban-rural divide in energy utilization. The above discussion thus brings into focus several critical challenges for the energy sector of Bangladesh. This is an urban-centric demand and, therefore, requires sustainable urban energy solutions to increase generation capacity, grid resilience, and efficiency in energy use in high-demand cities such as Dhaka and Chittagong. Simultaneously, however, the lower demands evidenced for Sylhet and Barisal hint that improving access and infrastructure in less developed areas will be necessary for the better distribution of energy resources throughout the country. Going forward with the sustainable development goals, more in line with Vision 2041 of Bangladesh, there is a greater need for the removal of these imbalances in energy consumption. Infrastructure investment and policy initiatives could be utilized to expand grid capacity in rural areas for agricultural and small-scale industrial activities, managing the rapid energy needs of urban centers. Besides, such a transition into renewable sources might be designed in a way to respond differently to high and low-consumption regions. This would therefore involve optimization of the energy mix in such a way that all areas would have electricity reliably and sustainably.

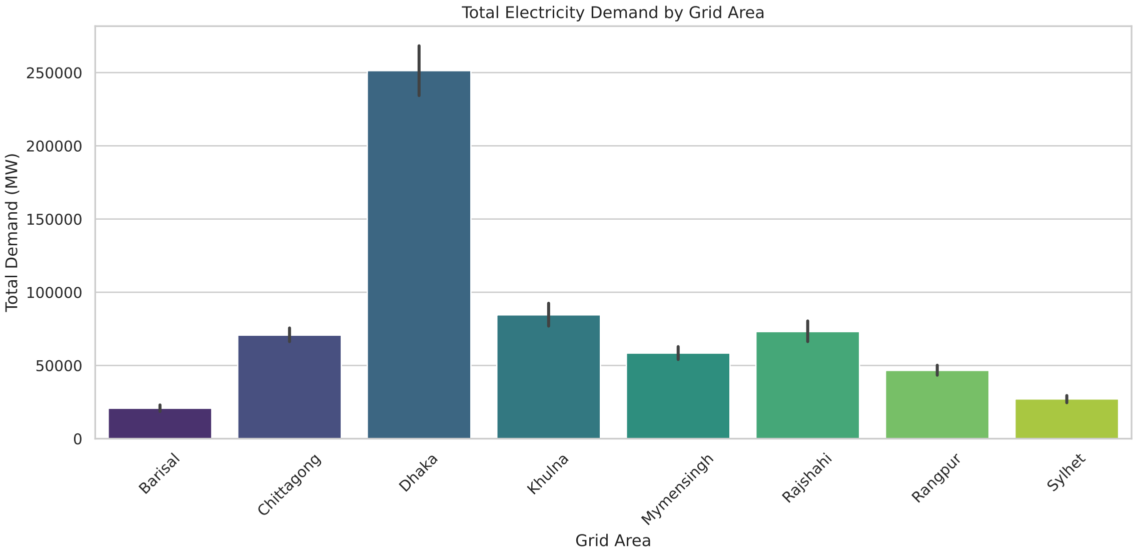


Fig. 4.7 Electricity demand distribution across divisions

Figure 4.8 illustrates in detail the total load shedding across different grid areas in Bangladesh. It highlights those areas where electricity demand could not be fulfilled due to constrained grids. It follows from the figure that Dhaka is in the worst situation when the value of load shedding amounts to almost 600 MW. This huge supply deficit of electricity is due to its heavy population and Dhaka being the economic hub of the country, which creates demand for high consumption of electricity. This acute load shedding disrupts daily life and hampers industrial productivity and economic activities in the region. Another major economic zone, Chittagong, also suffers from substantial load shedding-around 400 MW. This reflects that Chittagong, similar to Dhaka suffers from acute grid management issues, which in case of the former are worsening due to an expanding population and industrial growth. Being a port city, the implicit impact of continuous breakdowns in power would extend beyond the power sector to the national economy as a whole, requiring better infrastructure. In contrast, Rajshahi and Khulna report decidedly low levels of load shedding, at roughly 150 MW and 200 MW, respectively. These areas, though not free from power shortages, seem to manage their electricity distribution more efficiently, either due to lower demand levels or better grid efficiency. However, the lower figures may also show reduced industrial activities compared to Dhaka and Chittagong, which can have a reason for overall energy needs. Sylhet and Barisal, having much less urbanization and industrialization, depict even lower values of load shedding, hence showing regional disparities in electricity demand

and supply. These disparities suggest that, outside of the major metropolises, the power grid infrastructure and management are designed to cope with the demand. The same places a pointer that the focus for grid improvement efforts should be on the cities like Dhaka and Chittagong, where the demand grossly outstrips supply.

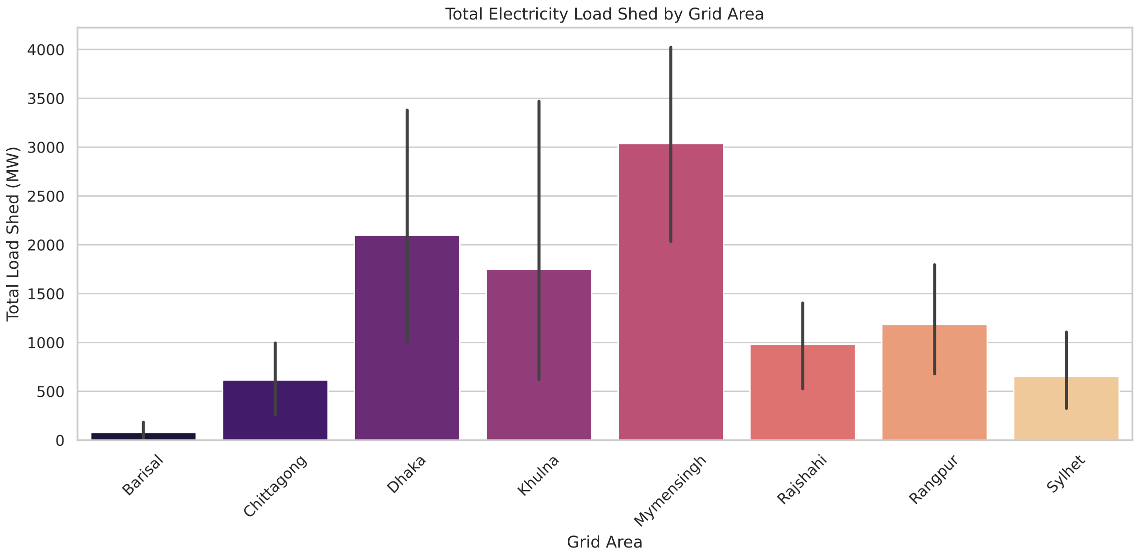


Fig. 4.8 Load shedding across divisions

In summary, the data gives a clear view of electricity demand and load shedding across different regions in Bangladesh over varying months. It shows that regional variation in both demand and load shedding is important. Dhaka has always topped the list in electricity demand and accounted for a sizeable share of load shedding due to its large population and high energy demand. In contrast, despite the relatively low total demands of Barisal and Sylhet, their respective magnitudes of load shedding are dissimilar. This variation tends to affirm the regional factors of influence in energy distribution and load management. The impact of population density on per capita demand and, thereby, load shedding can similarly be portrayed from the presented data, which is greater in Mymensingh and Khulna. These, in turn, result in seasonal variations of demand and load shedding, with months like April recording high energy needs and those like July recording high load shedding. The ratio of demand to load-shedding shows the efficiency of energy distribution and management policies in different areas. In conclusion, all these observations lead to the identification of specific energy policies and infrastructure development that addresses the issue of regional imbalance and further smoothes electricity distribution.

Chapter 5

Conclusion and Future Direction

5.1 Conclusion

For the energy sector in Bangladesh, the creation and application of the ElecKG knowledge graph marks a substantial advancement in the integration and analysis of electrical data. Through the conversion of conventionally non-FAIR, static data from the BPDB into a structured, machine-readable, and interoperable format, we have opened up hitherto unattainable multidimensional and comprehensive analysis capabilities. The data extraction, transformation, linking, and loading (ETL) process ensures that electricity data is correctly recorded and organized in a way that allows for sophisticated queries and informed decision-making.Throughout our efforts, we have demonstrated that ElecKG not only supports a variety of OLAP operations, such as roll-up, slice, dice, and drill-down, but also provides insights into key aspects of the country's energy infrastructure. These include patterns in electricity demand, load shedding studies, and energy generation trends. Importantly, ElecKG overcomes the limitations of traditional data formats by facilitating federated queries that connect to external datasets, such as GeoNames, enabling a more comprehensive understanding of the relationships between population density, geographic factors, electricity demand, and load shedding.

Moreover, the FAIRification of ElecKG adheres to international data management standards, ensuring that the dataset is findable, accessible, interoperable, and reusable. The RDF, OWL, and QB4OLAP vocabulary requirements are, hence, met in ElecKG, which puts it in harmony with other knowledge systems that apply semantic web technologies.Altogether, ElecKG provides the background for further research and practical enhancements related to the energy sector of Bangladesh. This allows new dimensions of detailed insight into energy distribution, demand, and generation that opens further avenues for the optimization of grid management, reducing load shedding, besides furthering the sustainable goals on energy of the nation. As Bangladesh is moving towards Vision 2041, ElecKG will be imperative in inculcating data-driven decision making, energy efficiency, and renewable energy efforts.

5.2 Future Directions

In this respect, further work on ElecKG covers some key extensions which could considerably widen its scope. First, there is the substantial domain of the inclusion of additional external data sources. ElecKG would be able to supply a much more comprehensive insight into the conditions at the basis of the supply and demand of electricity if it integrates relevant data about weather conditions, economic indicators, and specific measures of energy consumption. This will again widen the integration to improve strategic planning and decision-making in the energy sector due to enhancements of forecast accuracy, enabling analysis at more sophistication.

Another exciting path that can be pursued is enhancing the capability of federated queries. Expanding such federated queries into other external knowledge bases positions ElecKG to provide more correct and detailed insights. This will allow for detailed analyses of complex patterns and correlations within the data, therefore giving useful information about how different variables are combined to shape the energy landscape. This will enhance query capability, which, in turn, shall yield more sophisticated analytics and a more accurate, granular understanding of electrical data.

Machine learning and advanced analytics are a key frontier for ElecKG. The application of predictive analytics approaches will facilitate the capabilities of the system in accurately forecasting future electricity consumption, predicting any load-shedding events well in advance, and identifying the right energy generation strategy. These features will be required to improve resource allocation, strengthen grid management, and ensure reliability and efficiency in the energy systems. Advanced data analytics are used by machine learning models to find patterns in historical and real-time data, providing proactive recommendations that result in better outcomes in energy management.

Performance improvements are also needed in the core in order to maintain ElecKG's performance throughout its evolution. Optimizing query performance, as well as data processing efficiency, is essential for the system to remain responsive and capable of bearing higher volumes of data. These improvements will make ElecKG continuously usable and reliable, hence able to cater to the ever-growing demands of the energy sector. Finally, big perspectives open for ElecKG's future development. Advanced analytics, expansion of the sources of data, enhanced federated query capabilities, and performance improvements will further develop ElecKG to become a mighty tool in managing and analyzing electricity data. It is such developments that will go a long way toward achieving the goals of sustainable energy, besides enabling more informed decision-making within the energy sector, as well as enhancing grid management and resource efficiency.

References

[1] Ahlers, D. (2017). Linkage quality analysis of geonames in the semantic web. In *Proceedings of the 11th Workshop on Geographic Information Retrieval*, GIR'17, New York, NY, USA. Association for Computing Machinery.

[2] Arenas, M., Bertails, A., Prud'hommeaux, E., and Sequeda, J. (2012). A Direct Mapping of Relational Data to RDF. http://www.w3.org/TR/2012/REC-rdb-direct-mapping-20120927/. W3C Recommendation.

[3] Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data - the story so far. *Linking the World's Information.*

[4] Board, B. P. D. (2024). Daily generation report. Accessed: 2024-04-30.

[5] Bui, T. D. and Tseng, M.-L. (2022). A data-driven analysis on sustainable energy security: Challenges and opportunities in world regions. *Journal of Global Information Management*, 30:1–38.

[6] Böhms, M., Rieswijk, T., and Lijster, E. (2015). Linked energy data: Enabling monitoring and decision support for improved energy management. In *2015 IEEE International Conference on Engineering, Technology and Innovation/ International Technology Management Conference (ICE/ITMC)*, pages 1–8.

[7] Codd, E. F. (1970). A relational model of data for large shared data banks. *Commun. ACM*, 13(6):377–387.

[8] Cudre-Mauroux, P., Deb Nath, R. P., Romero, O., Pedersen, T. B., and Hose, K. (2022). High-level etl for semantic data warehouses. *Semant. Web*, 13(1):85–132.

[9] Deb Nath, R. P., Hose, K., Pedersen, T. B., Romero, O., and Bhattacharjee, A. (2020). Setlbi: An integrated platform for semantic business intelligence. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 167–171, New York, NY, USA. Association for Computing Machinery.

[10] Decker, S., Melnik, S., van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M., and Horrocks, I. (2000). The semantic web: the roles of xml and rdf. *IEEE Internet Computing*, 4(5):63–73.

[11] Etcheverry, L., Vaisman, A., and Zimányi, E. (2014). Modeling and querying data warehouses on the semantic web using qb4olap. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 45–56. Springer.

[12] Etcheverry, L. and Vaisman, A. A. (2012). Qb4olap: A new vocabulary for olap cubes on the semantic web. In *Proceedings of the Third International Conference on Consuming Linked Data - Volume 905*, COLD'12, page 27–38, Aachen, DEU. CEUR-WS.org.

[GO FAIR] GO FAIR. Fairification process. https://www.go-fair.org/fair-principles/fairification-process/. Accessed: 2023-09-05.

[14] Gomez-Perez, A. and Corcho, O. (2002). Ontology languages for the semantic web. *IEEE Intelligent Systems*, 17(1):54–60.

[15] Government of Bangladesh (2024). Power division, government of bangladesh. Accessed: 2024-09-06.

[16] Harth, A., Hose, K., and Schenkel, R. (2014). *Linked Data Management*. Emerging directions in database systems and applications. Taylor & Francis.

[17] Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2021). Knowledge graphs. *ACM Comput. Surv.*, 54(4).

[18] Jacobsen, A., de Miranda Azevedo, R., Juty, N. S., Batista, D., Coles, S. J., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T. A., Goble, C. A., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K. M., Heringa, J., Hooft, R. W., Imming, M., Jeffery, K. G., Kaliyaperumal, R., Kersloot, M. G., Kirkpatrick, C. R., Kuhn, T., Labastida, I., Magagna, B., McQuilton, P., Meyers, N., Montesanti, A., van Reisen, M., Rocca-Serra, P., Pergl, R., Sansone, S.-A., da Silva Santos, L. O. B., Schneider, J., Strawn, G. O., Thompson, M., Waagmeester, A., Weigel, T., Wilkinson, M. D., Willighagen, E., Wittenburg, P., Roos, M., Mons, B., and Schultes, E. A. (2020). Fair principles: Interpretations and implementation considerations. *Data Intelligence*, 2:10–29.

[19] Jensen, C., Pedersen, T. B., and Thomsen, C. (2010). *Multidimensional databases and data warehousing*. Morgan & Claypool Publishers.

[20] Nations, U. (2024). Sustainable development goal 7: Affordable and clean energy. Accessed: 2024-09-15.

[21] of Bangladesh, G. (2023). Integrated energy and power master plan (iepmp) 2023. Accessed: 2024-09-15.

[22] Penteado, B., Maldonado, J., and Isotani, S. (2022). Methodologies for publishing linked open government data on the web: A systematic mapping and a unified process model. *Semantic Web*, 14:1–26.

[23] Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Commun. ACM*, 45(4):211–218.

[24] Planning Division, Ministry of Planning (2021). Perspective plan (2021-2041). Accessed: 2024-09-06.

[25] Prud'hommeaux, E. and Seaborne, A. (2008). SPARQL Query Language for RDF. http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/. W3C Recommendation.

[26] Slootweg, J., Jordán Córdova, C., Montes Portela, C., and Morren, J. (2011). Smart grids - intelligence for sustainable electrical power systems. In *2011 IEEE 33rd International Telecommunications Energy Conference (INTELEC)*, pages 1–8.

[27] United Nations Framework Convention on Climate Change (UNFCCC) (2021). Bangladesh's nationally determined contribution (ndc) - cop26. Accessed: 2024-09-06.

[28] Vaisman, A. and Zimányi, E. (2014). *Data Warehouse Systems: Design and Implementation*. Data-Centric Systems and Applications (DC). Springer, 1st edition.

[29] Verborgh, R. and Wilde, M. (2013). *Using OpenRefine*.

[30] Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

[31] W3C (2011). Five star linked data. Accessed: 2024-09-06.

[32] W3C RDF Working Group (2014). Rdf 1.1 turtle: Terse rdf triple language. https://www.w3.org/TR/turtle/. W3C Recommendation, 25 February 2014.

[33] Wilkinson, M., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L. O., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., and Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3.