Ref. Ares(2022)2420791 - 31/03/2022

Grant Agreement N° 872592

PLATOON

Digital platform and analytic tools for energy

Deliverable D2.4
The PLATOON Unified Knowledge Base Creation

Contractual delivery date:
M12

Actual delivery date:
31/12/2020

Responsible partner:
P11: TIB, Germany

Project Title	PLATOON – Digital platform and analytic tools for energy
Deliverable number	D2.4
Deliverable title	The PLATOON Unified Knowledge Base Creation
Author(s):	Kemele M. Endris and Maria-Esther Vidal
Responsible Partner:	P11 – TIB
Date:	31.12.2020
Nature	R
Distribution level (CO, PU):	PU
Work package number	WP2 – Reference Architecture, Interoperability and Standardization
Work package leader	ENG, Italy

Abstract:	The PLATOON data sources are characterized not only by big data dominant dimensions like volume and velocity, which impact scalability but also in various formats and suffer from data quality issues affecting data exchange and integration. In the context of the task T2.4, the PLATOON data sources were analyzed to determine data interoperability that may hinder the creation of the PLATOON unified knowledge base and access and processing through the PLATOON architecture. Methodological and technological strategies are presented to assess and overcome the features of the data sources; computational methods to create materialized or virtual knowledge bases are discussed. Moreover, PLATOON pilots are examined to identify interoperability conflicts and the applicability of the discussed computational methods. The W3C standards recommended in the International Data Space (e.g., SHACL, RDF, and RDFS) correspond to building blocks of the PLATOON components that will enable either the materialized and virtualized creation of the PLATOON knowledge base.
Keyword List:	Data Integration, Unified Knowledge Graph, Curation and Integration, Data Sources, Federated Query Processing

The research leading to these results has received funding from the European Community's Horizon 2020 Work Programme (H2020) under grant agreement no 872592.

This report reflects the views only of the authors and does not represent the opinion of the European Commission, and the European Commission is not responsible or liable for any use that may be made of the information contained therein.

Editor(s):	Maria-Esther Vidal (TIB) Kemele M. Endris (TIB)
Contributor(s):	Valentina Janev (IMP)
Reviewer(s):	Philippe Calvez (ENGIE) – Platoon Coordinator Erik Maqueda (TECN) – Technical Coordinator Martino Maggio (ENG) – Work Package Leader
Approved by:	Philippe Calvez (ENGIE) – Platoon Coordinator Erik Maqueda (TECN) – Technical Coordinator
Recommended/mandatory readers:	WP3, WP4, WP5, and WP6

Document Description

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
0.1	03-08-2020	Table of Contents - draft	Kemele M. Endris (TIB)
0.11	04-08-2020	Table of Contents updated	Maria-Esther Vidal (TIB)
0.2	28-09-2020	LLUC2a data source analysis	Kemele M. Endris (TIB)
0.3	10-10-2020	LLUC3b-PI data source analysis and preliminaries added	Kemele M. Endris (TIB)
0.5	01-12-2020	Knowledge graph creation section added	Kemele M. Endris (TIB)
0.6	07-12-2020	Traversing the PLATOON knowledge graph section added	Kemele M. Endris (TIB)
0.61	10-12-2020	Review and update all sections	Maria-Esther Vidal (TIB)
0.62	17-12-2020	Description of pilots added	Kemele M. Endris (TIB)
0.63	17-12-2020	Pilot data source description summary	Maria-Esther Vidal (TIB)
0.64	18-12-2020	Data ecosystem and interoperability added to preliminaries	Maria-Esther Vidal (TIB)
0.65	18-12-2020	Section 3, 4 and 5 reorganized and added more section with examples. Interoperability section added with input from Deliverable D2.3 and pilot descriptions from T6.1	Kemele M. Endris (TIB)
0.66	20-12-2020	Example knowledge base creation process approaches explained using PLATOON data model and example data from D2.3	Kemele M. Endris (TIB)
0.7	22-12-2020	Review all sections.	Valentina Janev (IMP)
0.71	22-12-2020	Reviews all sections	Maria-Esther Vidal (TIB)
0.71	23-12-2020	Review	Erik Maqueda (TECN) and Martino Maggio (ENG)
0.8	23-12-220	Update document related to reviewers' (Erik's and Martino's) comments	Kemele M. Endris (TIB) and Maria-Esther Vidal(TIB)
0.81	23-12-2020	Updated Fig 13 (semantic data model domain) with latest version from D2.3 and related tables	Kemele M. Endris(TIB)

0.81	28-12-2020	Review	Philippe Calvez (ENGIE)
1.0	29-12-2020	Reviews all sections and finalized document	Kemele M. Endris (TIB)
2.0	14.03.2022	New version considering the comments given by the reviewers	Maria-Esther Vidal (TIB) All partners of PLATOON pilots

Table of Contents

List of Figures

List of Tables

Terms and abbreviations

APT	Application Program Interface
CA	Consortium Agreement
CAD	Computer-aided design
CO	Confidential
CSV	Comma Separated Values
DE	Data Ecosystem
DIS	Data Integration System
DM	Dissemination Manager
EC	European Commission
EM	Exploitation Manager
GA	Grant Agreement
GAM	General Assembly Meeting
Gb	Giga Byte
HVAC	Heating, ventilation, and air conditioning
Hz	Hertz (1 observation every second)
IDS	International Data Space
IMP	Institute Mihajlo Pupin
IRI	Internationalized Resource Identifier
JDBC	Java Database Connectivity
JSON	JavaScript Object Notation
KB	Kilo Byte
kHertz	Kilo Hertz (1,000 observations per second)
LLUC	Low-level Use Case
MB	Mega Byte

mHz	Mili Hertz (1 observation every 1,000 seconds)
PI	Poste Italiane
PM	Project Manager
PROV	Provenance Ontology
PU	Public
QA	Quality Assurance
R2RML	Relation to RDF mapping Language
RDB	Relational Database
RDF	Resource Description Framework
RDFS	RDF Schema
RES	Renewable Energy Source
REST	Representational State Transfer
RML	RDF Mapping Language
SDM-RDFizer	Engine to create RDF knowledge bases from DISs whose mappings are specified in RML mapping rules
SHACL	SHApe Constraint Language
SKOS	Simple Knowledge Organization System
SPARQL	An RDF Query Language
Tb	Tera byte
TDMS	Technical Data Management Streaming
W3C	The World Wide Web Consortium
WP	Work package
WPL	Work package Leader
XML	Extensive Markup Language
XSLx	File extension is a Microsoft Excel Open XML Spreadsheet (XLSX) file created by Microsoft Excel.

Executive Summary

This document reports on the outcomes of performing task T2.4 of WP2 - Reference Architecture, Interoperability, and Standardization. It presents the main characteristics of the PLATOON data sources, interoperability problems across them, and the data integration techniques to integrate heterogeneous data sources into the PLATOON unified knowledge base (graph). Data coming from different energy systems need to be integrated into a knowledge base to enable analytical tools and data-driven decision systems to exploit the knowledge represented in Big Energy Data using a unified semantic schema. These data sources include wind power systems, solar power systems, conventional power plants, cooling, heating, and lighting systems as well as smart grids. They represent measurements in different domains, e.g., energy consumption, energy generation, system outages, failures, weather, and energy transmission. These data sources are characterized by the dominant Big Data dimensions, i.e., volume, velocity, variety, veracity, and value; this document reports on the analysis of these sources. Furthermore, interoperability and heterogeneity problems are analyzed; these conflicts are usually caused by the various representations and interpretations of the data ingested from the project data sources. The outcome of this analysis confirms the diverse nature of the PLATOON data sources characterized as Big Energy Data, particularly in terms of volume, velocity, variety, and veracity. These results put in perspective the data complexity issues that need to be tackled in the project. They also state the requirements to be fulfilled during data sharing and integration to scale up large datasets and solve data heterogeneity and quality issues. In this document, the main interoperability issues present in PLATOON data sources, and the knowledge base creation and exploration techniques for different scenarios are described.

1. Introduction

1.1 Purpose and Scope of the Document

This document describes methodological and technical strategies for creating the PLATOON unified knowledge base (graph). The methodological strategies are devised to analyze the PLATOON data sources in terms of big data characteristics and heterogeneity problems. The outcome of the analysis states the requirements to be fulfilled to curate and integrate data into a unified knowledge base. This characterization of the PLATOON data sources has been collected from the PLATOON partners responsible for low-level use cases (LLUCs). A questionnaire characterizing data sources in terms of Big Data dominant dimensions, i.e., volume, velocity, variety, veracity, and value, is presented as part of this deliverable. These questionnaires have been filled by the partners to describe their data sources. Furthermore, the interoperability issues that arise between these datasets are described in detail. Such characterization of data sources provides the basis for creating and managing the PLATOON unified knowledge base using the techniques also described in this document. The interfaces and protocols to access and ingest data to the PLATOON platform and the access to the unified knowledge base are defined in task T2.1, where the PLATOON reference architecture is described. Moreover, the semantic data model for representing energy data from heterogeneous data sources into the unified knowledge base is depicted in task T2.3.

Nine sections compose this document. Section 2 presents preliminaries on concepts; it includes concepts like Big Data dominant dimensions, data spaces, data integration techniques, semantic interoperability issues, and Semantic Web technologies (e.g., RDF, SPARQL, SHACL, and RDF mapping languages). Section 3 sketches a methodology for data source characterization and highlights the questionnaire questions distributed to project partners. Section 4 reports on the overview of pilot and analysis of data sources available for integration to the unified knowledge base. Section 5 reports the interoperability conflicts among PLATOON data sources and description of the concepts in PLATOON semantic data model related to the available data sources in pilots. Section 6 presents the PLATOON data integration platform as a data ecosystem and the unified knowledge base creation pipeline and a description of each component in this pipeline with illustrating examples. Two basic knowledge base creation process scenarios are described in Section 7. In Section 8, presents the technique for exploration of the unified knowledge base. Finally, the conclusions and next steps are outlined in Section 9.

1.2 Relationship with Other Documents

This document is related to two deliverables in WP2: i) D2.1 [1] where the PLATOON reference architecture is defined; and ii) D2.3 [2] where the PLATOON common data models for energy are defined. It is also related to the deliverables of WP5, specifically D5.3, where the methods for data harmonization and knowledge extraction will be implemented, and to the second version of this deliverable which will be submitted on M27.

2. Preliminaries

2.1 Big Data

Structured, semi-structured, and unstructured data is being generated faster than before. Big data systems that integrate different data sources need to handle such characteristics of data efficiently and effectively. Generally, Big Data is defined as data whose volume, acquisition speed, data representation, veracity, and potential value overcome traditional data management systems [3]. Big data is an artifact of individual and collective intelligence generated and collected using technological environments. Virtually every real-world entity can be captured digitally and stored in data sources [4]. Data complexity of Big Data is characterized by data dimensions, usually known as the V's of Big Data [5]. Dominant dimensions of Big Data include Volume, Velocity, Variety, Veracity, and Value. Volume denotes that generation and collection of data are produced at increasingly prominent scales. It refers to the ability to ingest and store very large datasets; they may consist of terabytes, petabytes of data, or even more. This increase of data sets brings new challenges for integrating, managing, and analyzing. Velocity represents that data is rapidly and timely generated and collected, refers to difficulties in ingestion of high rate of data inflow with heterogeneous and evolving structures. Variety indicates heterogeneity in data types, formats, structuredness, and data generation scale. Data may not be consistent, nor does it follow a specific template or format; it is captured in diverse forms and diverse sources, e.g., weather data from sensors and power generation data from the RES plant. These different forms indicate that heterogeneity is a natural property of Big Data, and it is a big challenge to integrate, manage, and analyze such data sources. Veracity denotes noise and quality issues in the data. It refers in part to the biases, ambiguities, and noise in data and about understanding the data, as there are integral discrepancies in almost all the data collected. Thus, the necessity to deal with inaccurate and ambiguous data is another facet of Big Data, which demands to be tackled for ensuring the management and mining of unreliable data. Finally, Value denotes the benefit and usefulness that can be obtained from processing and mining big data. It concerns data quality, including trustworthiness, authenticity, provenance, accountability, and data availability. The challenge is extracting knowledge from vast amounts of structured and unstructured data without loss in their meaning and properties.

2.2 Data Ecosystems

Data ecosystems (DEs) are data-driven infrastructures that allow different stakeholders to exchange data [6, 7]. DEs are furnished with various computational methods to solve interoperability and integrate data, while preserving data privacy, security, and sovereignty. The design of DEs is considered a crucial technological building block for digitalization and the digital economy of the future. DEs aim at being aligned with European Data Strategy and facilitate the creation of data markets for ensuring Europe's competitiveness and data sovereignty. Several research initiatives and industry consortia have followed DEs; they contribute with reference architectures to tackle: (i) data governance according to regulations imposed by data providers; (ii) policies and computational frameworks to ensure a trusted and secure data exchange; (iii) semantic data models for representing main data concepts and relationships, as well as exchange formats and protocols, and (iv) software design principles for guiding the implementation of the components of the reference architectures.

DEs are flexible infrastructures able to fulfill requirements imposed by DE stakeholders. DEs can be centralized, and one single node maintains all the data sources shared by the providers. The node also hosts all the services implemented on top of the DE data sources. Contrary, whenever data cannot be moved to a single node and data privacy regulations hinter the materialized and complete data integration of the DE data sources, DEs will be decentralized, i.e., they will be composed of several nodes. Each DE node will be able to perform services and share data management and analytical results.

Data interoperability is a barrier in DEs; thus, semantic data models or ontologies describing the meaning of the data sources are also part of a DE. Moreover, mapping rules relating to how data sources are defined in terms of the semantic data models are included. Lastly, a DE can also be enhanced with a meta-layer that describes business models, data access regulations, and data exchange contracts. Figure 1 is taken from the Dagstuhl Seminar report on "Data Ecosystems: Sovereign Data Exchange among Organizations" (Figure 5, page 81).

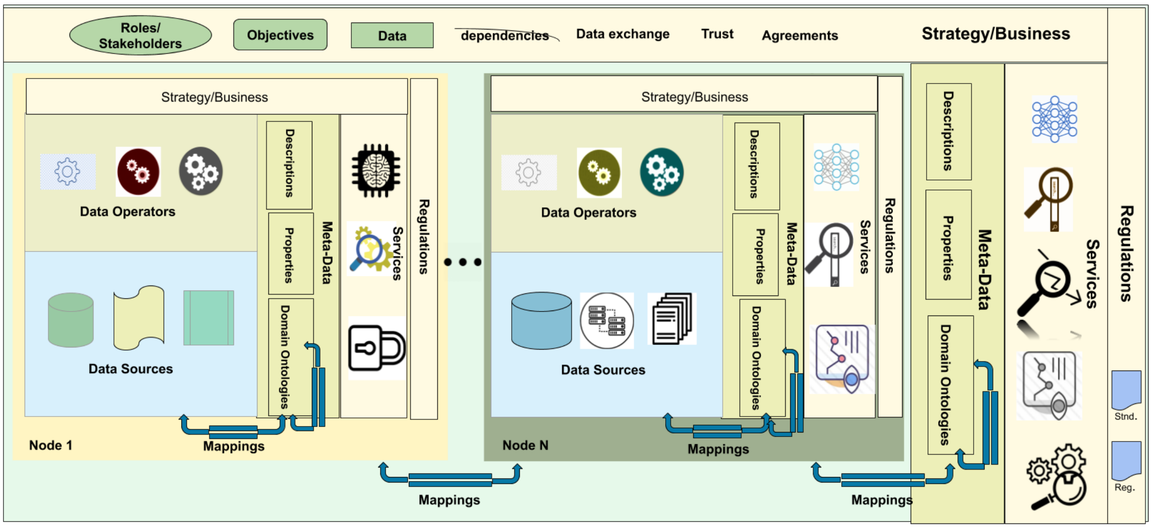

Figure 1: A Network of Data Ecosystems (taken from the report of the Dagstuhl Seminar on "Data Ecosystems: Sovereign Data Exchange among Organizations" [7]). A DE can be distributed and composed of several DEs. Each DE is defined in terms of data sources, data operator s, meta-data, services, business models, and regulations for data exchange.

The International Data Space (IDS) [8] exemplifies DEs; it proposes various standards, technologies, and governance models to facilitate secure and standardized data exchange and integration. Moreover, IDS provides building blocks for the development of data-driven services, while data sovereignty for data providers is guaranteed. IDS propose a message-based infrastructure to enable the communication of the different nodes and components in a DE. Moreover, IDS resorts to the Semantic Web standards to express the content and meaning of the shared data source. The Resource Description Framework (RDF) and ontologies defined using RDF is proposed to specify meta-data, and data control and protection in a decentralized or federated DE. The IDS shared information model states standards for representing Content, Concept, Community of Trust, Commodity, and Communication. Proposed W3C standards include SHACL are proposed to express content and integrity constraints; SKOS for modeling concepts and relationships; and PROV for representing data and service provenance. Figure 2 (taken from the article by Bader et al. [8]) depicts the different W3C standards recommended for semantically describing exchanged data.
This document builds upon the results presented on DEs and proposes a methodology to describe data sources regarding big data characteristics and interoperability issues. Moreover, data integration systems and data lakes are presented as frameworks for supporting data exchange and integration in DEs. Lastly, the main features of W3C standards RDF, RDFS, OWL, SHACL, and SPARQL are illustrated in the context of PLATOON pilots.

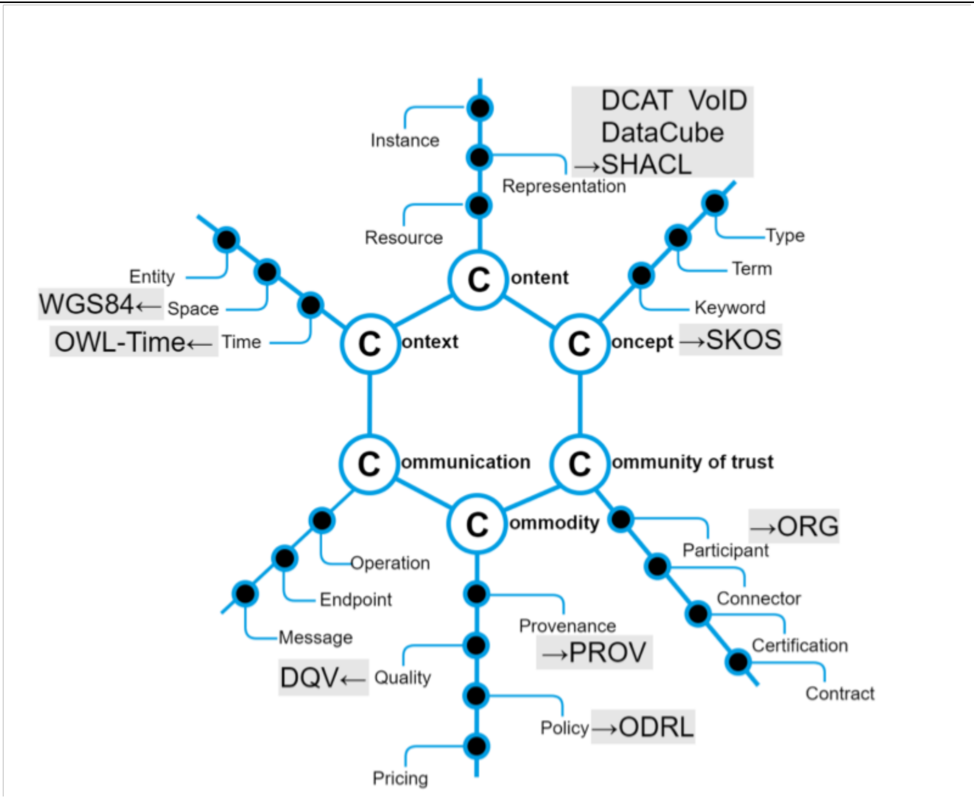

Figure 2: Proposed W3C standards to express meaning and content in International Data Spaces. Figure taken from Bader et. al. [8] : Standards like SHACL, SKOS, and PROV provide a unified way to describe DEs in terms of content, concepts, and provenance.

2.3 Data Interoperability

Due to the technological advances in data collection, generation, and storage in the last decade, different data sources are being generated and stored within enterprises. Various departments within an enterprise generate these datasets independently of each other. As a result, they have different data representation formats, coverage of the domain they represent, and different access methods and permissions. Such properties of the data sources hinder the usage of knowledge available in them. *Data interoperability* is defined as the process of providing uniform access to a set of distributed (or decentralized), autonomous, and heterogeneous data sources [9]. Data interoperability architectures include data integration systems and data lakes.

A data integration system (DIS) integrates two or more datasets. DISs provide a global schema (also known as mediated schema) to provide a reconciled view of all data available in different data sources it integrates. Mappings between the global schema and source schema should be established to combine data residing in data sources considered in the integration process. DISs can be executed to produce a materialized version of a data warehouse of the integrated data sources. The SDM-RDFizer [10] is a computational framework that enables the creation of materialized DISs; mapping rules are expressed in RDF Mapping Rule Language (RML) the generated data warehouses correspond to knowledge bases expressed in RDF. Besides, following the IDS reference architecture recommendations, SHACL can be utilized to specify data content and integrity constraints. Moreover, semantic data models defined in the context of task T2.3 correspond to the DIS global schema.

Data Lakes are data interoperability infrastructures that provide scalable and flexible data discovery, analysis, and reporting; data lakes provide a central repository for raw data. This data can be made available to the users immediately and defer any aggregation or transformation tasks to the data analysis phase. Thus, data lakes address disconnected information silos by storing non-integrated heterogeneous data sources with diverse schemas and query languages. Such a central repository may include different data management systems, such as distributed file systems, relational database management systems, graph data management systems, and triple stores for specialized data model and storage, i.e., preserving the rawness of data and constraints represented in them. Data lakes guarantee a standard access interface to available data for processing and analysis tasks without conveying the development costs of pre-processing and transformations. In addition to raw data, metadata describing the data sources can also be extracted during the ingestion phase. Metadata governance plays a vital role in Data Lakes to efficiently discover datasets and avoid data swamps. Ontario [11] is presented as an engine that enables virtual DISs from data lakes. Ontario builds on top of DISs as shown in Figure 1 where DE nodes comprise organization Data Lakes of data sources kept in raw formats (e.g., XML, CSV, RDB, and JSON), and the semantic data models correspond to the global schema. Virtual knowledge bases are created on the fly from the execution of SPARQL posed against a DE that interlink DEs of organization data lakes. SHACL rules can also describe the content and integrity constraints of data sources in data lakes; they can be validated during the on-the-fly creation of a virtual knowledge base.

Materialized DISs are appropriate for historical data, and whenever data sources can be fully integrated without violating data exchange and access regulation. A knowledge base is virtually created for data that changes frequently or whenever data protection laws hinter the creation of a unified and materialized integration of the data sources. This document reports on the analysis of the PLATOON pilots and discusses interoperability issues present in data sources available in the project pilots. Moreover, the convenience of developing a materialized, or virtual knowledge base for achieving each pilot objectives is discussed.

2.4 Interoperability Conflicts across Heterogeneous Data Sources

Semantic integration of big data entails data variety by enabling the resolution of several interoperability conflicts, e.g., structuredness, schematic, representation, completeness, domain, granularity, and entity matching conflicts. These conflicts arise because data sources may have different data models, follow various data representation schemes, and contain complementary information. Furthermore, a real-world entity may be represented using multiple properties or at multiple levels of detail. Thus, data integration techniques able to solve such interoperability issues while addressing data complexity challenges imposed by big data characteristics are demanded. To be able to integrate these sources in a unified way, semantic interoperability conflicts need to be identified [12]. In this document, interoperability conflicts are characterized in the following six categories:

Structuredness (C1): this interoperability conflict occurs whenever data sources are described at different levels of structuredness, e.g., structured, semi-structured, and unstructured. Structured data sources are represented using schemas of a particular data/knowledge model, e.g., the relational data model; all the represented entities are described in terms of fixed schema and attributes. Semi-structured data sources are also described using a data/knowledge model, e.g., the Resource Description Framework (RDF) or XML; however, in contrast to structured data, each modeled entity can be represented using different attributes and a predefined and fixed schema is not required to describe an entity. Finally, unstructured data sources represent data without following any structured or using a data model; typically, data is presented in various formats, e.g., textual, numerical, images, videos, or other files.

Schematic (C2): this interoperability conflict exists among data sources that are modeled using different schema. Conflicts include: i) different attributes representing the same concept in different sources; ii) the same concept modeled using different structures in the distinct data sources, e.g., attributes versus classes; iii) different types are used to represent the same concept, e.g., string versus integer; iv) the same concept is described at different levels of specialization/generalization; v) different names are used to model the same concept; and vi) different ontologies are used to annotate the same entity.

Domain (C3): this interoperability conflict occurs when various interpretations of the same domain are represented. Different interpretations include: i) Homonym: the same name is used to represent concepts with different meaning; ii) Synonym: distinct names are used to model the same concept; iii) Acronym: different abbreviations for the same concept; iv) Semantic constraint: different integrity constraints are used to model the characteristics of a concept.

Representation (C4): this interoperability conflict refers to different representations used to model the same concept. Representation conflicts include: i) different scales or units; ii) various values of precision; iii) incorrect spellings; iv) different criteria for identifiers; and v) various methods for encode values or representing the encoding.

Language (C5): this interoperability conflict occurs whenever different languages are used to represent the data or metadata (i.e., schema).

Granularity (C6): this interoperability conflict refers to the level of granularity used to collect and represent the data. Examples of granularity conflicts include: i) Samples of the same measurement observed at different time frequency; ii) various criteria of aggregation; and iii) data modeled at various levels of detail.

2.5 RDF and SPARQL

The Semantic Web provides formalism for representing and accessing data that are translated to a set of standards and technologies used to create data stores, vocabularies, and write rules for handling data. At the core of these standards is the Resource Description Framework (RDF) and its associated schema languages, RDF Schema (RDFS) and the Web Ontology Language (OWL). The Resource Description Framework (RDF) is a graph-based data model representing information on the Web. The RDF data model allows expressing information in the form of three element tuples, called RDF *triples*. An RDF triple consists of a *subject*, *predicate*, and an *object*. A *subject* of an RDF triple denotes a resource or entity that is being described, *a predicate* specifies a property or binary relation that associates the subject with the object of the triple and an *object* of a triple denotes a value of the predicate. A set of RDF triples is called an RDF graph, and a collection of RDF graphs form an RDF dataset. Nodes in an RDF graph can be resources or literals, and RDF resources are identified by IRIs (Internationalized Resource Identifier) or blank nodes (anonymous resources or existential variables). Literals can be enriched with data types (defined by XML Schema) and language tags in conformance with the RDF specification. RDF resources can be served via native web access interfaces such as dereferencing resource identifiers, and SPARQL endpoint via the SPARQL protocol.

SPARQL query language is a W3C Recommendation for querying RDF data. SPARQL is basically a graph pattern matching query language, as RDF is a directed graph data model. SPARQL queries can be seen having three parts; *pattern matching*, *solution modifiers*, and *output type*. The pattern matching part includes several features of pattern matching of graphs, such as JOIN, OPTIONAL, UNION, nesting, and FILTER parts. The “solution modifiers” part allows for changing the values computed by the pattern matching part by applying operators such as projection, DISTINCT, GROUP BY, ORDER BY, and LIMIT. Finally, the output type part can be ASK (yes/no), SELECT (selections of values of the variables matching the patterns), CONSTRUCT (construction of new RDF data from these values), and DESCRIBE (descriptions of resources). An evaluation of a SPARQL query Q over an RDF graph G, corresponds to the set instantiations of the variables in the SELECT clause of Q against RDF triples in G. The basic building block in the body of SPARQL query (i.e., in WHERE clause) is the triple pattern, or a triple with variables. A Basic Graph Pattern (BGP) is the conjunction of triple patterns, where a conjunction corresponds to the JOIN operator. Finally, BGPs can be connected with the JOIN, UNION, or OPTIONAL operators.

Let us consider the following question expressed in SPARQL: "*A list of power plants that have generated less than half of their capacity?*" (shown in Figure 3 below).

```
SELECT ?plant ?date ?measure ?gen_amount
WHERE {
     ?plant        a             cim:Plant .
     ?plant        cim:measure   ?measure .
     ?generation   a             cim:Production .
     ?generation   cim:plant     ?plant .
     ?generation   cim:measure   ?gen_measure .
     ?generation   plt:date      ?date .

     FILTER ( ?gen_amount < ?measure/2)
 }
```

Figure 3: SPARQL query expressing question 'A list of power plants that have generated less than half of their capacity

The query is composed of six (6) triple patterns; each triple pattern, e.g., "?plant cim:measure ?measure", corresponds to a subject, ?plant, predicate, cim:measure, and object ?measure. Values that start with a question mark (?) correspond to a variable, e.g., ?plant and ?measure. Each triple pattern is connected via the JOIN operator ("."). In addition to the triple patterns, the above query also includes a FILTER clause that filters values of the generation amount, ?gen_amount, to half of the generation capacity, i.e., ?measure/2.

2.6 Rule based RDF Data Mapping Language

Mapping languages defined by the Semantic Web community can be used to transform non-RDF data sources to RDF. The rules represent mappings that define the concepts of ontology in terms of heterogeneous data sources. Such transformation can also be used to transform legacy databases, data streams, as well as semi-structured data sources published on the Web. R2RML [13], RML [14], xR2RML [15], and SPARQL-Generate [16] are exemplar rule-based languages that are widely used for these tasks.

R2RML is a W3C Recommendation for transformation of relational databases to RDF. R2RML is a language for expressing customized mappings from relational databases to RDF datasets. Such mappings provide the ability to view existing relational data in the RDF data model, expressed in a structure and target vocabulary of the mapping author's choice. An R2RML mapping is represented as a Triple Map, a rule that maps each row in the logical table to a number of RDF triples.

RDF Mapping Language (RML) extends R2RML by generalizing to heterogeneous data sources. RML is a generic mapping language defined for expressing customized mappings from heterogeneous data sources, e.g., RDB, CSV, XML, JSON, to the RDF data model. Each mapping rule in RML is represented as a Triple Map which consists of one logical source, one subject map and zero or more predicate-object maps.

SPARQL-Generate is an expressive template-based language to generate RDF streams or text streams from RDF datasets and document streams in arbitrary formats; it extends SPARQL 1.1 leveraging Aggregates, Solution Sequences and Modifiers, SPARQL functions and their extension mechanism.

2.7 SHACL Constraint Language

The Shapes Constraint Language [17] (SHACL) is the W3C recommendation language to express integrity constraints over RDF graphs. SHACL defines rules over the attributes (i.e., DataTypeProperties) of RDF classes and entities using shapes. Moreover, SHACL enables the definition of constraints among relationships among types (i.e., ObjectProperties). These inter-class constraints induce a shape network used to validate the integrity and data quality properties of an RDF graph. The evaluation results of a shape network over an RDF graph are presented in validation reports using a controlled vocabulary. A validation report includes explanations about the violations, the severity of the violation, and a message describing the violation. SHACL is the language selected by the International Data Space to express the restrictions that state the integrity of an RDF graph [8]. Besides the integrity validation of an RDF graph, SHACL can describe data sources and the certification of a query answer.

2.8 Federated Query Processing

A federated query processing system provides a unified access interface to a set of autonomous, distributed, and heterogeneous data sources. While distributed query processing systems have control over each dataset, federated query processing engines have no control over datasets in the federation, and data providers can join or leave the federation at any time and modify their datasets independently. Query Processing in the context of data sources in a federation is more difficult than centralized systems because of the different parameters involved that affect the query processing engine's performance. Data sources in a federation might contain data fragments about an entity, have different processing capabilities, and support different access patterns, access methods, and operators. The role of federated query processing engines is to transform a query, i.e., the federated query, expressed in terms of the global schema into an equivalent query expressed in the schema of the data sources, i.e., the local query. The local query represents the federated query's actual execution plan by the federation's data sources. An essential part of query processing in the context of federated data sources is query optimization. Since many execution plans are correct transformations of the same federated query, the one that optimizes (minimize) resource consumption should be retained. The performance of query processors can be measured by the total cost that will be used in processing the query and the response time of the query, i.e., the time elapsed for executing the query.

3. Analyzing the PLATOON Data Sources

In this section, the methodology used to analyze the PLATOON data sources from each pilot is described. First the methodology for determining the interoperability issues within datasets is presented, then the template of the questionnaire used to collect the description of the data sources is described. The application of this methodology is presented in Section 4.

3.1 Methodology for Determining Data Interoperability in PLATOON

A methodology to analyze the main characteristics of the PLATOON data sources and interoperability issues has been followed. This methodology is iteratively applied because some sources may change over time. The integration methodology is composed of the following steps; Figure 4 depicts the steps of the methodology.

Figure 4: Methodology followed to describe the PLATOON data sources

1. **Description of the Vs of Big Data** of the PLATOON data sources. A questionnaire is shared with each of the partners of the project who are data providers. These questionnaires allow for determining the main characteristics of the data sources.
2. **Identification of the main concepts** represented by each PLATOON data source. The classes and relationships that compose the semantic data models defined in task T2.3 are utilized to identify the sources that will provide the data to populate the corresponding classes and relationships in the PLATOON knowledge base.
3. **Identification of the interoperability conflicts (C1-C6)**. Data sources that provide data for a given concept of relationship are compared to identify heterogeneity in how data is structured or presented, the level of granularity used in the data set, the language in which data is recorded, and the data meaning. The detected interoperability issues represent the input to task T5.3 where the techniques for data harmonization and integration will be implemented.
4. **Metadata describing PLATOON data sources**. Mapping languages like the RDF Mapping Language (RML) are used to describe the PLATOON data sources in terms of the semantic data models defined in task T2.3. The resulting mapping rules define classes and relationships in terms of the sources based on the outcomes of step 2. Furthermore, these mapping rules solve interoperability issues uncovered in step 3. These mapping rules will be defined by knowledge engineers in task T5.3 in collaboration of the data providers of pilot.

3.2 Questionnaires for Describing the PLATOON Data Sources

Questionnaires were defined in the context of task T2.4 to facilitate the description of the PLATOON data sources. The aim was to collect these specifications from the leaders of each pilot. A questionnaire is composed of the following five parts:

- **Overview:** allows to collect a general description of the data set.
- **Big data Vs**: the data set is described in terms of the dimensions big data models volume, velocity, variety, veracity, and value.

- **Data provider**: captures the protocols followed to access the data, who is the data owner and administrator, and permission status.
- **Data set detailed features**: outlines the main characteristics of the data in the source. These features include: data formats; language; assumptions and standards followed during data collection and harvesting; ontologies and vocabularies used to describe the data; accessibility, permissions, and anonymization, and data collection frequency.
- **Use cases**: presents the use cases where the described data set can be utilized and the coverage of the data set.

In total, the questionnaire comprises 30 questions and the partners need to fill in each questionnaire per data set that will be utilized in the pilot. Table 1 summarizes the main parts of the questionnaire. The pilot developments are still ongoing and some of them have not been described yet at the level required to complete this questionnaire. However, by the day this deliverable is submitted, five (5) PLATOON partners have filled in the description of the data sources based on these 30 questions. Appendix B presents the collected questionnaires for the pilots: 1a (VUB), 2a (IMP), 2b (SAMPOL), and 3b (PI and ROM). During the year 2021, workshops will be conducted to collect complete descriptions of all the PLATOON pilots.

Questionnaire Section	**Question Description**
Overview	Data source title
	Data source acronym
	Data set general description
	Temporal coverage
	Status/ Maintenance
Big Data Vs	Volume- Data size
	Velocity- Frequency of the observations (Longitudinal data)
	Variety- Various formats (CSV, JSON, XML, RDB)
	Veracity - Type of quality problems
	Value - Key Performance Indicators
Data Provider	Name of data provider
	Data provider URI
	Protocol Used to Access Data
	Experimental Strategy

	Data Owner
	Data administrator
	Permission status
Detailed Description	Data format
	Data language
	Data collection assumptions
	Standards
	Ontologies and Vocabularies
	Accessibility, Permissions, Anonymization
	Data collection frequency
	Data schema documentation
	Raw data sample
Use Case	Application scenario
	Possible scenario coverage

Table 1: Questionnaire for the description of the PLATOON data sources

Once the answers of the questionnaires are collected from the PLATOON partners, they are utilized to uncover interoperability conflicts present in the data sources. The outcome of the analysis feeds the steps 2, 3, and 4 of the methodology reported in Figure 4.

4. The PLATOON Data Sources

The PLATOON data sources to be used to create the PLATOON unified knowledge base will be provided by pilot owners of the PLATOON consortium. Additionally, external providers that serve data of a specific domain, such as meteorological data, will also be considered in the unified knowledge base. The PLATOON pilots are described in terms of their datasets; the 5Vs Big Data model describes the main characteristics of the PLATOON data sources.

4.1 PLATOON Pilot overview and available Data Sources

The PLATOON pilots, low-level use cases, and available data sources are discussed below.

Pilot 1a – Predictive Maintenance for Wind Farms

Pilot 1a focuses on offshore and onshore wind turbines equipped with a doubly fed induction generator; it aims at predicting maintenance of wind turbine electrical drivetrain components, i.e., generator and power converter. Pilot 1a will develop, implement, and validate accurate physical and data-driven models of the wind turbine electrical drivetrain components. Additionally, anomaly detection methods will be defined for identifying the unhealthy behavior of the components in scope. Further, an approach to convert the identified anomalies towards health metrics to create a diagnostic tool will be implemented. Lastly, the pilot aims at extracting the relevant events that the electrical drivetrain components are exposed to and have a potentially negative effect on the lifetime of the electrical components. The pilot makes use of two primary sources of data:

- **La Haute-Lys dataset** consists of data from a single, Onshore, General Electric 1.5 MW turbine (machine) placed at the La Haute-Lys wind farm in France. The dataset generated from this turbine focuses on high-frequency (500Hz) measurements of the sensors necessary to gain insights into the electric response/behavior of the wind turbine. This data source is useful for validating the physical models or for data-driven models to capture healthy behavior.

- **ENGIE fleet dataset** consists of data from numerous turbines located in different wind farms. The focus is on Supervisory Control and Data Acquisition system data (SCADA data) sampled at 10-minute intervals. Unlike the *La Haute-Lys* dataset, *the ENGIE fleet dataset* includes turbines with more sensor types for measuring temperature signals and sensors for measuring wind speed, wind direction, generator speeds, torque, etc. Furthermore, this dataset contains fault logs with different fault scenarios, e.g., a short circuit in generator winding. One use case is identified in this pilot: LLUC 1a-01 - Failure detection using a combined data-driven and physics-based model. The ENGIE fleet dataset is an extension of the ENGIE La Haute Borne open dataset. In the DoA, this dataset was described as two separate datasets. However, given that all project partners of Pilot 1a have access to the extended dataset, we opted to use the extended dataset in all specification documents.

- **Open wind speed dataset** consists of wind measurements distributed along the Belgian North Sea. Sensor data includes wind speed and wind direction. This dataset is used in LLUC 1a-01 to assess the typical ranges of wind speeds and directions that can occur in the field. These are used as basis of understanding for defining semantic labels describing wind conditions.

- **Offshore measurement campaign data** consists of acceleration measurements collected on an offshore wind turbine drivetrain. These measurements were in the end not used in Pilot 1a, given that a new dedicated measurement campaign is conducted during the project targeting current measurements that are more appropriate for the analytics methods developed in Pilot 1a.

- **Dedicated current measurement campaign data** consists of current signals that are acquired on an onshore wind turbine. These data are similar to the La Haute Lys dataset. As such they will be merged in further discussions on data handling and analytics with the La Haute Lys data as the same processing methodology applies.

The behavior of electrical components is predicted following various modeling approaches based on anomaly detection schemes. First, integrated digital twins combined with physical-based mode are able to learn links between the potential causes and the failures. Data-driven normal behavior models are also considered. Semantic-based reasoning is utilized to diagnose faults in terms of anomalies and root cause analysis. These predictive approaches will require the pilot data description using the semantic data models defined in task T2.3. Batch processing will be done at the cloud level, while edge computing will be used to perform processing of data in motion from the generator for signature extraction. Figure 5 depicts the structure of the data lake that will be supported at the ENGIE. Services of data cleaning, aggregation, and integration will be implemented on top of the pilot data lake. As a result, three datasets will be extracted. Dataset 1 will comprise high-frequency measurements of one turbine, while dataset 2 is composed of low-frequency data (SCADA) for multiple turbines; both datasets will be utilized for predicting healthy conditions. Moreover, dataset 3 -also composed of low-frequency data (SCADA) for numerous turbines- will be consumed for forecasting faulty conditions. Pilot 1a means for not only predicting turbine maintenance of wind farms effectively but also efficiently. Since both *La Haute-Lys* and *ENGIE fleet* datasets are characterized by the dominant dimensions of big data (i.e., volume, velocity, and variety), scalable data management techniques to curate, aggregate, and integrate these data sources demand to be developed. For this pilot, VUB and TECN are responsible for developing and validating Data Analytics Tools based on both datasets from ENGIE. Thus, these datasets will be exchanged between ENGIE, Vrije Universiteit Brussel (VUB) and Tecnalia (TECN). Furthermore, processed data by TECN will be exchanged to VUB. Hence, data representation and integration using the common semantic data model defined in T2.3 will facilitate the data exchange ensuring the interoperability between them.

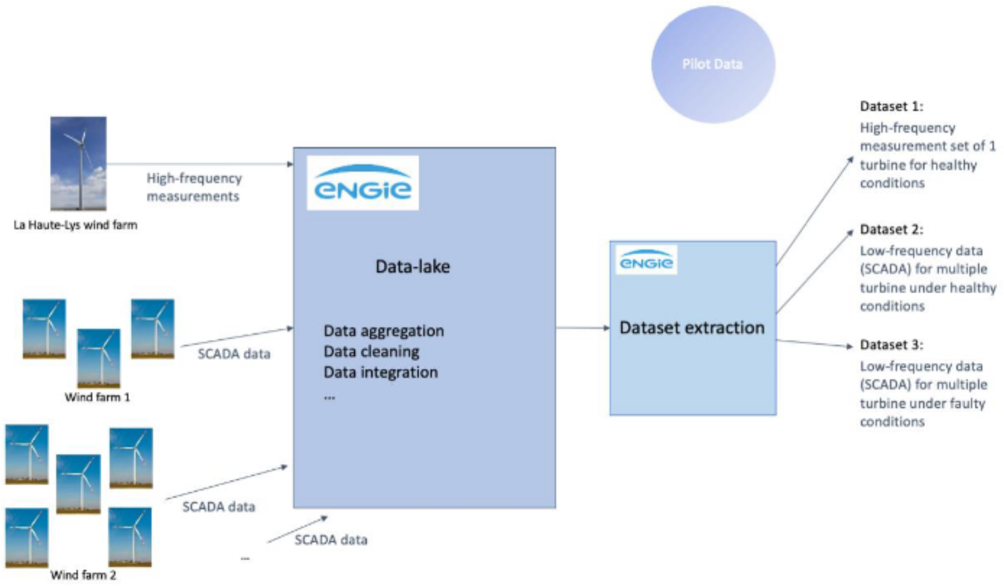

Figure 5: Pilot 1a Data Extraction

Pilot 2a – Electricity Balance and Predictive Maintenance

Pilot 2a focuses on integrating and deploying different PLATOON analytical services with the Institute Mihajlo Pupin (IMP) proprietary VIEW4 Supervisory control and data acquisition (SCADA) system deploys the energy value chain in Serbia. The VIEW4 system controls the production site in the large hydro and thermal power systems and RES, via transmission management to distribution and electricity dispatching. Six use cases are identified in this pilot: 2a-01 - Balancing on a regional level, 2a-02 - Balancing on a country level, reserve/energy exchange process, 2a-03 - Load/Demand forecasting, 2a-04 - RES forecasting, 2a-05 - RES Effects on the power system, and 2a-07 - Predictive maintenance in RES power plants. Uses cases are categorized into different Smart Grid domains: market-related domain (2a-01 and 2a-02), grid-related domains in transmission, distribution, micro-grids (2a-03), resources connected to the grid domains, i.e., Distributed Energy Resources (2a-04 and 2a-05), and support domain functions, i.e., Asset Management (2a-07).

Energy resources related to Renewable Energy Sources (RES) in this pilot include: wind power plants and PV power Plants. Electricity production from solar and wind plants is subject to forecast errors that drive demand for balancing. Use cases 2a-03, 2a-04, 2a-05 and 2a-07 aim at providing such forecasting based on historical data from energy sources. Figure 6 summarizes the interactions among the various data sources.

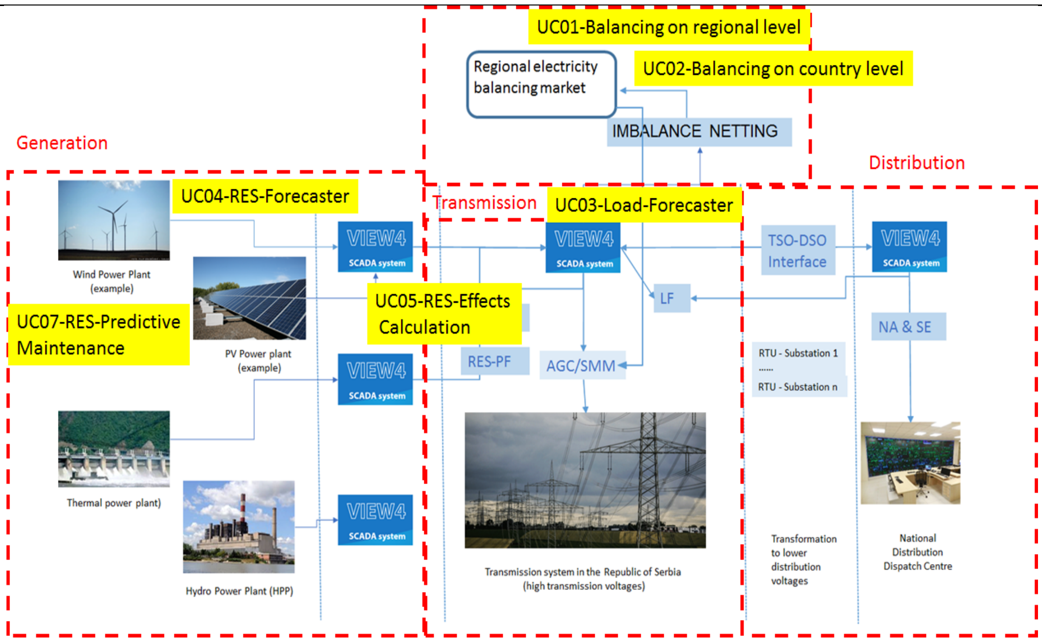

Figure 6: Pilot 2a - Serbian energy supply chain and pipeline for Energy Resources

Two types of data sources are available in this pilot: a) the Transparency Platform data source; and b) Renewable Energy Source (RES) data source.

The **transparency platform data source** contains data related to the electricity market that is published in accordance with the Energy Act, which transposes EU Regulation no. 543/2013 [18]. There are two data providers: 1) Joint Stock Company EMS AD, and 2) ENTSO-E transparency platforms. EMS AD will publish all relevant consumption, transmission and balancing data within the deadlines defined in the Key Market Data Disclosure Policy. In addition to the public data available from these platforms, additional details on each dataset can be provided by IMP. Transparency platform datasets include:

- **Generation (Production) dataset**: energy production and production forecasts data provided by ENTSO-E Transparency platform. It provides the installed capacity, actual generation and generation forecasts per generation unit.
- **Consumption (Load) dataset**: historic data about power consumption (System vertical load from Oct 2016) in the electric market. Power consumption data provided freely by the ENTSO-E Transparency Platform to pan-European electricity market data for all users.
- **Transmission dataset:** data about power transfer over between areas. It provides the current day, day ahead, month ahead, and year ahead data (no historical data), in the electric market.
- **Balancing (Load forecast) dataset**: data about regular energy used to keep the electricity transmission grid in balance.
- **Congestion dataset**: data about actions taken to relieve the overloaded transmission grid, includes Countertrading, Redispatching Internal, Redispatching Cross Border, Redispatching (legacy) and Costs of Congestion management.
- **Outage dataset**: data about planned maintenance and failures inside the electricity transmission grid provided by Transparency Platforms. It provides data about unavailability in transmission, offshore, production and generation units.
- **Short-time Load forecast dataset**: short time load forecast datasets needed for LLUC P-2a-03 load demand forecast on transmission level. Historical data with higher granularity for testing purposes is also available in **IMP SCADA archives**.

The RES data source provides the renewable energy source (RES) energy generation systems, such as wind and PV farms. Renewable energy datasets provided by IMP includes:

- **RES-PROD (Production)**: Historical Wind Power Production Measurements contains measurements of the production from the wind **power plant,** as well as **topology data**.
- **RES-PV (Predictive Maintenance)**: Data will be collected when the Phasor Measurement Unit is installed at IMP side.
- **RES-MET (Meteorological)**: Meteorological Data for RES Production (Generation) Forecasting Modelling Data. Meteorological dataset that will be utilized for RES production forecasting models training process as input data. Data is historical data (**private data** from WeatherBit).
- **RES-Effects**: Effects of Renewable Energy Sources on the Power System will be calculated with CS edge computing services based on the input by Phasor Measurement Unit installed at IMP side.

There are two interpretations of the data points; actual measure data and forecast (predicted) data. Actual data represent measurements and values for confirmed facts that are measured at the time of recording. On the other hand, forecast (predicted) data means data that is yet to be approved or realized, which is not a measurement of values but forecasting of value using experience or historical data. Care must be given when interpreting and integrating such different interpretations of data values. Data related to consumption, generation, weather, and fault data are the main datasets affected by such interpretation conflicts.

Pilot 2b - Electricity Grid Stability, Connectivity, And Life Extension

Pilot 2b takes place in ParcBit technological park located in Palma de Mallorca, Spain. ParcBit's grid is formed by 5 km long mid-voltage network and 5 km low-voltage network. Two different use cases are identified for this pilot: LLUC 2b-01 - Predictive Maintenance for MV/LV Transformers, and LLUC 2b-02 - Non-technical loss detection in Smart Grids. LLUC 2b-01 focus on transformer predictive maintenance, estimating transformer components health and its maintenance costs, planning maintenance actions, monitoring transformers and studying different grid scenarios in case of replacing old transformers or adding complementary transformers. LLUC 2b-02 focuses on quantification of losses in the distribution grid of a DSO and the detection of non-technical losses (NTL), using the available smart meter data from SAMPOL smart grid in ParcBit, in Mallorca, Spain. Pilot 2b resorts to three datasets; Figure 7 illustrates the pipeline for data generation, ingestion, and storage:

- **Power grid ZIV Power Meters**: consists of hourly measurements of active and reactive power delivered to the users (measured by **Smart Meters**), grouped by **concentrator** and identified by power meter.
- **Transformer Sensors data**: consists of data from 8 temperature sensors located at different positions of the transformers, 2 sensors for ambient temperature, humidity and pressure, 1 sensor for oil temperature.
- **Medium-voltage Network Analyzer data**: consists of Electrical Network analyzer for current transformers.
- **Lab test results**: This contains the results of the oil analysis performed on the electrical transformers.

In this pilot, data ingestion, integration, storage, processing, and visualization will be done by INDRA (IND), while Tecnalia (TECN) will develop Data Analytics tools. Thus, data will be exchanged with IND and TECN.

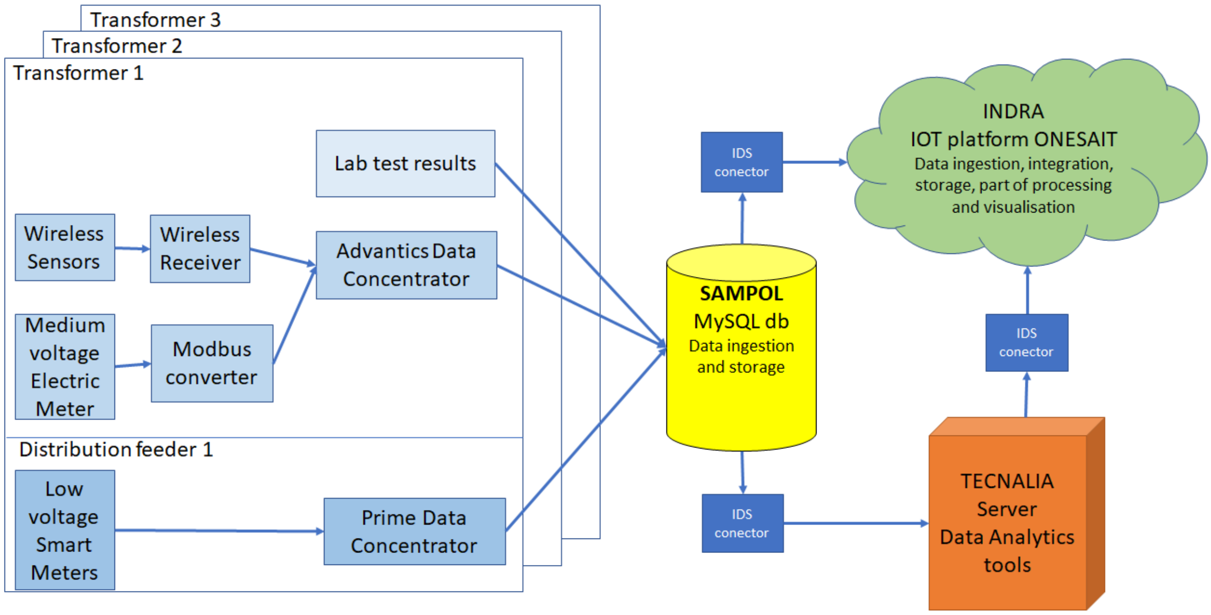

Figure 7: Pilot 2b pipeline for Managing SAMPOL datasets

Pilot 3a - Office Building: Operation Performance Thanks to Physical Models and IA Algorithms

Pilot 3a is related to an office building with a building management system (BMS) controlling the HVAC and comfort in different zones of the building. Two use cases are identified for this pilot: LLUC 3a-01 - Optimizing HVAC control regarding occupancy, and LLUC 3a-02 - Providing Demand Response Service through HVAC control. LLUC 3a-01 aims at providing a smart module for an office building that optimizes HVAC operation in function of real occupancy. Occupancy data are available via dedicated sensors, and the comfort and HVAC controls are available via BMS of the building. Using historical data, some learning algorithms are implemented to predict occupancy and anticipate heating and cooling period in the building and its different zones. LLUC 3a-02 aims at providing a smart module to supervise the implementation of Demand Response services in an office building using HVAC control and building inertia. The module will provide predictions of the HVAC load and the potential flexibility available in the building using the building parameters and weather forecast data. The dataset grows in the order of 100K records per day and the observations are with a resolution of minutes and hours. Figure 8 depicts the interaction among the steps of data generation, ingestion, integration, processing, and aggregation. Managing data ingested from IT connections by zones, Weather APIs, Building Management Systems demand the interaction of data collected from three distinct data sources, which is registered at different time frequencies.

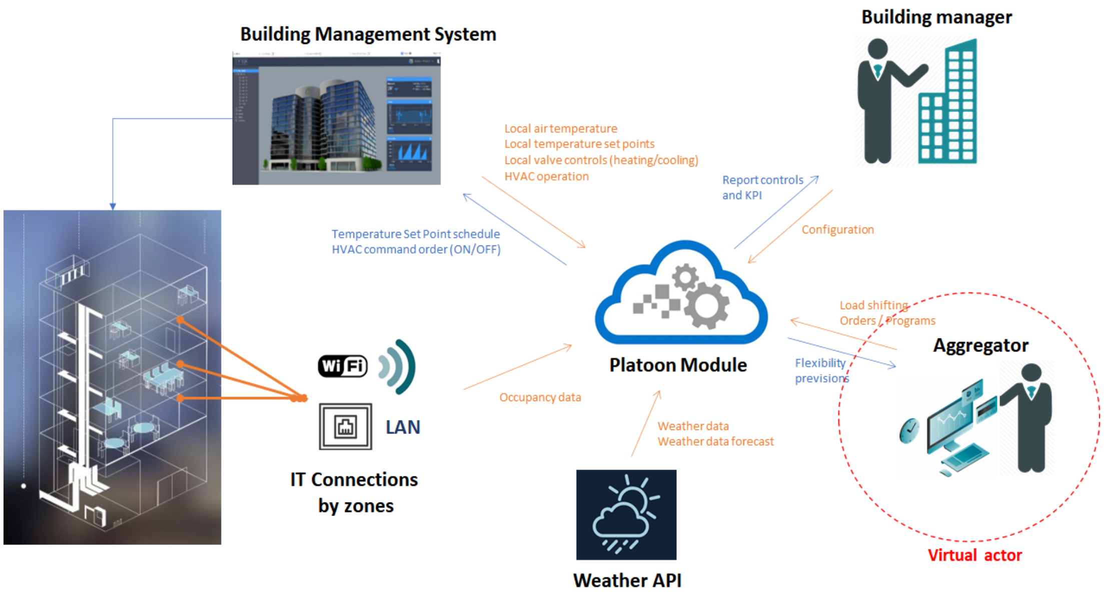

Figure 8: Pilot 3a. Interaction of the Datasets and Stakeholders

Pilot 3b - Advanced Energy Management System and Spatial (Multi-scale) Predictive Models in the Smart City

Pilot 3b will take place in Rome, Italy, and the overall goal of the pilot is to acquire, aggregate and process energy consumption and related data of different buildings to make energy domain specific data analysis such as consumption forecasting, predictive maintenance, benchmarking and so on. Four use cases are identified for this pilot: LLUC 3b-01 - Building Heating & Cooling consumption Analysis and Forecast, LLUC 3b-02 - Predictive maintenance of Cooling and Heating Systems, LLUC 3b-03 - Lighting Consumption Estimation and Benchmarking, and LLUC 3b -04 - Monitoring and Analysis of Energy Meters Data of ROM large Asset. A set of buildings from two different partners will be available for in this pilot; Poste Italiane building for LLUC 3b-01, LLUC 3b-02 and LLUC 3b-03 while Roma Capitale large asset of municipal building data for LLUC 3b-04.
For a better reading of the pilot, we have divided the pilot in two sub-pilot: #3b_PI and #3b_ROM.

Pilot #3b_PI

Poste Italiane buildings to be considered are located in Rome Municipality Area and four different destinations for the building spaces are considered: Datacenter, Logistics distribution and cross-docking (mail & parcels), Retail and Office (Directional), for a total of 16 buildings. Poste Italiane already collects and manages data related to energy use and consumption mainly from mid and big size buildings but is also going to increase the depth and detail of data collected through progressive integration of existing energy consumption devices (lighting, heating, cooling technical plants, …) in a centralized database to be supported with AI tools to determine benchmark, best practices, and areas of opportunities for initiatives to increase energy efficiency and reduce CO2 production.

PI's buildings data source provides datasets on heating, cooling and lighting consumption. The datasets provided by Poste Italiane (PI) includes:

- **Building dataset:** details of static data for each building, such as location, climate zone, surface area, etc).
- **Building Occupancy dataset**: daily number of employees and customers in each of the building of the pilot.
- **Weather Data (meteorological)**: Meteorological dataset that will be utilized for consumption forecasting models training process as input data. (Data recovered by external systems) **Systems Anomalies**: Information based on monitoring of in-building temperature measurements. The system provides alarms when it detects temperatures outside defined thresholds.
- **Calendar:** Information on office openings and shifts.
- **Consumption on Building**: Energy consumption onbuilding (smart and traditional) and internal climate information: information on active energy consumption (kWh) both of building or line and internal temperature and humidity. It will be used for many purposes, such as consumption prediction, consumption benchmarking, anomalies recognition, and lighting consumption esteem. Climate sensors info will be used for many purposes, such as consumption predictions which guarantee a given comfort level, and proper consumption benchmarking.
- **Building Energy Systems**: Information on kind and characteristics of heating, cooling and lighting systems. Building HVAC plant information will be used for many purposes, such as consumption prediction and consumption benchmarking. . Building lighting plants info will be used, for example, lighting consumption benchmarking, and lighting consumption estimation.

Figure 9 shows the interaction among the Pilot 3b_PI datasets.

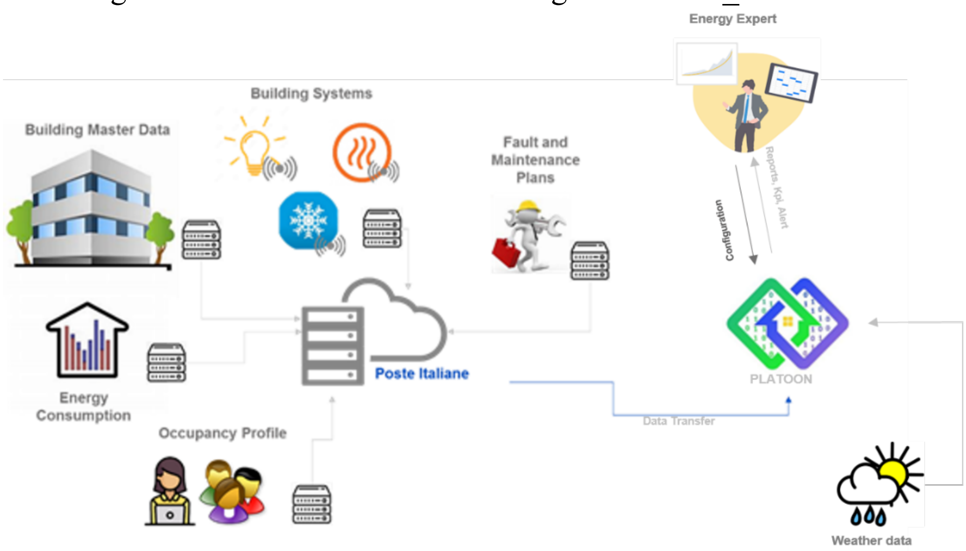

Figure 9: Pilot 3b (PI) Interaction of the Datasets and Stakeholders

Pilot #3b_ROM

The Public Works and Infrastructures Department of Roma Capitale (SIMU Department) includes Plants Division with at least 3 offices managing energy issues: the Energy Manager Office of Roma Capitale (EMO), the Utilities Meters Office (UMO) and the Thermal Plants Office (TPO). This Unit manages around 8,950 energy meters (6,500 electric meters and 2,450 gas meters) related to almost 2,000 buildings and complexes of buildings owned by the municipality (Figure 10).

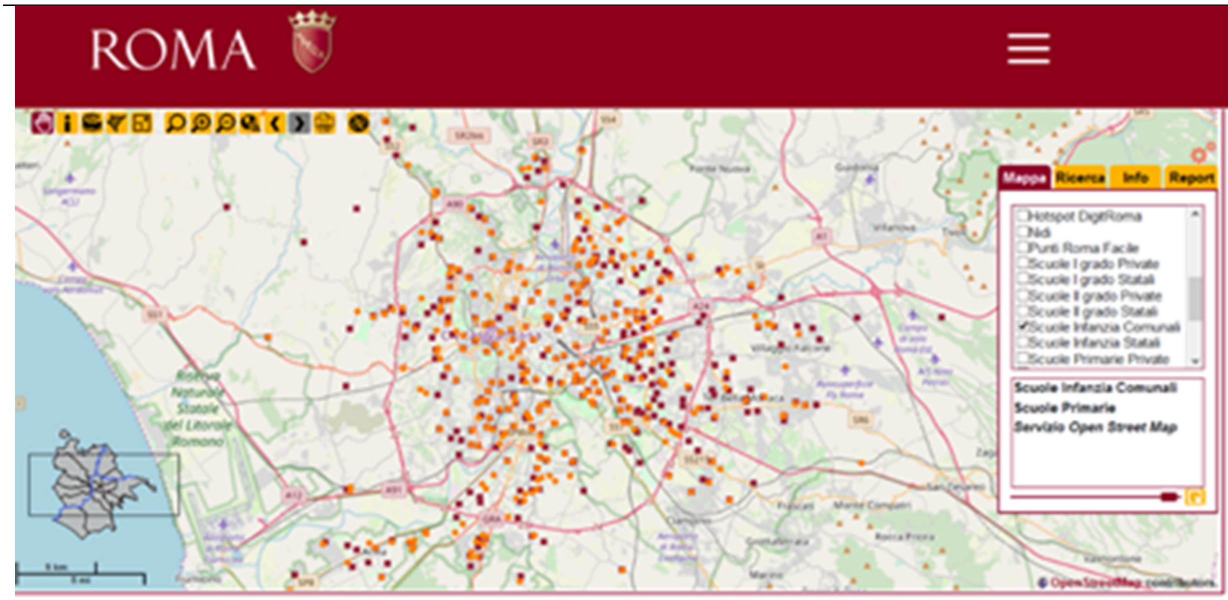

Figure 10 Pilot 3b (ROM) large asset of municipal buildings

To help the offices in this activity, considering the huge amount of data coming from the meters each month, an integrated monitor and analysis system shall be implemented. The proposed tools will increase knowledge and awareness of energy consumption profiles, anomalies, forecasting, and efficiency measures potentialities.

In the initial context the management of these data is fragmented and far from being fully integrated in coherence with a set of general objectives, while the energy consumption datasets, for electricity and gas, are quite heterogeneous. The data should be cleaned, correlated, and analyzed automatically in order to produce a benchmarking focusing on Energy Performances (EP), to highlight anomalies, to generate reports for different purposes also on spatial basis, to produce forecasts in terms of energy consumption (EC) and other reports and assessments in order to tackle energy efficiency activities more effectively. The use cases from 01 to 03 are dedicated to all meters (gas and electric) supplying energy to the large asset of buildings. The fourth use case focuses on the PV installation potentialities on all the asset buildings, calculating the power peak and total PV energy productions on the basis of the available surfaces, comparing with buildings self-consumptions (electric meters data) and PV productions on going where PV plants are already existing, resulting in automatic balances in terms of potential Energy Community needs that each building can serve.

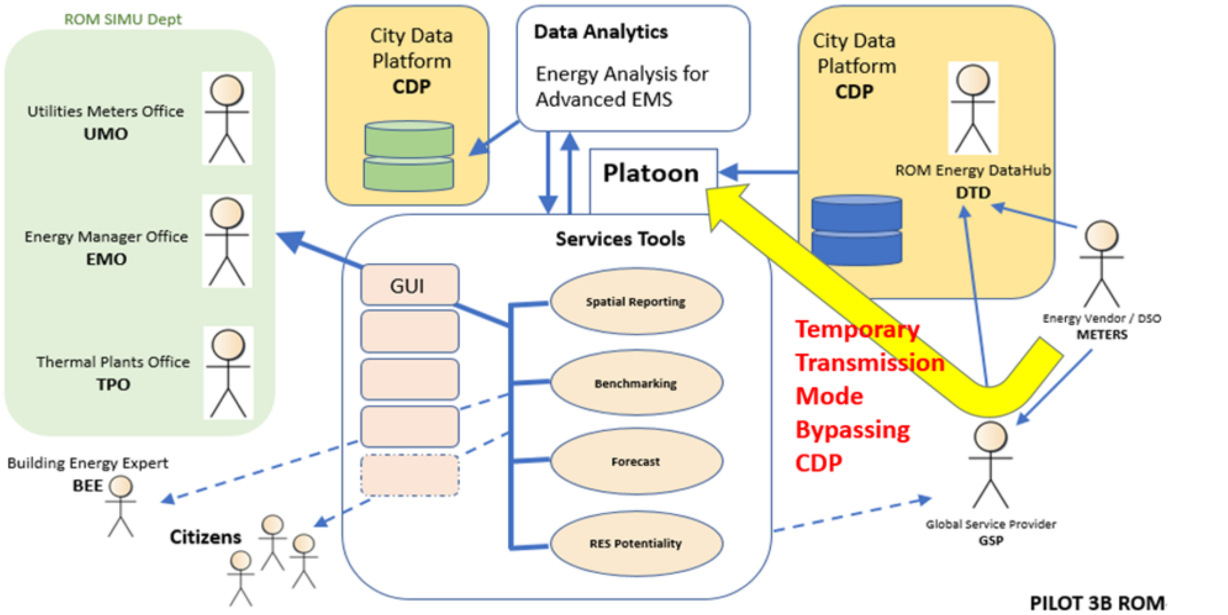

Figure 11 depicts relationships among datasets and users.

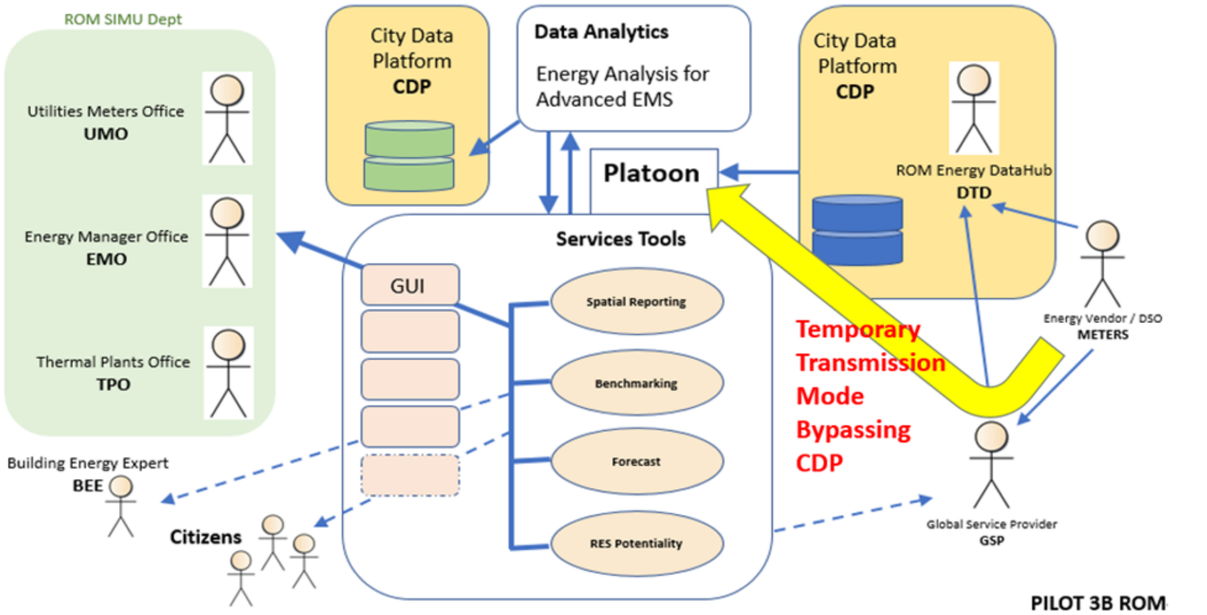

Figure 11 Pilot 3b (ROM) Relations among Datasets and Users

Pilot 3c - Energy Efficiency and Predictive Maintenance in the Smart Tertiary Building Hubgrade

Pilot 3c takes place in Nanogune, a tertiary sector smart building dedicated to nanotechnology research, based in San Sebastian, Spain. This building is divided into different areas, such as offices and laboratories, and has both thermal and electrical meters to differentiate the areas. Two use cases are identified in this pilot: LLUC 3c-01 - Advanced EMS for tertiary Buildings, and LLUC 3c-02 - Predictive Maintenance in Smart Tertiary Building Assets. In LLUC 3c-01 the Advanced EMS will optimize the local renewable energy resources (RES) and HVAC operation as a function of building and RES characteristics, building comfort constraints, ambient conditions, and energy market price following a multi-objective pattern which targets to reduce the overall energy bill and maximize the usage of RES. LLUC 3c-02 aims for the development and implementation of predictive maintenance tools for the thermal control assets of smart tertiary buildings (specifically chillers, pumps, and distribution rings). The objective is to improve the maintenance policy, increasing the availability and useful life of these assets and reducing the general maintenance costs. Pilot 3c is built on the following three datasets stored in an SQL server database, and the scope is the North of Spain. Figure 12 illustrates the pipeline of how data ingested, stored, and shared stakeholders.

- **Energy Huybgrade dataset**: consists of SCADA data about buildings (up to 300) with observations collected from thermal, electric, and gas meters. The values collected correspond to temperatures, water, electricity and thermal consumption, position of the valves, dumpers. Moreover, the weather data and forecasts are part of this dataset. The observations are registered every 10 minutes and the database grows 1.5MB per day per building.
- **PRISMA software dataset**: contains maintenance logs about events compliant with norms of preventive maintenance; they include data on work orders and executions. The growth trend is not stable and is expected to have a volume less than 1MB per year per building.
- **Condition monitoring dataset**: comprises real time computing capabilities in critic devices of a building. It is expected to be registered at 10kHz.

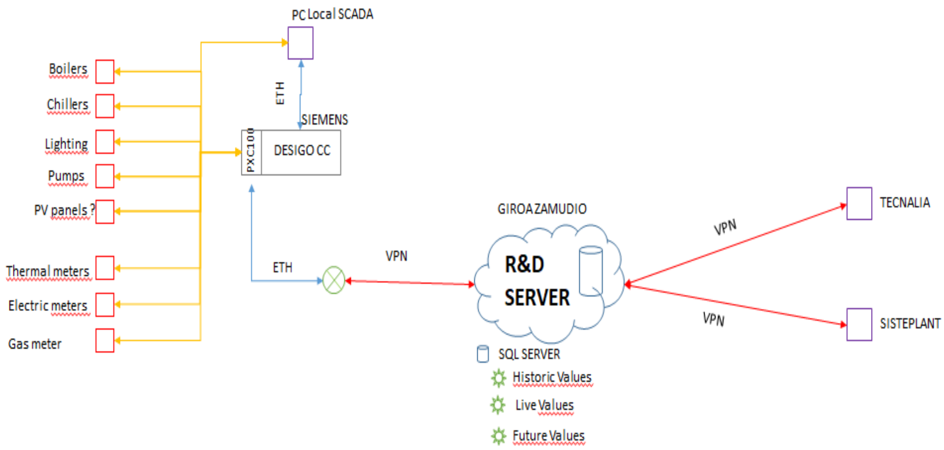

Figure 12: Pilot 3c. Interaction of Building Devices, Database Storage, and Stakeholders

Pilot 4a - Energy Management of Microgrids

Pilot 4a takes place at the Politecnico di Milan's Multi-Good Microgrid Laboratory (MG2lab) in Milano, Italy. MG2lab is an experimental facility for real-life scale research, simulation and test purposes, thus, allowing to study new data-driven paradigms for energy management able to deal with increased complexity of the energy systems and to assess the advantages of innovative strategies. The main use case identified in this pilot is: LLUC 4a-01 - Energy Management of Micro-grids, where the goal is to study data-driven energy management able to deal with increased complexity of the energy systems and to assess the advantages of innovative strategies, by means of EMS with real-time processing and optimization for small-scale/renewable electricity generation, based on power generation and load forecasts. Pilot 4a consists of four datasets from the area of Milan, Italy; the first three are CVS files while the one referring to full sky imaging is JPEG. Figure 13 depicts the energy flow from the energy sources to endpoints in the microgrids.

- **Microgrid PV power production and forecast**: consists of forecasting and modeling of Photovoltaic (PV) power. The dataset is expected to grow with more than 30K records per day, and the updates are per minute.
- **Microgrid battery**: comprises observations of batteries described in terms of State of Charge (SOC), State of Health (SOH), Direct Current (DC), and Alternate Current (AC). Current and voltage are registered, as well as average cell temperature and average ambient temperature. This dataset grows in 86K records per day, and new observations arrive per 1 sec.
- **Microgrid potable water production**: contains relevant measurements of a plant for potable water production. The dataset collects active and reactive power values, frequency of pump rotation, feed and permeate water conductivity, concentrate and permeate water flow rate, and temperature and pressure in the hydraulic circuit. It has a growth trend of 1,440 records per day, and updates are per minute.
- **Microgrid weather parameters:** consist of observations sensed by a weather station. It reports ambient temperature, wind speed, wind direction, relative humidity, rain, and irradiance (diffuse, total horizontal, and total on the tilted plane). The growth trend is 65K records per day, and observations are registered every 10 seconds.
- **Microgrid full skype imaging:** comprises full-sky images in JPEG format. It grows in more than 250 records per day every 5 minutes.

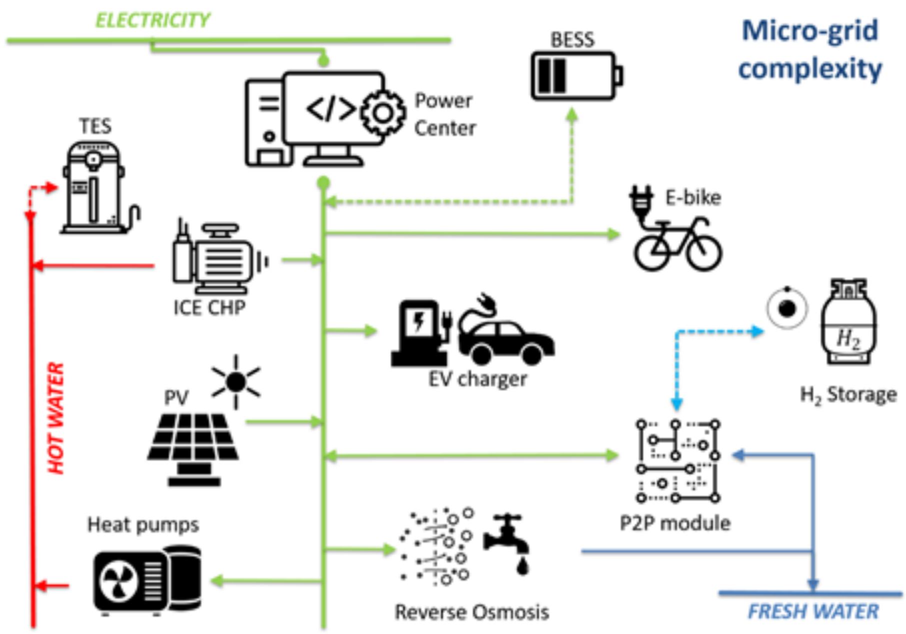

Figure 13: Pilot 4a. Diagram of Energy Flow

4.2 The 5V's of the PLATOON Data Sources

The 5Vs big data model is used to describe the main characteristics of the PLATOON data sources. Table 2 summarizes the main characteristics of datasets in Section 00. Note that velocity is reported in several datasets in terms of the units mHz, Hz, and kHZ. These metrics are defined as follows. Consider a sensor observing measurements at 1 mHz, then the sensor makes 1 observation every 1,000 seconds, i.e., the dataset will maintain an observation every 1,000 seconds. Suppose a sensor observes a measurement at 500Hz, then the observed measurements are stored every 1/500 sec, i.e., every 2 msec. Lastly, a sensor that observed at 5 kHz, makes a new observation every 0.2 msec, i.e., the dataset will store an observation every 0.2 msec. As observed all the datasets can be considered as big data because they enclose data that meet the 5-Vs of the big data model. It is worthy to mention that the datasets of pilots 1a and 2a can also consider very large according to their growth trends, which range from 100K per day to 5KB per second, respectively.

Pilot	Volume	Velocity	Variety	Value	Veracity
1a	Datasets vary from Gb to Tb per year. **Growth Trend > 100K per day**	From **mHz to Hz** depending on the dataset. 10-min averages and statistics	SCADA data: ~50 tags/WT Status codes: ~600 status codes/WT. Diverse formats: CSV, TDMS, .mat Acceleration signals of 10 accelerometers and 2 encoders. Signals of current probes as well as turbine controller parameters (collected at lower frequency (1Hz)).	Predictive Maintenance	Missing Values and Observations
2a	Logs with a total size in **Petabyte** order of magnitude. Up to 3 million new entries a day. **Growth Trend > 5 KB per second**	Updates in the order of Hertz (Hz). Per Minute, hourly, daily, weekly, monthly and yearly basis. Depending on the dataset	SCADA data. Weather data from Weather APIs. Diverse formats: CSC, XML, CAD, columnar, Different language: English, Serbian, Russian	Electricity Balancing and Predictive Maintenance	Missing Values and Observations
2b	Logs in the order of Gb; initially 420 MB. Data from October 2016 to 2022.	Updates in the order of Hz. **Updates hourly,** for a total of 77 power meters. Depending of the datasets, values can be received every 5 minutes	MySQL Database. Languages: English, Spanish	Predictive Maintenance	Missing Values and Observations
3a	In the order of Gb. **Growth Trend > 100K per day**	Updates in the order of **minutes and hours**	Data collected from IT connections by zones, Weather APIs, Building Management Systems. Language: Italian	Optimizing Heating, ventilation, and air conditioning (HVAC)	Missing Values and Observations
3b	Monitoring data up to Gb per year. **Growth Trend > 2 MB per year.**	Updates range from **minutes, day, and year**	Diverse data models: CSV and XSLx files and relational. Language: Italian	System Anomalies	Missing Values and Observations

3c	Initial size 1Gb aprox. **Growth Trend > 1.5 MB/day**	Updates in the order of **minutes and Hertz (Hz)**	SCADA data. Diverse formats: Relational database and JSON	Energy Efficiency and Predictive Maintenance	Missing Values and Observations
4a	In the order of Gb. **Growth Trend > 86K per day**	Updates in the order of **seconds and minutes.**	Diverse formats: CSV and JPEG	Energy Management	Missing Values and Observations

Table 2: Big Data Characteristics of the PLATOON data sources

5. Energy Big Data and Interoperability Conflicts among Energy Data Sources

As discussed in section 4, there are seven pilots involved in the PLATOON project. Each pilot has specific KPIs to evaluate different use cases and use their data. Nevertheless, it is possible that they share several concepts, as identified by task T2.3, and external data from other providers (e.g., weather forecast data) or between partners in the project. Solving the interoperability issues within and between pilots in this project through semantic modeling and integration is critical to validate, and different techniques and tools developed at a high level. For instance, building energy management and predictive models designed and validated by one pilot can be applied to another pilot or production environment.

5.1 Interoperability Issues among PLATOON Data Sources

Based on the description of data sources provided by PLATOON partners, the interoperability conflicts among energy sources are analyzed. Descriptions are collected from partners who are data providers; the questionnaire (in 0) is used to gather data providers' answers.

Structuredness (C1): The PLATOON data sources have two levels of structuredness. The *structured* data sources include: Building energy consumption, and RES data sources are stored in MySQL database. Contrary, Transparency platform data is semi-structured. Data sources such as RES-PROD, RES-PV, and BEMS are structured data sources stored in a relational database (MySQL or SQL Server), while weather data MET-RES and Building power consumption (PI, ROM) are semi-structured data sources, structured as XML and CSV, respectively. Based on the description of data sources provided by PLATOON partners, there are no *unstructured* data sources. The data integration techniques defined in this task will integrate these structured (relational data in MySQL) and semi-structured (CSV, XML, JSON, and XLSx) data sources into the PLATOON unified knowledge graph.

Schematic (C2): schematic interoperability conflicts exist among PLATOON data sources from pilots as the data is generated independently of each other. For instance, the concepts that represent the HVAC and its subcomponents from different pilots might have different semantics. Another issue is the semantic of temperature measures from weather data sources; e.g., temperature represented from external sources such as Weather forecasting and temperature measures from Temperature Sensors. Such interoperability issues will be solved by using the PLATOON common data model.

Domain (C3): this interoperability conflict occurs when various interpretations of the same domain are represented. Domain conflicts exist because different energy generation domains and consumption are included in the PLATOON unified knowledge graph. Wind, Solar, Nuclear, and other energy generation plants and heating, cooling, and lighting consumption have different domain data that will cause conflict among different domains.

Representation (C4): this interoperability conflict refers to different representations used to model the same concept. Energy consumption data from the Transparency platform and Smart Building has representation conflict because of the other measurement units used. MegaWatts is used in the Transparency Platform, while Kilowatts is used in Smart Building energy consumption. Furthermore, these data sources use different date representations: "DD/MM/YYYY", DD-MON-YY, "YYYY-MM-DD", "YYYY", and "YYYY-MM-DD:HH:MM:SS". Additionally, interoperability issues arise between date-time values for failure events, maintenance planning, etc. For instance, system alarms for building data from pilot 3b_PI represent dates using 'DD/MM/YYYY' format, while Office opening dates from

Calendar data is represented in 'YYYY-MM-DD:HH:MM:SS' format. To solve these conflicts attributes of these concepts should be standardized to the same representation format. To solve these representation conflicts, attributes will be modeled using the same types and properties into the PLATOON unified knowledge graph.

Language (C5): this interoperability conflict occurs whenever different languages are used to represent the data or metadata (i.e., schema). Data from JSC EMS Transparency platform contains some data points represented in English and others in Serbian and Russian. English language is used to represent data in the ENTSO-E Transparency platform, while Italian is used to represent data in Smart Building. Sensor data from PI and ROM are represented in Italian while it is represented in Serbian, Russian and English from IMP (Serbian SCADA and Transparency platforms: JSC EMS and ENTSO-E, respectively). Entity linking and matching techniques will be deployed to solve the language interoperability conflicts among these data sources.

Granularity (C6): this interoperability conflict refers to the level of granularity used to collect and represent the data. Transparency platforms and smart building energy consumption data are described in different levels of detail. Transparency platform energy consumption is provided at control area, bidding zone, and country level. Smart building energy consumption, on the other hand, is provided at the building and department level. Similarly, the energy consumption of transparency platforms is an aggregation of all energy usage (load) of control area, bidding zone, and country, respectively. In contrast, energy consumption from the smart building is the aggregation of energy systems like cooling, heating, and lighting usage (load) at the department and building level, respectively. Furthermore, interoperability conflicts occur between data sources related to measures from weather data, for instance, temperature measure at country, city, or specific GPS location. Furthermore, energy consumption and generation data from GasMeter or ElectricMeter can have different levels of granularity: all gas heating plants, each gas heating plant, all cooling, ventilation, lighting, etc. at department level, zone level, or building level. In addition, different frequency or velocity of measures from energy consumption and production pilots (3a, 3b, 3c and 4a) have different granularity. The PLATOON unified schema will be modeled to solve the granularity conflicts, handle different aggregations' levels, and integrate to the same semantic concept as defined in task T2.3 – Data Models.

5.2 PLATOON Semantic Data Model for Interoperability among Data Sources

The PLATOON semantic data model unifies the meaning of the data collected from the PLATOON data sources; it is defined in WP2 task *T2.3 – Data Models*. This data model comprises concepts and their relationships of the energy domain ontology. For instance, Figure 14depicts the building system data model that will be used to represent Smart City buildings data. More detail on the semantic data models for PLATOON unified knowledge base can be found in deliverable *D2.3 – PLATOON Common Data Models for Energy* [2].

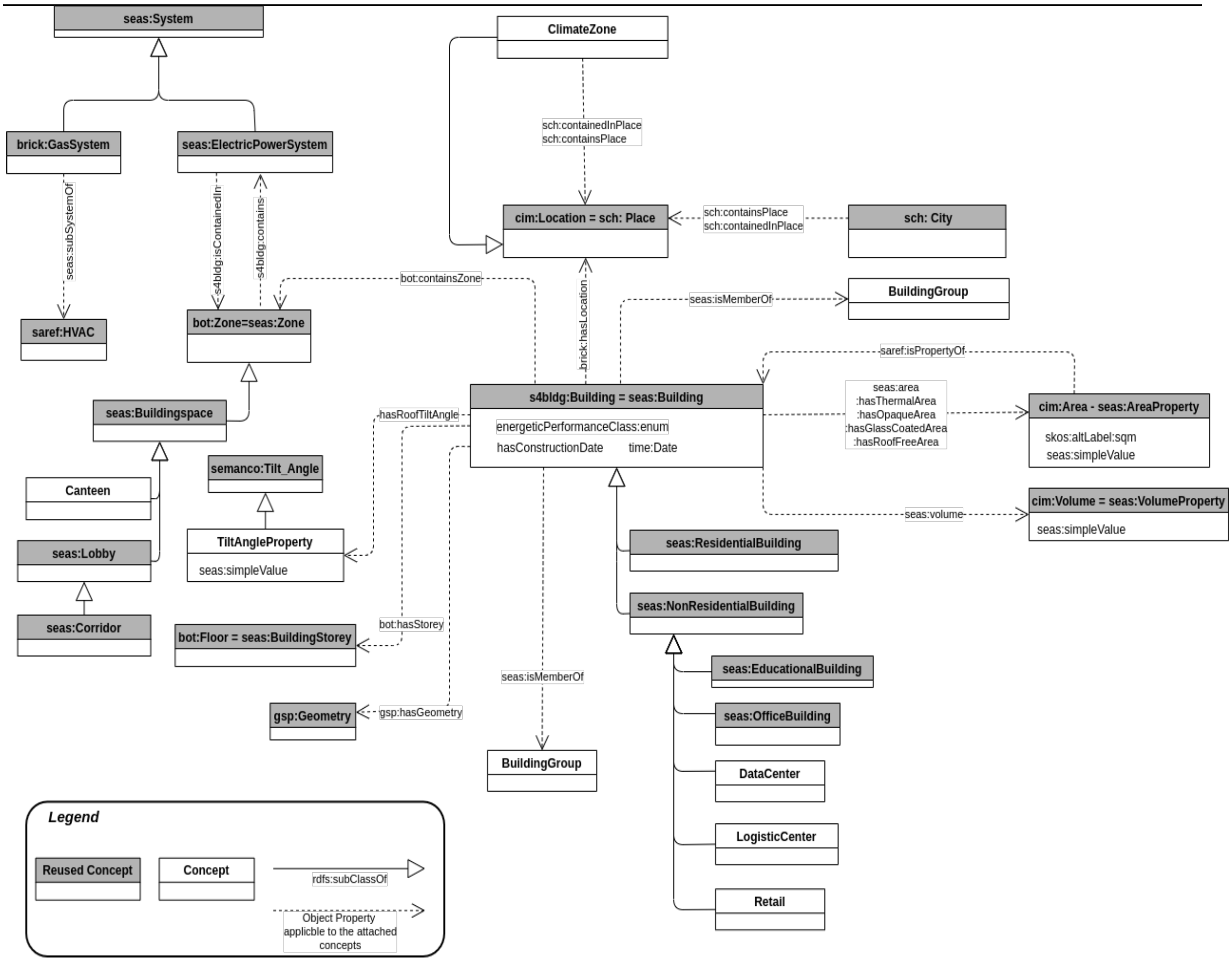

Figure 14: Building Systems Semantic Data Model (taken from D2.3 [2])

The relationship between the PLATOON semantic data models and the data sources is described to outline interoperability issues across the PLATOON datasets. The PLATOON domains (identified in task T2.3 and reported in D2.3 [2]) are used to characterize concepts. They include: i) Common domain, ii) electricity generation from Wind power production and electricity generation domain, iii) Smart grid/microgrid, electricity generation and balancing domain, iv) Buildings and Zones domain, and v) HVAC equipment and its subsystems domain. Figure 15 (taken from D2.3 [2]) summarizes the relationships among these domains. The reported analysis is built on top of these domains to elaborate on the interoperability conflicts existing among the PLATOON datasets.

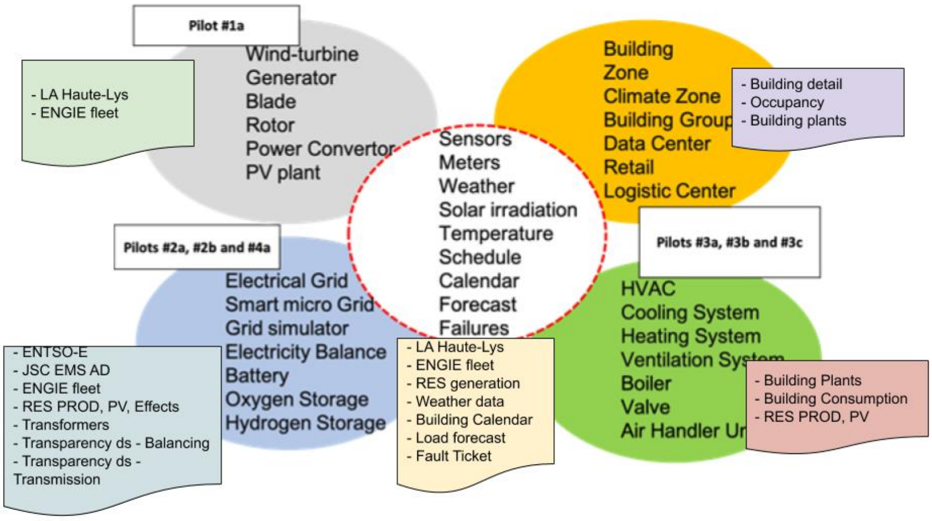

Figure 15: The PLATOON domains. (Based on Figure 16 in D2.3)

Common domain: contains a set of concepts that are common to most of the pilot use cases related to Sensors, Meters, and Meteorological data. Table 3 shows the description of the main concepts represented in common domain with respect to the available data sources for the PLATOON pilots.

Name of the Concept	Related Data Sources	Description of the Concept
Temperature	Building Weather data, ENGIE fleet dataset, Microgrid weather parameters (MG2lab), MicroGrid potable water production (MG2lab), Energy Huygrade (GIROA)	Temperature of building (inside and outside), wind turbine generators, wind turbine converters measured from TemperatureSensor.
Metering GasMeter, ElectricEnergyMeter	Building energy consumption, La Haute-Lys, ENGIE fleet dataset, Energy Meter Gas (ROM), Energy Meter Electric (ROM), RES-PROD (IMP), RES-PV (IMP), Energy Huygrade (GIROA), Power grid ZIV Power Meters (SAMPOL)	Meter or evaluation of energy consumption or production; Gas and Electric.
SolarRadiation	PV plant, RES-PV (IMP)	Solar radiation measures
SensorEvaluation	Building Weather data, External Weather data, La Haute-Lys, ENGIE fleet dataset, RES-PROD (IMP), RES-PV (IMP)	Sensor evaluation is a measurement/observation value of a certain property, such as temperature, wind, humidity, power quality, etc.
WindSpeed, WindSpeedEvaluation	La Haute-Lys, ENGIE fleet dataset, RES-PROD (IMP),	Wind speed measured near the location of wind turbines, wind

	Microgrid weather parameters (MG2lab)	farm, etc
WindDirection, WindDirectionEvaluation	La Haute-Lys, ENGIE fleet dataset, RES-PROD (IMP), Microgrid weather parameters (MG2lab)	Wind direction measured near the location of wind turbines, wind farm, etc
Humidity	Building Weather data, Microgrid weather parameters (MG2lab), Energy Huygrade (GIROA)	humidity measured or forecast of weather data
AirTemperature	Building Weather data, MET-RES (IMP)	Air temperature measured by a temperature sensor or forecast
Sensor, Anemometer, Pyranometer PMU	Building Weather data, Occupancy (ROM), La Haute-Lys, ENGIE fleet dataset, RES-PROD (IMP), RES-PV (IMP), PMU Power quality (CS)	Sensors are devices that measure certain properties, such as temperature, wind speed, wind direction, humidity, power quality, etc.
FailureEvent	) System Anomalies (PI), Outage (IMP-ENTSO-E), Condition monitoring (GIROA), La Haute-Lys, ENGIE fleet dataset	Failure or damage to devices, transmission, or any system in the grid or building systems
Maintenance	, Maintenance (ENTSO-E), Condition monitoring (GIROA)	Maintenance in the grid or building systems that are scheduled or executed.
SolarPanel	RES-PV (IMP), PV plant, Microgrid PV power production and forecast (MG2lab)	Solar panel for energy production.

Table 3: main concepts represented in common domain with respect to the available data sources

Electricity generation from Renewable Energy Source (RES) domain: contains a set of concepts related to wind turbine power production and electricity generation. This domain is mainly related to pilot 1a, and in part pilot 2a for RES production dataset. Table 4 shows the available data sources that can be represented using concepts in this domain.

Name of the Concept	Related Data Sources	Description of the Concept
WindTurbine	La Haute-Lys, ENGIE fleet dataset	Wind turbine for RES energy generation. Two types of wind turbine: Onshore (on land) and Offshore (on body of water). Three key components of wind turbines: Blade, Nacelle, and Converter.
Nacelle	La Haute-Lys, ENGIE fleet dataset	Nacelle is one of the key components of a wind turbine. Nacelle includes the gearbox: the

		Controller and the Generator.
Generator	La Haute-Lys, ENGIE fleet dataset, RES-PROD (IMP)	Generator is one of the subcomponents of the Nacelle for electric generation.
Power converter	La Haute-Lys, ENGIE fleet dataset	Power convert is one of the subcomponents of a wind turbine that is capable of adjusting the generator frequency and voltage to the grid.
Wind Speed	La Haute-Lys, ENGIE fleet dataset, RES-PROD (IMP), Microgrid weather parameters (MG2lab)	Wind speed measured via sensors or forecasts.
Wind direction	La Haute-Lys, ENGIE fleet dataset, RES-PROD (IMP), Microgrid weather parameters (MG2lab)	Wind direction measured via sensors or forecasts.
Transformer	La Haute-Lys, ENGIE fleet dataset, Transformer Sensor (SAMPOL), Medium-voltage network analyzer (SAMPOL)	Transformer is capable of transforming electricity within a power network that is connected to a wind turbine.

Table 4: Main concepts in Electricity generation from wind power production and electricity generation domain and related data sources

Smart grid/microgrid, electricity generation and balancing domain: consists of concepts related to managing smart electric grids, electricity generation and balancing domain. Data sources available from pilots 2a, 2b and 4a are aligned with these concepts in Table 5.

Name of the Concept	Related Data Sources	Description of the Concept(s)
Grid topology	Microgrid PV power production and forecasting (MG2lab)	Topology of the energy grid/microgrid which includes units for generation, consumption, transmission, distribution, transformation, etc and the connection between them.
Grid Power Properties	SLTF – Short Time Load Forecast, ENTSO-E Load Balancing (forecast), Medium-voltage network analyzer (SAMPOL)	Characteristics of the energy grid (Smart grid/microgrid)
Electricity generation	RES-PROD (IMP)	Electricity generation from RES and other production plants, including Nuclear, Hydro, Wind, Solar, Geothermal
Electricity balancing, BalanceSupplier,	ENTSO-E Load Balancing (forecast) and JSC EMS	Balancing is a process of activating secondary and tertiary

BalanceResponsibleParty, ReserveReq ImbalanceSettlement	Load Balancing (forecast) data	reserves in order to maintain the sum of power exchange with the regional power systems and frequency at the planned value.
Transmission, ScheduledTransmission, Agreement	Transparency Platform transmission data	Transmission of power between provider and consumers according to an agreement between transmission system operator and the balancing service provider.
Distribution	SLTF – Short Time Load Forecast, ENTSO-E Load Balancing (forecast) (on national level)	Distribution of electric power between generation units to consumers via transformers and substations from distributed energy generation (DER).
Transformer	Transformer Sensor (SAMPOL), Medium-voltage network analyzer (SAMPOL)	Transformer is capable of transforming electricity within a power network between the primary connection point and a secondary connection point.
Power System Resources	ENTSO-E	Power system resources can be an item of equipment such as switches and containers such as substations, charging stations, electric vehicles, electric bikes, electric hubs, etc.
Energy Market, Regional Transmission Operator (RTO), Bid, Balancing Supplier	ENTSO-E and JSC EMS Transparency platform data sources	Energy market for power bidding, transfer and balancing at different control areas, units and countries through registered suppliers, balancing operators, etc.
Producer, Consumer, Prosumer	ENTSO-E and JSC EMS Transparency platform data sources	Power producer, consumer and hybrid (producer and consumer) organization in the Energy market.
Consumption	ENTSO-E Consumption (LOAD) data, Smart Meter data (SAMPOL)	Energy consumption at different levels: regional, country, control area, etc.
Substations	RES-Effects (IMP)	Substraitions are part of power system resources in energy grid for transmission and distribution to consumers or storage
ActivePower, ReactivePower, VoltageProperty, CurrentProperty	Smart Meter data (SAMPOL), Concentrator data (SAMPOL) Power Quality data (CS)	Subsystems of a transformer
StorageSystem	Microgrid Battery (MG2lab)	Storage for power includes: ElectricPowerStorage system, HydrogenPowerToPower system,

		OxygenStorage system, Hydrogen storage system, ThermalStorage system and Battery.
ElectricalGrid	Transformer Sensor (SAMPOL), Medium-voltage network analyzer (SAMPOL) PMU Power quality (CS)	Electric grid is an interconnected network for distributing electricity from producers to consumers.
SmartMicrogrid	Transformer Sensor (SAMPOL), Medium-voltage network analyzer (SAMPOL)	Smart Microgrid is one type of electric grid which is connected to distributed electric power producers of photovoltaic plants (PVSystem)

Table 5: Main concepts in Smart grid/microgrid, electricity generation and balancing domain and related data sources

Buildings and Zones domain: is characterized by a set of concepts that describe buildings, zones, and characteristics in terms of occupancy, calendar, and comfort level. Table 6 maps the data sources available from pilots 3a, 3b and 3c to the main concepts in this domain.

Name of the Concept	Related Data Sources	Description of the Concept
Building	Building details (PI, ROM, ENGIE), Energy Huygrade (GIROA)	Building characteristics
Zone	Building details (PI, ROM, ENGIE), Energy Huygrade (GIROA)	Zones of a building
Heating and Cooling	Building Energy Systems (PI), Building Energy Systems (ENGIE), Energy Huybgrade (GIROA)	Heating and Cooling systems (Gas or Electricity) in a building
Occupancy, OccupancySensor, OccupancyForecast	Building Occupancy (PI), Occupancy Estimation data from Sensors (ROM), Occupancy data through Sensor Estimation (ENGIE), Energy Huygrade (GIROA)	Occupancy of the building by customers or employees. Occupancy can be the actual number of customers/employees or it can be estimates or forecasts. Number of customers/employees can be measured using sensors or can be found from a building schedule/calendar.
ComfortLevel	Building data (PI, ROM, ENGIE, GIROA), Internal Comfort level(ENGIE), Energy Huygrade (GIROA)	Comfort level of heating, cooling, and lighting depending on temperature and humidity measures of a building/zone. Thresholds of comfort level can be set such as minimum, basic and best comfort levels.
Calendar	Building Calendar, Energy Huygrade (GIROA)	Calendar for opening and closing hours/dates of a building/zone.

ElectricPowerSystem	Building Plant (Building Energy Systems),	Power systems in a building.
GasSystem	Building Plant (Building Energy Systems)	Gas systems for heating, cooling and lighting in a building
Location	Building details, Building Energy Systems	Location of a building and zone as well as location of building sensors, valves, heating systems, cooling systems, lighting.
NonResidentialBuilding, ResidentialBuilding, Department	Building details – Building Type	Type of building as residential and non-residential. Non-residential buildings also can have different departments; retail, data center, logistic center, office, etc.

Table 6: Main concepts in the Buildings and Zones domain and related data sources

HVAC equipment and its subsystems domain: consists of concepts that represent HVAC systems and its subcomponents for energy consumption and generation in pilots 3a, 3b and 3c. Table 7 describes the main concepts in this domain and the available data sources in pilots that can be represented.

Name of the Concept	Related Data Sources	Description of the Concept
HVAC	Building Plant (Building Energy Systems), Building management system (BMS) HVAC operations (ENGIE), Energy Huygrade (GIROA)	HVAC is a set of components such as boiler, pump, fan, filter, and valves. It has three subsystems: heating, cooling and ventilation systems. HVAC is located in buildings controlled by building management systems. HVAC has three kinds of operations: heating execution, cooling execution, and ventilation execution at a specific time.
HVAC Parts, AirHandleUnit, Boiler, HeatingCoil, Chiller, CoolingCoil Fan, Coil	Building management system (BEMS) HVAC operations (ENGIE), Energy Huybgrade (GIROA)	HVAC system subsystems: air handle unit, boiler, heating coil, chiller, and cooling coil. Air handle unit is connected to heating, cooling and ventilation systems and has fan and coil subsystems.
Energy Consumption	Energy Huygrade (GIROA), BEMS consumption data (ENGIE)	Energy consumption of heating, cooling and ventilation systems
AirFlow, AirFlowSensor,	Energy Huygrade (GIROA), Heating and Cooling room data (ENGIE)	Flow of air from a Fan (SupplyFan and ReturnFan) measured by AirFlowSensor.
Contract, GasContract, ElectricityContract	Energy Huygrade (GIROA)	Contract is agreement of transactions between competent parties to which parties agree to be

		legally bound. Two types of contracts: GasContract and ElectricContract characterized by selling and buying prices.

Table 7: Main concepts in the HVAC equipment and its subsystems domain and related data sources

6. The PLATOON Data Integration Platform

The PLATOON data integration platform is presented as an instantiation of the Data Ecosystem (DE) introduced in Section 2. Then, the PLATOON unified knowledge base creation pipeline per pilot is described and illustrated with an example.

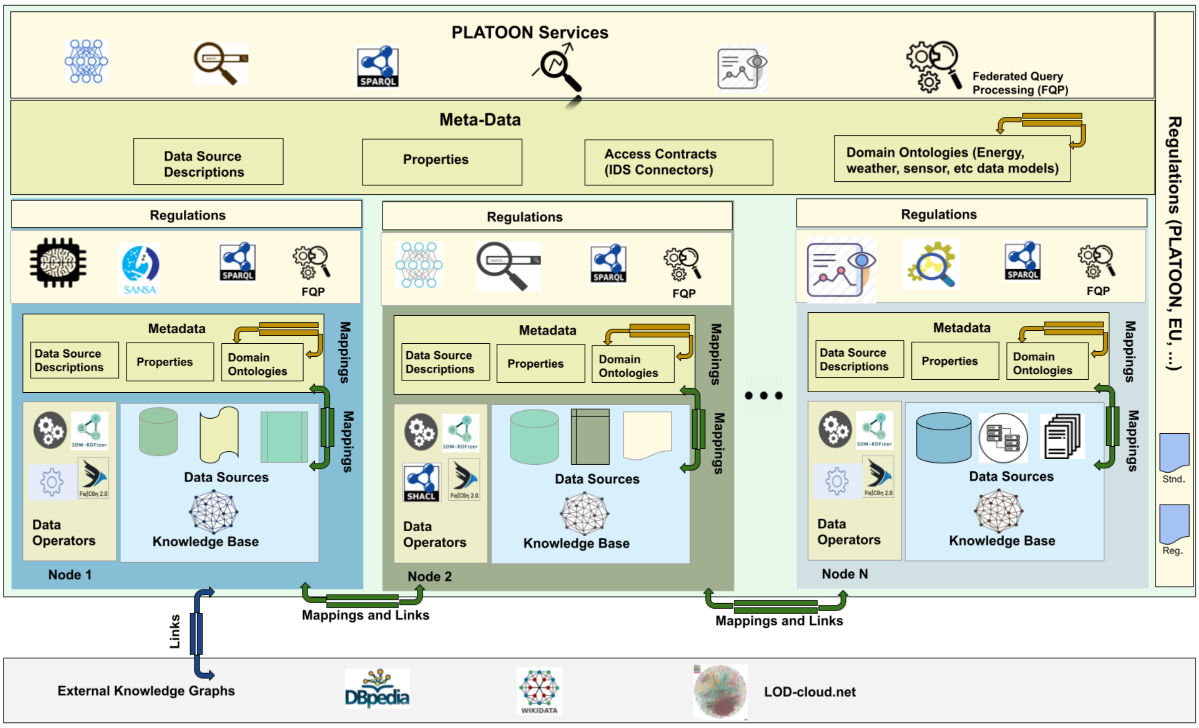

Figure 16: The PLATOON data integration platform as a Data Ecosystem

Figure 16 depicts the PLATOON data integration platform as DE composed of data integration platforms per each pilot (Node i). Each node corresponds to a DE and can be integrated on the PLATOON level through mappings among pilots, data sharing, and service agreements. Each node (in the figure denoted by Node 1 and Node2) applies a data integration process on a specific PLATOON pilot and can deploy its services for query processing, analytics as well as dashboards. Communication between nodes needs to be through an access agreement and can employ data connectors (IDS connectors) to secure data exchange according to data access contracts and regulations. Nodes have control over their data and may have data integrated in a unified knowledge base. Moreover, each individual knowledge base can be linked to knowledge bases in other nodes, or to external knowledge bases like DBpedia [19], Wikidata [20], or others in the Linked Open Data cloud [21]. Metadata is expressed using the PLATOON semantic data models, and diverse mapping rule languages (e.g., RML or SPARQL) are utilized to define each pilot datasets in terms of the semantic data models. This platform enables pilots to preserve data sovereignty, privacy, and protection of data and analytical outcomes. More importantly, it represents an alliance-driven decentralized infrastructure empowered with the components that pave the way for interoperability across the PLATOON pilots.

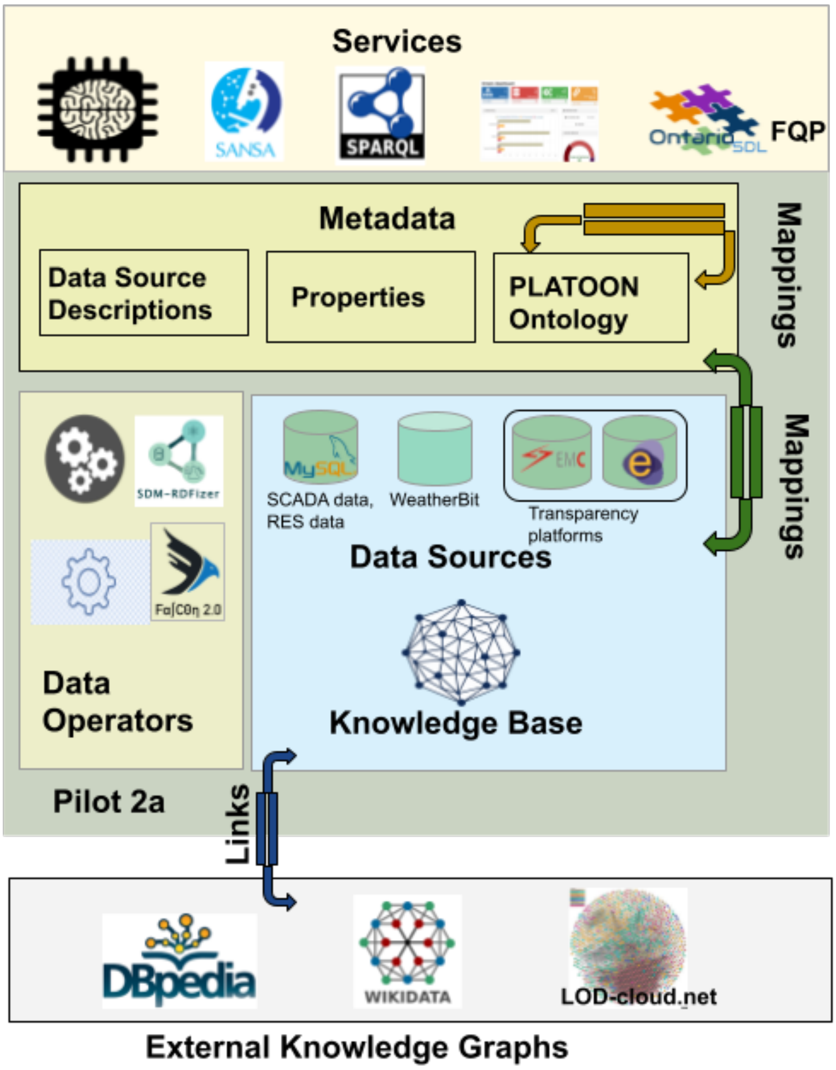

Figure 17: Example Instantiation of the PLATOON data integration platform for Pilot 2a

The main features of the PLATOON data integration platform are illustrated in the instantiation of a node as a pilot level for Pilot 2a; Figure 17 details the components. In the Serbian pilot, four main data sources are available as follows i) JSC EMS AD transparency platform; ii) ENTSO-E transparency platform; iii) Meteorological data from WeatherBit; and iv) data from SCADA system (archive data for RES production and aggregated load). The WeatherBit forecasting data is available to IMP through API. JSC EMS AD and ENTSO-E transparency data with higher granularity are available in SCADA archives maintained by IMP, while SCADA RES data is available in real time through a MySQL database internal to IMP. Data Operators for preprocessing, mapping, linking, transformation, and validation are applied to the pilot data sources for creating a materialized version of the knowledge base. Mappings between data sources and the PLATOON semantic data models (Ontology) are part of the node. Furthermore, mappings between concepts from different semantic data models are part of the node. Data sources are also described in terms of provenance and main properties; these descriptions are utilized for the creation of the knowledge base (e.g., by using SDM-RDFizer) and during query processing (e.g., by using Ontario) Entities in the pilot four data sources as well as external data sources can be done by performing entity linking. Tools like Falcon2.0 [22] can be applied to linking the pilots' datasets with external knowledge graphs like DBpedia and Wikidata. RDF data from the unified knowledge base will be fed to the Semantic based analytics engine SANSA [23]. The tools SANSA, SDM-RDFizer, Falcon2.0, and Ontario will be deployed in the pilot node at IMP. As a result, IMP will preserve autonomy and data sovereignty. Simultaneously, it will take advantage of the connections with other pilots at the level of cross-domain data sources and the usage of the PLATOON services.

6.1 The PLATOON Unified Knowledge Base Creation Pipeline

The PLATOON unified knowledge base creation pipeline receives heterogeneous data from different energy data sources and transforms these data sources into a unified knowledge graph using the PLATOON semantic data model, defined in D2.3. It applies the Semantic Data Lake approach where data is represented in raw format and defines a semantic layer to transform and integrate depending on the use cases. This pipeline is deployed at the node level in the PLATOON data integration platform shown in Figure 16. The main steps of the pipeline in the pipeline are depicted in Figure 18. First, data is ingested and preprocessed with the aim of assessing data quality, overcoming quality issues, and aggregating values. Moreover, database normalization processes may need to be performed to reduce duplicates in raw files (e.g., to transform tabular data into 3 Normal Form) and to annotate data with terms from existing ontologies (e.g., the Ontology of Units of Measure [24] or Data Quality Ontology). Next, data is enriched with the semantic data models; this semantic enrichment represents the input for knowledge integration and for linking. Existing tools for knowledge graph creation (e.g., SDM-RDFizer and SPARQL-Generate) are utilized to create the knowledge base in the RDF graph model. Engines for exploring the knowledge base or for answering queries (e.g., Ontario) provide the basis for the development of computational methods for discovery and prediction.

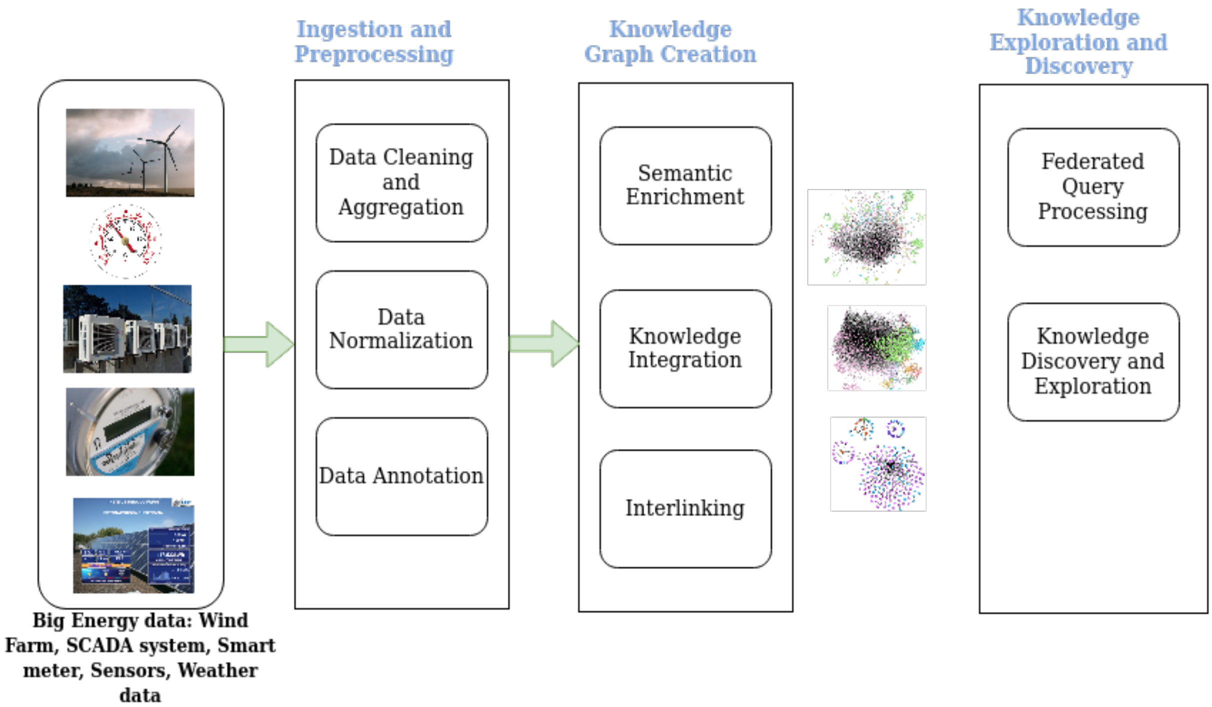

Figure 18: PLATOON Unified Knowledge Base Creation Pipeline

Data sources can be integrated using the materialized approach, where datasets are transformed to a common data representation model and stored in a data management system(s), or using the virtual approach, where data sources remain in their original format and transformed (integrated) on the fly using mappings between the original data to a common data model (see Section 7).

Various data extraction methods ingest and preprocess data in diverse formats into a Data Lake with annotations using the PLATOON data model. Data from legacy systems, IoT devices, and external web services are collected from PLATOON data source providers via IDS connector, if supported, and data connectors for external sources. Example data from the energy value chain in Serbia, i.e., from production, distribution, and forecasting is available from Institute Mihajlo Pupin (IMP) via different MySQL databases. Weather data is collected from REST APIs available from external data sources. Furthermore, data about smart buildings and their energy usage through sensors is provided through flat files and streaming services by Poste Italiane (PI). Interoperability issues will arise when data coming from different sources are integrated. Different measurement units and scales can be used as well as different data aggregation might be applied. Data normalization and standardization techniques are applied to the data ingested to the Data Lake. After data normalization and standardization, data will be annotated using common data vocabulary. Each data source has different requirements and interoperability issues that need to be handled by the **ingestion and preprocessing components**.

Once extracted data is annotated, knowledge graph creation tasks are performed to semantically describe and integrate annotated data into the PLATOON unified knowledge graph. Entity linking techniques are applied to connect equivalent entities in the knowledge graph. Moreover, rule-based mapping languages are utilized by the semantic enrichment component to create RDF triples that populate the knowledge base. The **Semantic Enrichment component** transforms annotated data into RDF; it relies on rules in a mapping language, e.g., RML, to generate the RDF triples that correspond to the input's semantic description data. The mapping rules and constraints need to be manually defined by knowledge engineers and domain experts. The PLATOON semantic data models and properties from existing RDF vocabularies like RDFS and OWL will be utilized as predicates and classes. Annotations in the input data are also represented as RDF triples. The RDF representations of these annotations are linked to the corresponding entities in the knowledge graph. Moreover, equivalences and semantic relations between annotations are represented in the knowledge graph. These relationships allow for detecting entities annotated with equivalent annotations, and that may correspond to the same real-world entities, i.e., they are duplicates; thus, equivalent annotations represent the input to the tasks of knowledge integration. Notable tools for semantic enrichment include: SDM-RDFizer, RMLMapper, and SPARQL-Generate.

The Knowledge Integration component receives an initial version of the PLATOON knowledge graph that may include duplicates, and it outputs a new version of the knowledge graph from where duplicates are removed. To detect if two entities correspond to the same real-world entity, i.e., they are duplicates, similarity measures are utilized, e.g., Jaccard; all the entities in an RDF class of the knowledge graph are compared pairwise. A 1-1 perfect weighted matching algorithm is performed to identify duplicates in the class; thus, if two entities are matched, they are considered equivalent entities and merged in the knowledge graph. Fusion policies are followed to decide how equivalent entities are merged in a knowledge graph; the fusion policies include: 1) *Union* - creates a new entity with the union of the properties of the matched entities. 2) *Semantics based Union* - creates a new entity with the union of the properties of the matched entities. Only most general properties are kept in case of properties related by the subproperty relationship; furthermore, if two properties are equivalent, only one of them is kept in the resulting entity. 3) *Authoritative Merge* - creates a new entity with the properties of the entity with the data provided from an authoritative source.

Interlinking component receives the PLATOON knowledge graph and a list of existing knowledge graphs, e.g., DBpedia or Wikidata, and outputs a new version of the PLATOON knowledge graph, where entities are linked to equivalent entities in the input knowledge graphs. Entity linking tools like Falcon [25] and DBpedia Spotlight [26] are used for linking. Additionally, link traversal techniques are performed to further identify links with other knowledge graphs.

Once a knowledge graph is created, it can be explored and traversed using a **federated query processing** engine. Additionally, data exploration and knowledge discovery services can be employed. Results of executing a federated query can be used as input of **Data Analytics or Knowledge Discovery** tasks.

6.2 Example of Illustrating Data Integration Pipeline in the context of Pilot 2a

In this section, the pipeline for creating the PLATOON unified knowledge graph is illustrated with an example related to Pilot 2a - transparency platform (and data from Germany). We suppose that the data describing the *installed power generation capacity per production type* of Germany for 2020 is received in a tabular format, e.g., CSV file, as in Table 8 below.

Production type	MW
Nuklear	8114
Wind Onshore	53184
Geothermal	4

Table 8: Input CSV Data: Installed Energy Generation Capacity per Production types of Germany in 2020

Step 1: Ingestion and preprocessing:
First, the CSV file is stored to the raw data repository and the provenance is recorded to the metadata store. If preprocessing scripts are pre-registered for this type of data, then it will be triggered. For instance, the production type column of the tabular data below contains textual names that might be represented in different languages or synonyms. In such cases, the entity linking step needs to be triggered to uniquely identify the same entities represented in different names or languages. Therefore, the first step is to find unique IDs for production types from, e.g., Wikidata, as shown in Table 9.

Production type	Production_type_ID	ID	MW
Nuklear	wd:Q12739	PL001	8114
Wind Onshore	wd:Q43302	PL002	53184
Geothermal	wd:Q3215493	PL003	4

Table 9: Entity Linking and Production Annotation: Installed Generation Capacity

Step 2: Knowledge graph creation
Once the linking of the production types is added, the next step is the knowledge graph creation. Mapping rules are defined to describe the semantic meaning of raw files. An RDF graph representing the production plants in the file is created. These RDF graphs are called simple RDF molecules or groups of RDF triples that share the same subject. RML mapping

rules are defined and executed to transform raw data into the RDF triples that comprise the resulting RDF molecules. Furthermore, these mapping rules indicate the format of the URIs of the resources that appear as subjects or objects of the RDF molecules created during their execution. In this case, two URIs are created, i.e., for the production type and country. The same process is repeated for all the RML mappings that define the RDF classes that represent the RDF classes in the PLATOON knowledge graph in terms of the available data sources.

Step 2.1: Data Transformation (Materialized vs Virtual): Mapping rules defined (Figure 19) for each dataset will be stored in a metadata store. They will be used for performing data transformation to RDF in two ways: the materialization approach and virtual integration approach (see Section 7). If the materialization approach is intended for this data source, then the semantic enrichment (or RDFization), validation, and other steps will be performed upon the pipeline's ingestion phase. On the other hand, if the virtual approach is preferred, then the mapping rules will be used during query time of the data source for RDFization, validation, and enrichment.

```
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix rml: <http://semweb.mmlab.be/ns/rml#>.
@prefix ql: <http://semweb.mmlab.be/ns/ql#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix platoon: <http://www.w3id.org/platoon/> .

<#GenerationCapacityMapping>
    rml:logicalSource [
        rml:source      "/data/entso-e/germany/2020/production_capacity_DE.csv";
        rml:referenceFormulation  ql:CSVQ;
    ];
    rr:subjectMap [
        rr:template     "http://www.plantoon.eu/GC/DE/{ID}";
        rr:class        platoon:GenerationCapacity
    ];
    rr:predicateObjectMap [
        rr:predicate    platoon:productionType;
        rr:objectMap  [
            rml:template  "{Production_type_ID}";
            rr:termType   rdf:IRI.
        ]
    ];
    rr:predicateObjectMap [
        rr:predicate    platoon:measure;
        rr:objectMap  [
            rml:reference "MW";
            rr:datatype  xsd:nonNegativeInteger
        ]
    ];
    rr:predicateObjectMap [
        rr:predicate    platoon:measureUnit;
        rr:object       "Mega Watt"
    ];
    rr:predicateObjectMap [
        rr:predicate    platoon:agg_year;
        rr:object       "2020"
    ];
    rr:predicateObjectMap [
        rr:predicate    rdfs:label;
        rr:objectMap [
            rml:reference "Production type";
            rr:language "DE"
        ]
    ];
    rr:predicateObjectMap [
        rr:predicate  platoon:country;
        rr:object     <http://www.platoon.eu/Country/Germany>
    ].
```

Data Source
Class Subject
Class Properties

Figure 19: RML Mapping Rule for Generation Capacity per Production Type CSV Data

Figure 20 shows the RDF graph after the RDFization process. It employs the mapping rules defined in Figure 6 to create three RDF molecules identified by three unique subject URIs; `pl:PL001`, `pl:PL002`, and `pl:PL003` (pl prefix refers to: platoon resource), which correspond to each unique row from annotated input CSV data in Table 9.

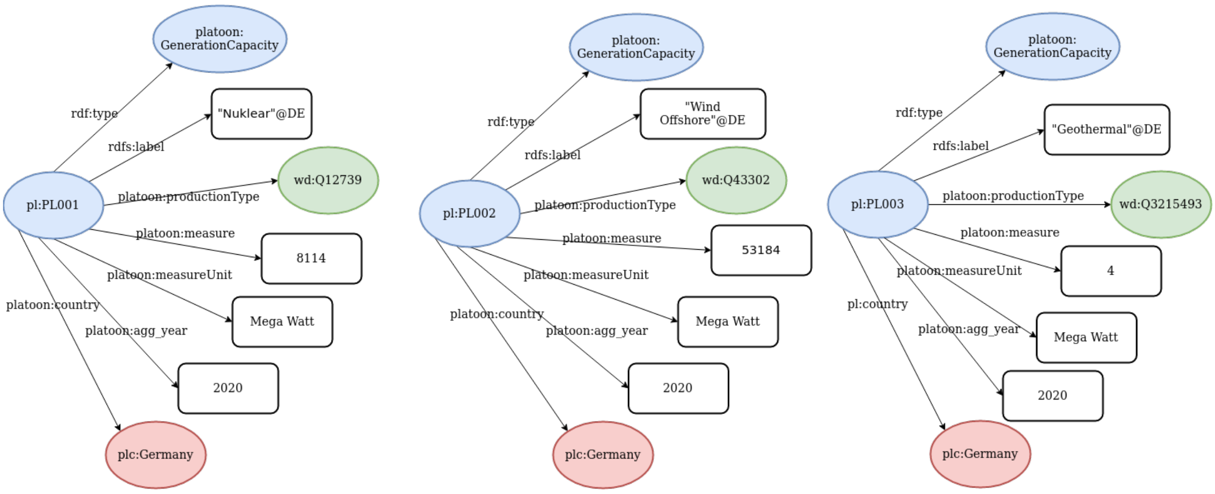

Figure 20: RDF Molecules Created by RDFizer Component for Generation Capacity per Production Type

Step 2.2: Data Validation and Constraint Checking: SHACL Constraints: The next task in the knowledge graph creation step is to validate the RDF molecules generated by the RDFization process against defined constraints (at level of data shapes). Let the constraint on the installed generation capacity measure per production type is set to be a minimum of 5 MegaWatt. Such constraints are defined using the SHACL constraint language as follows:

```
@prefix sh:       <http://www.w3.org/ns/shacl#> .
@prefix ex:       <http://example.com/ns#> .
@prefix platoon: <http://www.w3id.org/platoon/> .

ex:GenerationCapacityShape
   a                sh:NodeShape;
   sh:targetClass   platoon:GenerationCapacity;

   sh:property [
         sh:path          platoon:measure;
         sh:datatype      xsd:positiveInteger;
         sh:minInclusive 5
      ];

   sh:property [
         sh:path      platoon:measureUnit;
         sh:pattern   "Mega Watt"
      ].
```

Figure 21: Generation Capacity Shape Constraint

Validation of the data graph, containing three RDF molecules generated by the RDFization process, with respect to the shape constraint in Figure 21, yields a validation report; which reports the output of conformance checking, in Figure 22. As the report shows, production capacity of geothermal (PL003) measure does not conform to the constraint, i.e., measure value should be a minimum of 5 MegaWatts. This report guides the integration process of the RDF molecules to the unified knowledge graph. In this case, pl:PL003 RDF triples will be disregarded when integrating the RDFized data.

```
@prefix sh:      <http://www.w3.org/ns/shacl#> .
@prefix ex:      <http://example.com/ns#> .
@prefix platoon: <http://www.w3id.org/platoon/> .

[  a              sh:ValidationReport;
   sh:conforms    false;
   sh:result   [
        a                        sh:ValidationResult;
        sh:resultSeverity        sh:Violation;
        sh:focusNode             platoon:PL003;
        sh:resultPath            platoon:measure;
        sh:value                 "4"^^xsd:positiveInteger;
        sh:sourceConstraintComponent sh:MinInclusiveConstraintComponent;
        sh:sourceShape           ex:GenerationCapacityShape

      ]
]
```

Figure 22: SHACL - Generation Capacity Shape Constraint Validation Report

Step 2.3: Integrated RDF graph: Once the validation of RDF molecules created by the RDFization process is completed, the next task is enrichment and fusion of RDF molecules that conforms to the constraints. For instance, different fusion techniques can be employed to merge two or more similar RDF molecules based on semantic similarity measures with existing entities in the unified knowledge graph as well as external data sources. Figure 23 shows the integrated RDF graph of two RDF molecules of type platoon:GenerationCapacity; excluding pl:PL003 since it violates the constraint on minimum measure value of 5 MegaWatt.

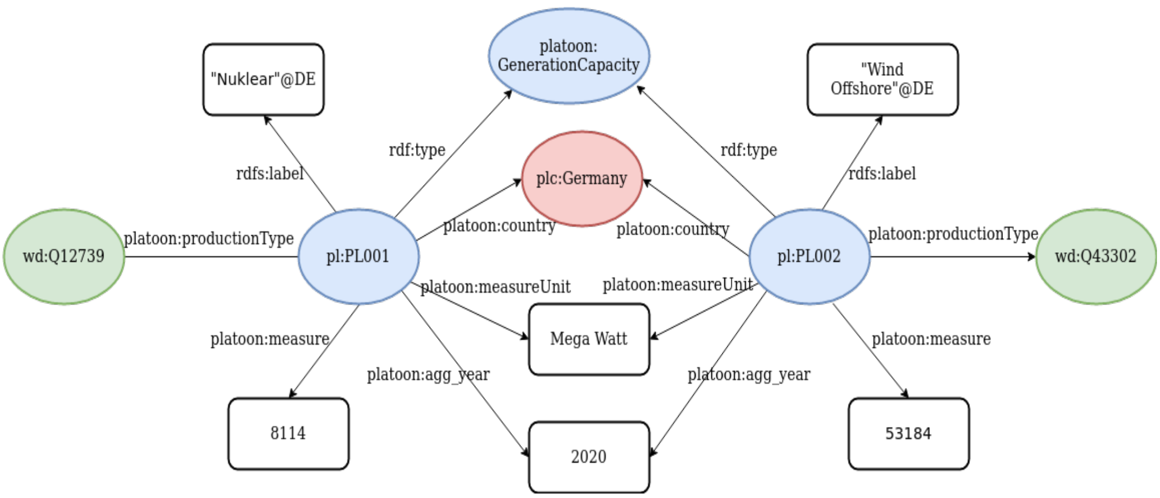

Figure 23: Integrated RDF Graph after running the Knowledge Graph Creation Pipeline

Step 3: Exploration and Discovery

The last step (Step 3) of this pipeline is to query and explore the integrated knowledge graph. Querying the integrated knowledge graph can be performed using the SPARQL query language posed over the query processing engine. Such an engine can be integrated to the data management system (e.g., Virtuoso) or can be through a federated query processing engine that is able to access data stored in centralized as well as distributed storage systems. In Section 8, we present an example of a federated query processing engine that is able to combine the output of this example (in Figure 23) with an external knowledge graph (Wikidata).

7. Knowledge Graph Creation Process

In this section, two scenarios of the knowledge graph creation process and their pros and cons are discussed.

Creating a knowledge graph from heterogeneous data sources requires the description of the entities in the data sources using RDF vocabularies, as well as the performance of curation and integration tasks in order to reduce data quality issues, e.g., missing values or duplicates. Two types of knowledge graph creation strategies: materialized (i.e., data warehousing) and virtual (i.e., Semantic Data Lake). Both strategies are applicable for the PLATOON unified knowledge base and can be deployed at different levels of the platform.

To compare the two strategies for knowledge base creation described below, we will use the following example raw data, in a MySQL database table named *building_temperature*, that describe the temperature in a building (example taken from deliverable D2.3 [2]). There are five columns (Table 10):

1. **BuildingID**: includes all the identifiers of buildings.
2. **ZoneID**: includes all the identifiers of zones.
3. **TempSensor:** includes all the names of sensors.
4. **Value C°**: includes all temperature values in degrees Celsius.
5. **Date**: includes all dates of temperature measurement.

BuildingID	ZoneID	TempSensor	Value C°	Date
1	1	S1	22	20201008:11h40
1	1	S1	21	20201008:11h50
1	2	S2	20	20201008:11h40
1	2	S2	19	20201008:11h50
2	1	S1	17	20201008:07h00
2	1	S1	20	20201008:08h00

Table 10: Datasets of temperature in a building (taken from D2.3 [2])

The mapping rule in Figure 24 is defined for this database table. The mapping rule is defined using RML language and contains a total of six TripleMaps representing RDF molecules of concepts Building, Zone, Temperature Sensor, Air Temperature Property, Air Temperature Evaluation, and Instant time. All these concepts are defined by the PLATOON data model for energy. One logical source is defined to populate data from a MySQL database called 'BUILDINGDB' and table building_temperature.

```
@prefix rr:       <http://www.w3.org/ns/r2rml#>.
@prefix rml:     <http://semweb.mmlab.be/ns/rml#>.
@prefix bot:      <http://w3id.org/bot#>.
@prefix xsd:     <http://www.w3.org/2001/XMLSchema#>.
@prefix rdfs:    <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf:      <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@base              <http://platoon.eu/mapping/base/> .
@prefix d2rq:    <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
@prefix s4bldg:  <http://saref.etsi.org/saref4bldg/> .

<#ENGIEBuildingDataset2020>   a   d2rq:Database;     d2rq:jdbcDSN "BUILDINGDB";  d2rq:jdbcDriver "com.mysql.cj.jdbc.Driver";   d2rq:username "root";    d2rq:password "" .

<#ENGIEBuildingDataset2020_BuildingMapping>
          rml:logicalSource [ rml:source       <#ENGIEBuildingDataset2020>;  rr:sqlVersion    rr:SQL2008;        rr:tableName    "building_temprature";    ];
          rr:subjectMap [     rr:template      "http://platoon.eu/resource/engie/building/{BuildingID}";          rr:class        s4bldg:Building       ];
          rr:predicateObjectMap [ rr:predicate      bot:containsZone;
                                  rr:objectMap [  rml:template   "http://platoon.eu/resource/engie/building/{BuildingID}/zone/{ZoneID}"  ]  ];
          rr:predicateObjectMap [ rr:predicate     prov:dataset;       rr:object          <#ENGIEBuildingDataset2020>   ].

<#ENGIEBuildingDataset2020_ZoneMapping>
          rml:logicalSource [     rml:source       <#ENGIEBuildingDataset2020>;   rr:sqlVersion     rr:SQL2008;    rr:tableName       "building_temprature";        ];
          rr:subjectMap [         rr:template      "http://platoon.eu/resource/engie/building/{BuildingID}/zone/{ZoneID}";            rr:class   bot:Zone          ];
          rr:predicateObjectMap [ rr:predicate      rdfs:label;          rr:objectMap [       rml:template "Zone {ZoneID}";    rr:termType   rr:Literal   ]  ];
          rr:predicateObjectMap [ rr:predicate      seas:temperature;
                                  rr:objectMap [    rml:template   "http://platoon.eu/resource/engie/building/{BuildingID}/zone/{ZoneID}/airtemperature/property"  ]      ];
          rr:predicateObjectMap [ rr:predicate     prov:dataset;     rr:object        <#ENGIEBuildingDataset2020>    ].

<#ENGIEBuildingDataset2020_TempSensorMapping>
        rml:logicalSource [  rml:source      <#ENGIEBuildingDataset2020>;    rr:sqlVersion       rr:SQL2008;      rr:tableName    "building_temprature";   ];
        rr:subjectMap [      rr:template      "http://platoon.eu/resource/engie/sensor/{TempSensor}";    rr:class     saref:TemperatureSensor      ];
        rr:predicateObjectMap [ rr:predicate    rdfs:label;   rr:objectMap [     rml:template "Sensor {TempSensor}";    rr:termType   rr:Literal      ]      ];
        rr:predicateObjectMap [ rr:predicate    s4bldg:isConnectedIn;
                               rr:objectMap [  rml:template "http://platoon.eu/resource/engie/building/{BuildingID}/zone/{ZoneID}"  ]  ];
        rr:predicateObjectMap [ rr:predicate    ssn:measures;
                               rr:objectMap [  rml:template   "http://platoon.eu/resource/engie/building/{BuildingID}/zone/{ZoneID}/airtemperature/property";  ]     ];
        rr:predicateObjectMap [rr:predicate    prov:dataset;  rr:object         <#ENGIEBuildingDataset2020>     ].

<#ENGIEBuildingDataset2020_TempSensorPropertyMapping>
        rml:logicalSource [    rml:source       <#ENGIEBuildingDataset2020>;  rr:sqlVersion     rr:SQL2008;       rr:tableName    "building_temprature";     ];
        rr:subjectMap [        rr:template   "http://platoon.eu/resource/engie/building/{BuildingID}/zone/{ZoneID}/airtemperature/property";  rr:class  platoon:AirTemperatureProperty  ];
        rr:predicateObjectMap [ rr:predicate      rdfs:label;         rr:objectMap [       rml:template "Zone {ZoneID} Temperature";    rr:termType   rr:Literal      ]      ];
        rr:predicateObjectMap [ rr:predicate      saref:isPropertyOf;  rr:objectMap [      rml:template "http://platoon.eu/resource/engie/building/{BuildingID}/zone/{ZoneID}"  ]  ];
        rr:predicateObjectMap [ rr:predicate      seas:evaluation;
                               rr:objectMap [      rml:template      "http://platoon.eu/resource/engie/building/{BuildingID}/zone/{ZoneID}/airtemperature/evaluation/{Date}";   ] ];
        rr:predicateObjectMap [ rr:predicate     prov:dataset; rr:object         <#ENGIEBuildingDataset2020>     ].

<#ENGIEBuildingDataset2020_TempSensorPropertyEvaluationMapping>
        rml:logicalSource [    rml:source       <#ENGIEBuildingDataset2020>;  rr:sqlVersion     rr:SQL2008;       rr:tableName    "building_temprature";     ];
        rr:subjectMap [  rr:template   "http://platoon.eu/resource/engie/building/{BuildingID}/zone/{ZoneID}/airtemperature/evaluation/{Date}";  rr:class   platoon:AirTemperatureEvaluation ];
        rr:predicateObjectMap [    rr:predicate  rdfs:label;     rr:objectMap [       rml:template "Zone {ZoneID} Temperature Evaluation on {Date}";    rr:termType   rr:Literal    ]      ];
        rr:predicateObjectMap [    rr:predicate     seas:evaluatedSimpleValue;     rr:objectMap [       rml:reference "Value C0"; rr:datatype xsd:integer ]  ];
        rr:predicateObjectMap [    rr:predicate     qudt:unit;   rr:object unit:DEG_C];
        rr:predicateObjectMap [    rr:predicate    seas:hasTemporalContext; rr:objectMap [   rr:parentTriplesMap <#ENGIEBuildingDataset2020_TempSensorPropertyEvaluationContextMapping>; ] ];
        rr:predicateObjectMap [    rr:predicate     prov:dataset;  rr:object         <#ENGIEBuildingDataset2020>     ].

<#ENGIEBuildingDataset2020_TempSensorPropertyEvaluationContextMapping>
        rml:logicalSource [     rml:source       <#ENGIEBuildingDataset2020>;  rr:sqlVersion     rr:SQL2008;       rr:tableName    "building_temprature";     ];
        rr:subjectMap [         rr:template   "_:{BuildingID}zone{ZoneID}airtemperatureEval{Date}";     rr:class         time:Instant; rr:termType rr:BlankNode      ];
        rr:predicateObjectMap [ rr:predicate      time:inXSDDateTime;  rr:objectMap [      rml:reference "Date"; rr:datatype  xsd:datetime      ]      ];
        rr:predicateObjectMap [ rr:predicate     prov:dataset;        rr:object         <#ENGIEBuildingDataset2020>     ].
```

Figure 24: RML mapping rules for representing building temperature table to PLATOON data model

7.1 Materialized Knowledge Graph Creation Process

In a materialized knowledge graph creation process, data from individual data sources are loaded and materialized into an RDF format and stored in a physical database, the so-called triplestore. Figure 25 below shows the data curation and integration sub-components for creating the PLATOON unified knowledge graph. The ingestion and preprocessing component is the gateway to the knowledge graph creation process. Input data from PLATOON data sources first will be stored in a raw data repository, i.e., staging repository. Any preprocessing steps, such as cleaning, normalization, and aggregation, that are predefined for input data are applied and provenance is recorded. The data integrator component then orchestrates the knowledge graph creation process according to the data source's configuration by invoking the Linking and Enrichment, RDFizer/Semantifier, and Data Validation sub-components and finally integrating data to the PLATOON unified knowledge graph. The Linking and Enrichment component performs entity linking and enrichment using external as well as existing materialized knowledge graphs. The RDFizer/Semantifier component transforms non-semantic, i.e., raw, data to RDF graph based on mapping rules. Data validation component checks data constraint conformance.

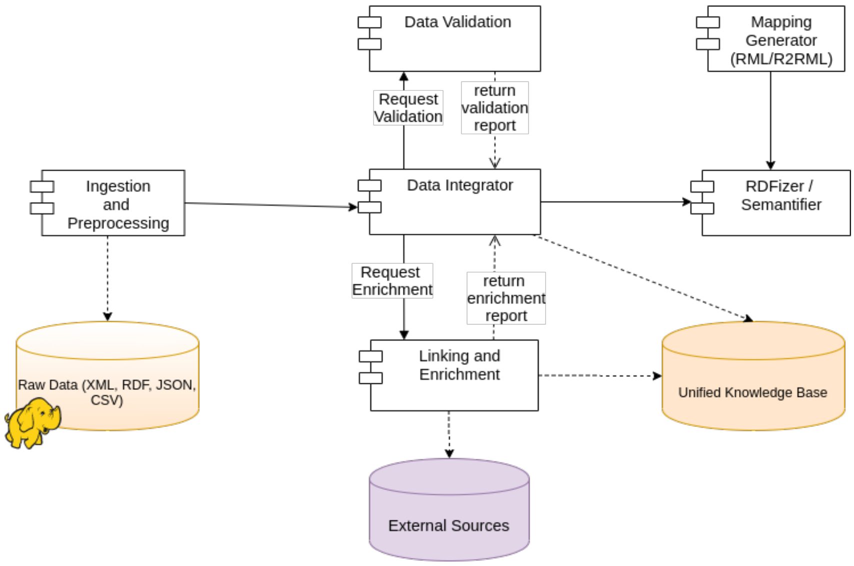

Figure 25: Knowledge Graph Creation Process

Applying the materialized knowledge base creation approach utilizes the RML mapping rules defined in Figure 24 and transforms data from relational data model in '`building_temperature`', Table 10, to RDF data. The result of this transformation will give the knowledge graph (part of it) shown in Figure 26.

```
@prefix bot:        <http://w3id.org/bot#>.
@prefix xsd:       <http://www.w3.org/2001/XMLSchema#>.
@prefix rdfs:      <http://www.w3.org/2000/01/rdf-schema#> .
@prefix s4bldg:  <http://saref.etsi.org/saref4bldg/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix seas: <https://www.w3.org/seas/> .
@prefix ssn: <https://www.w3.org/ns/ssn/> .
@prefix saref: <https://saref.etsi.org/core/> .
@prefix qudt: <http://www.qudt.org/2.1/schema/qudt/> .
@prefix time: <http://www.w3.org/2006/time#> .
@prefix platoon: <http://w3id.org/platoon/> .
@prefix :  <http://platoon.eu/mapping/base/#> .
@prefix engiepltr: <http://platoon.eu/resource/engie/building/> .

engiepltr:1 a s4bldg:Building ;
                    bot:containsZone engiepltr:1/zone/1, engiepltr:1/zone/2 ;
                    prov:dataset :ENGIEBuildingDataset2020 .
engiepltr:1/zone/1   a   bot:Zone ;
                     rdfs:label    "Zone 1";
                     seas:temperature   engiepltr:1/zone/1/airtemperature/property ;
                     prov:dataset :ENGIEBuildingDataset2020 .
engiepltr:1/sensor/S1   a    saref:TemperatureSensor ;
                      rdfs:label   "Sensor S1" ;
                      saref:isConnectedIn engiepltr:1/zone/1 ;
                      ssn:measures    engiepltr:1/zone/1/airtemperature/property ;
                      prov:dataset   :ENGIEBuildingDataset2020 .
engiepltr:1/zone/1/airtemperature/property  a platoon:AirTemperatureProperty ;
                      rdfs:label   "Zone 1 Temperature";
                      saref:isPropertyOf   engiepltr:1/zone/1 ;
                      seas:evaluation  engiepltr:1/zone/1/airtemperature/evaluation/2020-10-08T11_40_00Z ;
                      prov:dataset  :ENGIEBuildingDataset2020 .
engiepltr:1/zone/1/airtemperature/evaluation/2020-10-08T11_40_00Z  a  platoon:AirTemperatureEvaluation ;
                      rdfs:label "Zone 1 Temperature Evaluation on 2020-10-08T11:40:00Z" ;
                      seas:evaluatedSimpleValue  "22"^^xsd:integer ;
                      qudt:unit  unit:DEG_C ;
                      seas:hasTemporalContext  [
                              a time:Instant ;
                              time:inXSDDateTime  "2020-10-08T11:40:00Z"^^xsd:datetime ;
                              prov:dataset  :ENGIEBuildingDataset2020
                          ]
...
```

Figure 26: Representation of building temperature tabular data into RDF using PLATOON data model (partial view)

7.2 Virtual Knowledge Graph Creation Process

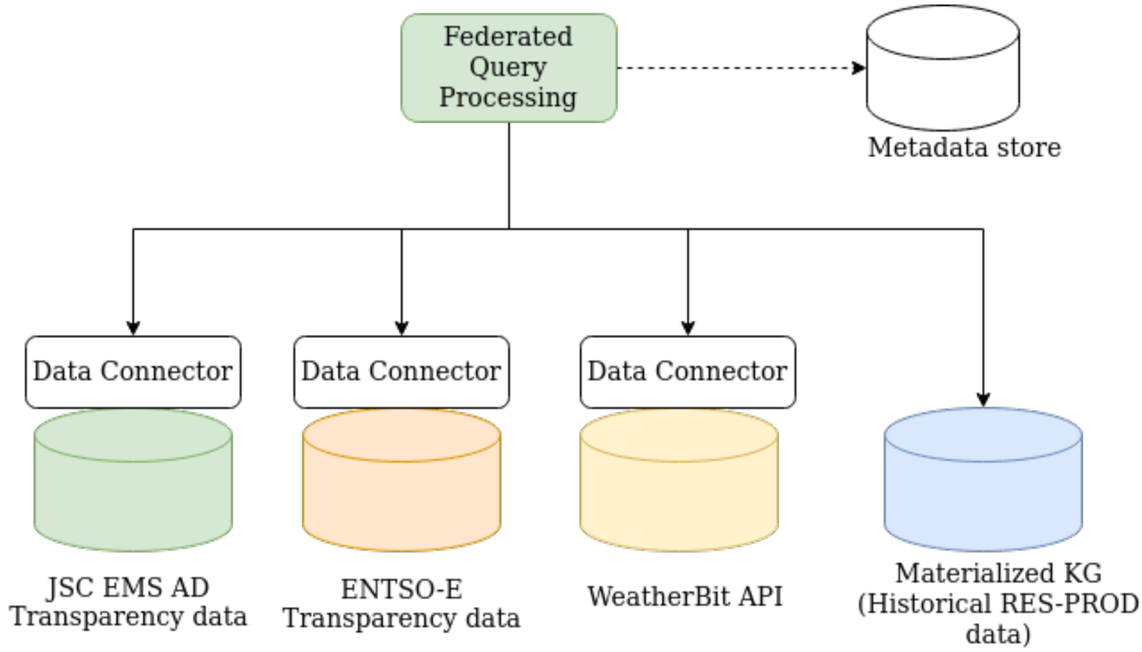

Figure 27: Federated Query Processing as Virtual Knowledge Graph Creation Process illustration using Pilot 2a data sources

In a virtual knowledge graph creation process, data remains in the sources (in raw format) and is accessed as needed during query time. The federated query processing component can handle this process. The federated query processing component employs the data source descriptions stored in the metadata store to perform the integration during query time. Metadata about the number of data sources available, the provenance of the datasets, and mapping rules to transform data to RDF graph are stored in a separate data store available for both materialized and virtual data integration processes. If the datasets are already included in the materialized knowledge graph, then the federated query processing component can directly access them without performing data transformation at query time. However, if the data sources are stored in raw format, then the data transformation rules will be applied only for the part of the dataset required to answer the query. Figure 27 shows the basic components of the virtual knowledge graph creation process through a federation system. The federated query processing component users will use SPARQL query language to access the unified knowledge graph, as described in Section 5.

Virtual knowledge base creation approach applied on 'BUILDINGDB.building _temperature' table receives a SPARQL CONSTRUCT query, Figure 28, and then consult the RML mapping rule defined for this dataset, in Figure 24 to transform the relational table data to RDF. The result of this CONSTRUCT query is the same as the result from the materialized approach in Figure 26. Contrary to the materialized approach, the result of the virtual integration approach is always timely data, while the result of the materialized approach needs to be updated if the source dataset is changing. On the other hand, querying the materialized version will be faster if data cleaning, linking, validation and other tasks are needed to be performed on the raw data.

```sparql
PREFIX s4bldg: <http://saref.etsi.org/saref4bldg/>
PREFIX bot: <http://w3id.org/bot#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX seas: <https://www.w3.org/seas/>
PREFIX ssn: <https://www.w3.org/ns/ssn/>
PREFIX saref: <https://saref.etsi.org/core/>
PREFIX qudt: <http://www.qudt.org/2.1/schema/qudt/>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX platoon: <http://w3id.org/platoon/>
PREFIX :  <http://platoon.eu/mapping/base/#>

CONSTRUCT
{
 #  same triple patterns as in WHERE clause
 #  omitted for readability
}
WHERE {
        ?building   a                  s4bldg:Building .
        ?building   bot:containsZone   ?zone .
        ?building   prov:dataset       :ENGIEBuildingDataset2020 .

        ?zone       a                  bot:Zone .
        ?zone       rdfs:label         ?zoneLabel .
        ?zone       seas:temperature   ?tempProp .
        ?zone       prov:dataset       :ENGIEBuildingDataset2020 .

        ?sensor     a                  saref:TemperatureSensor .
        ?sensor     rdfs:label         ?sensorLabel .
        ?sensor     s4bldg:isConnectedIn  ?zone .
        ?sensor     ssn:measures       ?tempProp .
        ?sensor     prov:dataset       :ENGIEBuildingDataset2020 .

        ?tempPorp   a                  platoon:AirTemperatureProperty .
        ?tempProp   rdfs:label         ?tempPropLabel .
        ?tempProp   saref:isPropertyOf ?zone .
        ?tempProp   seaas:evaluation   ?evaluation .
        ?tempProp   prov:dataset       :ENGIEBuildingDataset2020 .

        ?evaluation a                          platoon:AirTemperatureEvaluation .
        ?evaluation rdfs:label                 ?evaluationLabel .
        ?evaluation seas:evaluatedSimpleValue  ?tempValue .
        ?evaluation qudt:unit                  ?tempUnit .
        ?evaluation seas:hasTemporalContext    ?context .
        ?evaluation prov:dataset               :ENGIEBuildingDataset2020 .

        ?context    a                  time:Instant .
        ?context    time:isXSDDateTime ?datetime .
        ?context    prov:dataset       :ENGIEBuildingDataset2020 .
}
```

Figure 28: SPARQL CONSTRUCT query for virtual data transformation from building temperature tabular data to RDF (body of CONSTRUCT is omitted for readability as it is similar to body of the WHERE clause)

8. Traversing the PLATOON Unified Knowledge Base

This section presents techniques for traversing the PLATOON unified knowledge base. Once the knowledge graph creation process is established, exploring the knowledge base will be possible via a query engine. As the knowledge base is defined through mapping to semantic data models for energy, the query processing engine is able to process queries posed using the SPARQL query language. If the materialization approach is applied and data is stored in a centralized triple store, e.g., Virtuoso, then the knowledge base can be accessed using SPARQL query over the query engine embedded in the triple store. However, if the size (in terms of volume) of the materialized knowledge base is big, then partitioning and distribution is necessary for timely response from the query engine and handling the resource requirements to store such large data in expensive servers. Such distribution of data needs to be accessed through a federated query engine that is able to distribute the posed query to each partition and merge data returned from them. Virtual integration approach can also be applied over heterogeneous data sources. In this case, the query processing engine not only query each data source and merge results but also should be able to transform raw data to the semantic models specified in the mappings during query time. Below we present, Ontario, a federated query processing engine over heterogeneous data in a Semantic Data Lake. We anticipate the materialized approach could be applied on parts of the data sources from most of the pilots. The next part of this task, i.e., Task 5.3, will be able to decide per pilot basis.

8.1 Ontario: Federated Query Processing

Ontario [11] is a federated query engine that enables the exploration of the PLATOON unified knowledge base. Queries can be written in SPARQL and Ontario decides the subqueries that need to be executed over each data source to collect the data required for the query answer. Ontario executes physical operators, e.g., symmetric join and gjoin [27], and is able to combine non-RDF data with RDF triples stored in different knowledge bases (Figure 29).

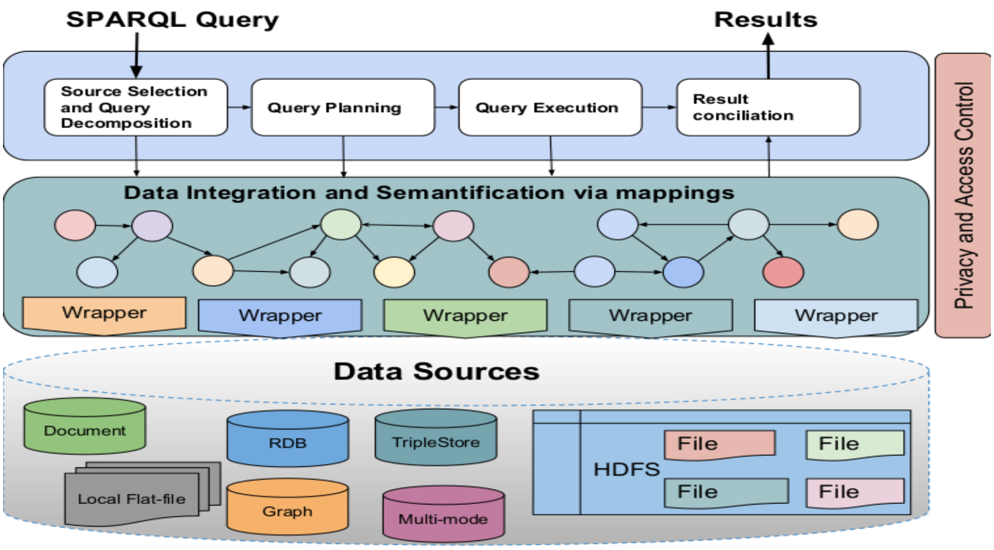

Figure 29: Ontario: Federated Query Processing over Heterogeneous Data Sources in a Data Lake

To describe heterogeneous data sources, Ontario employs **RDF Molecule Templates** (RDF-MTs), an abstract description of entities in a unified schema and their implementation in the federation of data sources. RDF-MTs describe a set of entities that belong to the same semantic concept and the relationships between them, i.e., within a data source and between different data sources. In other words, they are templates that represent a set of RDF molecules that share the same semantic concept. RDF-MTs provide a way to analyze the properties of a single data source and set of data sources in a federation, which provide an insight on how dense or sparse the connection of data elements appears in those data sources and the federation as a whole. At query time, such descriptions are consulted to answer the given query.

8.2 Example of Federated Query Processing in the context of Pilot 2a

To demonstrate the features of Ontario, as a federated query processing over multiple knowledge graphs, consider the following SPARQL query, Q, that represents the following data request: "*A list of countries, their renewable energy plants, and respective installed generation capacity for the year 2020*".

SPARQL Query Q:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt:  <http://www.wikidata.org/prop/direct/>
PREFIX platoon: <http://w3id.org/platoon/>
PREFIX pl: <http://project-platoon.eu/resource/>

SELECT DISTINCT ?country ?productionType ?measure
WHERE {
     ?genCapacity    a           platoon:GenerationCapacity .
     ?genCapacity    platoon:productionType ?productionType .
     ?genCapacity    platoon:country        ?country .
     ?genCapacity    platoon:measure   ?g_measure .
     ?genCapacity    platoon:agg_year       "2020" .
     ?productionType wdt:P279               wd:Q12705 .
}
```

To execute this query, both the PLATOON unified knowledge graph data sources and the external knowledge graphs, i.e., Wikidata, need to be consulted. Ontario maintains metadata about these knowledge graphs, represented in RDF molecule templates, and it is able to select them as relevant sources for the query. Then, once the RDF molecule templates are selected, Ontario decomposes the query into subqueries **SQ1** and **SQ2**, and executes them over the PLATOON and Wikidata knowledge graphs, respectively.

SQ1: Execute over PLATOON Unified Knowledge Graph

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX platoon: <http://w3id.org/platoon/>
PREFIX pl: <http://project-platoon.eu/resource/>
SELECT DISTINCT ?country ?productionType ?measure
WHERE {
     ?genCapacity    a            platoon:GenerationCapacity .
     ?genCapacity    platoon:productionType ?productionType .
     ?genCapacity    platoon:country        ?country .
     ?genCapacity    platoon:measure   ?g_measure .
     ?genCapacity    platoon:agg_year       "2020" .
}
```

SQ2: Execute over Wikidata Knowledge Graph

```
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt:  <http://www.wikidata.org/prop/direct/>

SELECT DISTINCT ?productionType
WHERE {
     ?productionType     wdt:P279          wd:Q12705 .
}
```

Executing **SQ1** over the PLATOON unified knowledge graph, created in Figure 8, will produce the following result:

country	productionType	measure
pl:Germany	wd:Q12739	8114
pl:Germany	wd:Q43302	53184

Executing **SQ2** over the Wikidata knowledge graph produces 80 answers (80 different renewable energy plants). The top three rows are:

productionType	productionTypeLabel
wd:Q43302	Wind power
wd:Q184037	Tidal energy
wd:Q40015	Solar energy
…	…

Ontario then performs join over two tables on `productionType` as a join column. As a result, only one renewable energy plant can be matched, i.e., the wind power:

country	productionType	measure
pl:Germany	wd:Q43302	53184

Notice that without the integration of the transparency platform data and the linking of the corresponding production types with Wikidata, this federation query could not be executed.

9. Conclusions and Next Steps

This document reports on the outcomes of performing task T2.4 – Data Integration of WP2, that started in month seven (M7) of the PLATOON project. The PLATOON data sources are characterized in terms of the 5Vs model of Big Data and interoperability issues. From the analysis using the 5Vs model of Big Data, it can be observed that the PLATOON data sources meet the characteristics of Big Data. Mainly, datasets of the pilots 1a, 2a, 2b, and 3a are large and have a high growth trend. Moreover, the datasets of the pilots 1b, 2a, 3a, 3b, 3c, and 4a are frequently updated, corroborating that the PLATOON data is in motion. Furthermore, datasets are present in diverse formats (e.g., CSV, JSON, RDB, JPEG), and various data management systems are utilized for data storage (e.g., MySQL pilot 2b). Lastly, because many of the datasets comprise data collected from multiple devices (e.g., wind turbines in pilot 1b or microgrid assets in 4a), faulty or noisy measurements may be ingested. Therefore, scalable data management and analytical tools are demanded to scale up. The data source analysis reported in this document represents a building block for elucidating the PLATOON reference architecture requirements in terms of scalability.

Additionally, interoperability conflicts are expected to be present across PLATOON datasets. The variety of the formats (e.g., CSV, JSON, RDB, JPEG), data management tools (e.g., SCADA and MySQL), and the diversity of languages (e.g., English, Spanish, Serbian, Italian, Russian) and measurement granularity (e.g., seconds, minutes, hours, months, and years) hinter interoperability during data exchange and integration. Thus, novel data management techniques are demanded to empower PLATOON data-driven components with strategies to enable the integration (materialized or virtual) of the PLATOON data sources into the PLATOON unified knowledge base. A hybrid knowledge graph creation process seems to be appropriate based on the complexity of data sources integrated into the PLATOON unified knowledge base.

The PLATOON pilots are developed incrementally, and some pilots are not still at the development level to complete the questionnaire presented in this document. The submitted questionnaires (i.e., from pilots 1a, 2a, 2b, and 3b) will be utilized as examples to guide the partners in the description and analysis of their data sources. Moreover, they will be used to illustrate the opportunities for following the PLATOON data integration platform and the benefits that it will bring in terms of data sovereignty and secure data exchange. During the next months of 2021, the T2.4 participants will organize workshops to guide the pilot owners into a more in-depth analysis of their developments. The outcomes of this collaborative work will be reported in the second version of this deliverable in month 27 (D2.4 V2 in M27).

The T2.4 participants will also utilize the workshops with the pilot owners to identify the portions of the data sources in the cross-domain of the semantic data models that will be integrated into the PLATOON unified knowledge base. Additionally, the best data integration approach (i.e., materialized and virtualized) will be discussed and selected according to the pilots' needs. The outcomes of this collaboration and the results reported in this document represent the input to T5.3 which will start in month 19 (M19) and define the techniques for data collection and harmonization.

References

[1] "PLATOON D2.1: PLATOON Reference Architecture," 2020.

[2] "PLATOON D2.3: PLATOON Common Data Models for Energy," 2020.

[3] M. Chen, S. Mao and Y. Liu, "Big data: A survey," *Mobile networks and applications,* vol. 19, pp. 171-209, 2014.

[4] H. V. Jagadish, J. Gehrke, A. Labrinidis, ros, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan and C. Shahabi, "Big data and its technical challenges," *Communications of the ACM,* vol. 57, pp. 86-94, 2014.

[5] U. Sivarajah, M. M. Kamal, Z. Irani and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *Journal of Business Research,* vol. 70, no. Elsevier, pp. 263--286, 2017.

[6] M. I. S. Oliveira and B. F. Loscio, "What is a data ecosystem?," in *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, 2018, pp. 1-9.

[7] C. Capiello, A. Gal, M. Jarke and J. Rehof, "Data Ecosystems: Sovereign Data Exchange among Organizations (Dagstuhl Seminar 19391)," in *Dagstuhl Reports*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2020.

[8] S. Bader, J. Pullmann, C. Mader, S. Tramp, C. Quix, A. W. Muller, H. Akyurek, M. Bockmann, B. T. Imbusch, J. Lipp and others, "The International Data Spaces Information Model--An Ontology for Sovereign Exchange of Digital Content," in *International Semantic Web Conference*, Springer, 2020, pp. 176--192.

[9] A. Doan, A. Halevy and Z. Ives, Principles of Data Integration, Elsevier, 2012.

[10] E. Iglesias, S. Jozashoori, D. Chaves-Fraga, D. Collarana and M.-E. Vidal, "SDM-RDFizer: An RML interpreter for the efficient creation of rdf knowledge graphs," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 3039--3046.

[11] K. M. Endris, P. D. Rohde, M.-E. Vidal and S. Auer, "Ontario: Federated Query Processing against a Semantic Data Lake," in *International Conference on Database and Expert Systems Applications*, Linz, Springer, 2019, pp. 379--395.

[12] R. Bellazzi, "Big data and biomedical informatics: a challenging opportunity," *Yearbook of medical informatics,* vol. 9, 2014.

[13] "R2RML," [Online]. Available: https://www.w3.org/TR/r2rml/.

[14] A. Dimou, V, e. S, M. e, P. Colpaert, R. Verborgh, E. Mannens and R. Van de Walle, "RML: a generic language for integrated RDF mappings of heterogeneous data," in *Linked Data on the Web (LDOW)*, 2014.

[15] F. Michel, L. Djimenou, C. F. Zucker and J. Montagnat, "xR2RML: Relational and non-relational databases to RDF mapping language," 2017.

[16] M. Lefrancois, A. Zimmermann and N. Bakerally, "A SPARQL extension for generating RDF from heterogeneous formats," in *European Semantic Web Conference (ESWC)*, Springer, 2017, pp. 35-50.

[17] "SHACL," [Online]. Available: https://www.w3.org/TR/shacl/.

[18] "the Energy Act, which transposes EU Regulation no. 543/2013," [Online]. Available: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:163:0001:0012:EN:PDF.

[19] "DBpedia," [Online]. Available: https://wiki.dbpedia.org/.

[20] "Wikidata," [Online]. Available: https://www.wikidata.org/.

[21] "LOD-Cloud," [Online]. Available: https://lod-cloud.net/.

[22] A. Sakor, I. O. Mulang, o, K. Singh, S. Shekarpour, M. E. Vidal, J. Lehmann and S. Auer, "Old is gold: linguistic driven approach for entity and relation linking of short text," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 2336--2346.

[23] J. Lehmann, G. Sejdiu, L. Buhmann, P. Westphal, C. Stadler, I. Ermilov, S. Bin, N. Chakraborty, M. Saleem, A.-C. N. Ngomo and others, "Distributed semantic analytics using the sansa stack," in *International Semantic Web Conference (ISWC)*, Springer, 2017, pp. 147-155.

[24] "Ontology of Units of Measure," [Online]. Available: https://enterpriseintegrationlab.github.io/icity/OM/doc/index-en.html.

[25] "Falcon2.0," [Online]. Available: https://labs.tib.eu/falcon/.

[26] "DBpedia Spotlight," [Online]. Available: https://www.dbpedia-spotlight.org/.

[27] M. Acosta, M.-E. Vidal, T. Lampo, J. Castillo and R. Edna, "ANAPSID: an adaptive query processing engine for SPARQL endpoints," in *International Semantic Web Conference*, Springer, 2011, pp. 18-34.

[28] M. Lenzerini, "Data integration: A theoretical perspective," in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002, pp. 233--246.

Appendix A. Summary of PLATOON Data Source Description with respect to the 5V's of Big Data

Pilot 1a – Predictive Maintenance for Wind Farms

Data Source Title	Wind turbine SCADA data	Acronym	
Type	Sensor Data	Provided by	ENGIE
Description	The data set contains sensors data of the Supervisory Control and Data Acquisition system of a wind turbine. This system contains sensors at the most important subcomponents of the wind turbine		
Volume	Probably in the Gb range. Exact size not clear yet		
Velocity	10-min averages and statistics		
Variety	Typically coming from multiple turbines. Each turbine brand has its own data structure and tag names		
Veracity	Data is not clean. Unrealistic values can be present as well as missing data		
Value	Pilot 1a all use cases		
	Data should contain turbines for periods where faults happened to the electrical subcomponents		

Table 11 Wind turbine SCADA data

Data Source Title	High-frequency Data	Acronym	
Type	Production (Generation) data	Provided by	ENGIE
Description	High frequency electric measurements and a limited set of turbine operational parameters (e.g., wind speed). This data originates from a dedicated measurement campaign on onshore turbines		
Volume	In Tb		
Velocity	All months of data were collected at 500Hz while for two weeks collection was done at 5kHz.		
Variety	Data is sampled using the same data acquisition system so all sensors are in the same file		
Veracity	Uncleaned data		
Value	Pilot 1a use cases		
	Covers only one year data		

Table 12 High-frequency data

Data Source Title	Open wind speed data	Acronym	

Type	Sensor data	**Provided by**	Vlaamse Meetbanken
Description	The data set contains data of environmental measurements (wind speeds, wind directions, wave heights…) across the Belgian North Sea.		
Volume	In Gb		
Velocity	10-min averages and statistics		
Variety	Typically, coming from multiple measurement locations. Data is standardized to one format.		
Veracity	Uncleaned data		
Value	Pilot 1a semantic reasoning		

Table 13 Open wind speed data

Data Source Title	**Offshore measurement campaign**	**Acronym**	
Type	Sensor data	**Provided by**	VUB
Description	The data set contains sensors data of accelerometers placed on the drivetrain of an offshore wind turbine.		
Volume	In Tb		
Velocity	20kHz measurements		
Variety	Acceleration signals of 10 accelerometers and 2 encoders		
Veracity	Uncleaned data		
Value	In the end not used since new current data is collected during dedicated measurements campaign which suits the project goals better.		

Table 14 Offschore measurement campaign

Data Source Title	**Dedicated current measurement campaign**	**Acronym**	
Type	Sensor data	**Provided by**	VUB-ENGIE
Description	Data of current sensors on one French onshore wind turbine is collected using the VUB data acquisition and edge processing hardware		
Volume	In Tb		

Velocity	20kHz and some channels 1Hz measurements
Variety	Signals of current probes as well as turbine controller parameters (collected at lower frequency (1Hz))
Veracity	Uncleaned data
Value	Pilot 1a edge case

Table 15 Dedicated current measurement campaign

Pilot 2a –Electricity Balance and Predictive Maintenance

Data Source Title	**Transparency PLATFORM Transmission Data**	**Acronym**	
Type	Transmission Data	**Provided by**	Joint Stock Company EMS AD
Description	Data about power transfer over between areas. It provides the current day, day ahead, month ahead, and year ahead data (no historical data), in the electric market of Serbia.		
Volume	Power transfer from different countries in the EU described in hourly and daily basis; from KBs to MBs of data per day.		
Velocity	Hourly, daily, monthly and yearly basis		
Variety	Heterogeneous data format and representations including: data format 1) XML, 2) CSV and 3) JSON; languages 1) EN and 2) RS		
Veracity	The platform is under development and data might not be published within the deadline (hourly or daily on time).		
Value	Relevant for KPI-2 – Saving from tertiary reserve trading, and Use Case #2a Electricity Balancing and Predictive Maintenance; balancing on regional and country level.		

Table 16 Transparency Platform Transmission Data

Data Source Title	**Transparency PLATFORM Consumption (LOAD) Data**	**Acronym**	
Type	Consumption (LOAD) Data	**Provided by**	Joint Stock Company EMS AD
Description	Historic data about power consumption (System vertical load from Oct 2016) in the electric market of Serbia.		
Volume	Power consumption on an hourly and daily basis; from KBs to MBs of data per day.		
Velocity	Hourly, daily, monthly and yearly basis		
Variety	Heterogeneous data format and representations including: data format 1) XML, 2) CSV and 3) JSON; languages 1) EN and 2) RS		
Veracity	The platform is under development and data might not be published within deadline (hourly or daily on time).		
Value	Relevant for KPI-2 – Saving from tertiary reserve trading, and Use Case		

	#2a Electricity Balancing and Predictive Maintenance; balancing on regional and country level.

Table 17 Transparency Platform Consumption (LOAD) Data Source

Data Source Title	Transparency PLATFORM Balancing (Load forecast) Data	Acronym	
Type	Balancing (Load forecast) Data	Provided by	Joint Stock Company EMS AD
Description	Data about regular energy used to keep the electricity transmission grid in balance.		
Volume	Power consumption forecasts described in min and max values on an hourly and daily basis; from KBs to MBs of data per day.		
Velocity	Weekly (since June 2013)		
Variety	Heterogeneous data format and representations including: data format 1) XML, 2) CSV and 3) JSON; languages 1) EN and 2) RS		
Veracity	The platform is under development and data might not be published within the deadline (hourly or daily on time).		
Value	Relevant for KPI-2 – Saving from tertiary reserve trading, and Use Case #2a Electricity Balancing and Predictive Maintenance; balancing on regional and country level.		

Table 18 Transparency PLATFORM Balancing (LOAD forecast) Data Source

Data Source Title	ENTSO-E Transparency Platform Consumption (LOAD) Data	Acronym	ENTSO-E C
Type	Consumption (LOAD) Data	Provided by	ENTSO-E
Description	Power consumption data is provided freely by the ENTSO-E Transparency Platform to pan-European electricity market data for all users.		
Volume	Power consumption on an hourly and daily basis; from KBs to MBs of data per day.		
Velocity	Hourly, daily, and yearly (since 2015)		
Variety	Heterogeneous data format and representations including: data format 1) XML, 2) CSV; language EN		
Veracity	Missing values in cases the country has not provided it.		
Value	Relevant for KPI-2 – Saving from tertiary reserve trading, and Use Case #2a Electricity Balancing and Predictive Maintenance; balancing on regional and country level.		

Table 19 ENTSO-E Transparency Platform Consumption (LOAD) Data

Data Source Title	ENTSO-E Transparency Platform Generation Data	Acronym	ENTSO-E G
Type	Generation (Production) Data	Provided by	ENTSO-E

Description	Energy production and production forecasts data provided by ENTSO-E Transparency platform. It provides the installed capacity, actual generation and generation forecasts per generation unit.
Volume	Power generation data and forecast in hourly and daily basis; from KBs to MBs of data per day.
Velocity	Hourly, daily, and yearly (since 2015)
Variety	Heterogeneous data format and representations including: data format 1) XML, 2) CSV; language EN
Veracity	Missing values in cases the country has not provided it
Value	Relevant for KPI-2 – Saving from tertiary reserve trading, and Use Case #2a Electricity Balancing and Predictive Maintenance; balancing on regional and country level.

Table 20 ENTSO-E Transparency Platform Generation Data

Data Source Title	**ENTSO-E Transparency Platform Transmission Data**	**Acronym**	**ENTSO-E T**
Type	Transmission Data	**Provided by**	ENTSO-E
Description	Data about power transfer over borders between areas. It provides scheduled commercial exchanges, cross-border physical flows, day ahead, week ahead, month ahead, and year ahead data.		
Volume	Power transfers data and forecasts on an hourly and daily basis; from KBs to MBs of data per day.		
Velocity	Hourly, daily, and yearly (since 2015)		
Variety	Heterogeneous data format and representations including: data format 1) XML, 2) CSV; language EN		
Veracity	Missing values in cases the country has not provided it		
Value	Relevant for KPI-2 – Saving from tertiary reserve trading, and Use Case #2a Electricity Balancing and Predictive Maintenance; balancing on regional and country level.		

Table 21 ENTSO-E Transparency Platform Transmission Data

Data Source Title	**ENTSO-E Transparency Platform Balancing (LOAD forecast) Data**	**Acronym**	**ENTSO-E B**
Type	Balancing (Load forecast) Data	**Provided by**	ENTSO-E
Description	Data about Regulation energy used to keep the electrical transmission grid in balance. It provides general rules on balancing, energy bids, capacity and imbalances.		
Volume	in KBs per min, hour, daily		
Velocity	per minute		
Variety	CSV		
Veracity	missing values		
Value	Use Case #2a Electricity Balancing and Predictive Maintenance; balancing on regional and country level.		

Table 22 ENTSO-E Transparency Platform Balancing (LOAD forecast) Data

Data Source Title	ENTSO-E Transparency Platform Outages Data	Acronym	ENTSO-E O
Type	Outages Data	Provided by	ENTSO-E
Description	Data about planned maintenance and failures inside the electricity transmission grid provided by ENTSO-E Transparency Platform. It provides data about unavailability in transmission, offshore, production and generation units.		
Volume	in KBs		
Velocity	variable time		
Variety	CSV		
Veracity	Missing values		
Value	Use Case #2a Electricity Balancing and Predictive Maintenance; balancing on regional and country level.		

Table 23 ENTSO-E Transparency Platform Outages Data

Data Source Title	SLTF – Short Time Load Forecast Data	Acronym	SLTF
Type	Balancing (Load Forecast) Data	Provided by	IMP (owner Joint Stock Company EMS AD)
Description	Short time load forecast datasets needed for LLUC P-2a-03 load demand forecast on transmission level.		
Volume	Data volume in MBs.		
Velocity	Hourly		
Variety	Heterogeneous data formats (XML and CSV export from SCADA). Data language in RS.		
Veracity	Missing values.		
Value	KPI-1 cost efficient distribution and transmission; KPI-3 better demand response. Exploited for demand forecasting model training.		

Table 24 SLTF – Short Time Load Forecast Data

Data Source Title	MET-RES - Meteorological Data for RES Production (Generation) Forecasting Modelling Data	Acronym	MET-RES
Type	Generation (Production) forecast data	Provided by	IMP (owner WeatherBit)
Description	Meteorological dataset that will be utilized for RES production forecasting models training process as input data. Data is historical data (no update will be needed)		
Volume	Approximately ~200 MB		
Velocity	Weather parameters are obtained with the hourly time resolution		

Variety	Data is organized in tables of CSV files and EN language.
Veracity	Potential missing values.
Value	Used in RES production forecasting model training process; LLUC P-2a-04
	Data will not be stored as part of PLATOON knowledge base

Table 25 MET-RES - Meteorological Data for RES Production (Generation) Forecasting Modelling Data

Data Source Title	RES-PROD - Historical Wind Power Production Measurements	Acronym	RES-PROD
Type	Generation (Production) Data	Provided by	IMP
Description	Contains measurements of the production from the wind power plant.		
Volume	Approximately ~30 MB		
Velocity	Hourly time resolution (not streaming any further – is now historic)		
Variety	Data is organized in tables of CSV files and EN language.		
Veracity	Missing measurements		
Value	Used in RES production forecasting model training process; LLUC P-2a-04		
	Data will not be stored as part of PLATOON knowledge base		

Table 26 RES-PROD - Historical Wind Power Production Measurements

Data Source Title	Effects of Renewable Energy Sources on the Power System (distribution level)	Acronym	RES Effects
Type	Effects of renewable energy sources on the power system	Provided by	CS (data owner IMP)
Description	-		
Volume	10 KB per 15 min		
Velocity	Each 15 min a report is generated by edge computing unit		
Variety	XML data format and languages in EN and RS		
Veracity	Missing data from meter		
Value	Electricity balance and predictive maintenance (LLUC P-2a-05 effects of renewable energy sources on the power system)		

Table 27 Effects of Renewable Energy Sources on the Power System (distribution level)

Data Source Title	RES PV Predictive Maintenance	Acronym	RES-PV
Type	Predictive Maintenance	Provided by	CS (data owner IMP)
Description	Data will be collected when the PMU is installed at IMP side.		
Volume	5 KB per second		
Velocity	Every second		
Variety	JSON data format and in EN language		
Veracity	Missing data		

Value	Electricity balance and predictive maintenance (LLUC P-2a-07 predictive maintenance in RES power plants)

Table 28 RES PV Predictive Maintenance

Pilot 2b - - Electricity Grid Stability, Connectivity, and Life Extension

Data Source Title	Power grid ZIV power meters	Acronym	
Type	Power Meter	Provided by	SAMPOL
Description	Hourly measurements of active and reactive power delivered to the users, grouped by concentrator and identified by power meter.		
Volume	300 MB		
Velocity	1 value each hour for 77 power meters		
Variety	Relational Table (MySQL), Languages: EN and ES		
Veracity	Not known		
Value	-		

Table 29 Power grid ZIV power meters

Data Source Title	Transformer sensors	Acronym	
Type	Observation data	Provided by	SAMPOL
Description	8 temperature sensors located at different positions of the transformers, 2 sensors for ambient temperature, humidity and pressure, 1 sensor for oil temperature		
Volume	60MB		
Velocity	Values are received every 5 minutes		
Variety	Relational Table (MySQL)		
Veracity	Hight		
Value	Temperature values read at the power transformer and power transformer center.		
Comment	Those devices will be installed probably in January 2021		

Table 30 Transformer sensors

Data Source Title	Medium voltage Network analyzer	Acronym	
Type	Observational Data	Provided by	SAMPOL
Description	Electrical Network analyzer for current transformers, not yet installed		
Volume	60MB		
Velocity	Values are received every 5 minutes		
Variety	Relational Table (MySQL); Languages: EN and ES		
Veracity	Hight		

Value	Medium voltage values measured at each power transformer
Comment	This device will be installed probably in January 2021

Table 31 Medium voltage Network analyzer

Pilot 3b - Advanced Energy Management System and Spatial (Multi-scale) Predictive Models in the Smart City

Pilot #3b_PI

Data Source Title	Building Data	Acronym	ANAG
Type	Buildings data	Provided by	Poste Italiane SPA
Description	Detailed data about each building characteristics and general (ID Office, address, destination use, smq, climate zone, etc). It will contain info regarding 'Historical Data Line consumptions Coefficient', i.e., the esteemed ratio of consumption due to the specific lines (cooling, heating, lighting)		
Volume	(30) KB		
Velocity	One shot, if changes occur The database does not increase regularly.		
Variety	Excel tables		
Veracity	High		
Value	ALL KPIs in LLUC-3b LLUC P-3b-01 Buildings Heating and Cooling consumption analysis and Forecast LLUC P-3b-02 Predictive maintenance of cooling and heating plats LLUC P-3b-03 Lighting Consumption Estimation and Benchmarking		
	This source is used as a common base and reference for all use cases.		

Table 32 Building Master Data Source

Data Source Title	Calendar	Acronym	CALE
Type	Calendar data	Provided by	Poste Italiane SPA
Description	Information on office openings and shifts.		
Volume	1.2 MB/Year		
Velocity	Daily		
Variety	CSV and Italian		
Veracity	High		
Value	ALL KPIs in LLUC-3b LLUC P-3b-01 Buildings Heating and Cooling consumption analysis and Forecast LLUC P-3b-02 Predictive maintenance of cooling and heating plats LLUC P-3b-03 Lighting Consumption Estimation and Benchmarking		

Table 33 Calendar Data Source

Data Source Title	**Customers Occupancy**	Acronym	**OCCU_C**
Type	Occupancy/population data	Provided by	Poste Italiane SPA
Description	Information on numbers of customers in the building		
Volume	2 MB/year		
Velocity	Daily		
Variety	CSV		
Veracity	High (Potential Missing values)		
Value	OCCU_C: KPI - PI_03_K01 LLUC P-3b-01 Buildings Heating and Cooling consumption analysis and Forecast		

Table 34 Customers Occupancy Data Source

Data Source Title	**Employees Occupancy**	Acronym	**OCCU_E**
Type	Occupancy/population data	Provided by	Poste Italiane SPA
Description	Information on numbers of employees in the building		
Volume	2 MB/year		
Velocity	Daily		
Variety	CSV and languages in IT		
Veracity	High		
Value	ALL KPIs in LLUC P-3b-01 Buildings Heating and Cooling consumption analysis and Forecast, LLUC P-3b-03 Lighting Consumption Estimation and Benchmarking		

Table 35: Employees Occupancy Data Source

Data Source Title	**Energy Data Consumption on building and internal climate information**	Acronym	**EC_TOT**
Type	Consumption (LOAD) Data	Provided by	Poste Italiane SPA
Description	Information on building (total) active energy consumptions (kWh) in Multi Distr Buildings , DL 102 Buildings and Smart Buildings		
Volume	EC_TOT: Start up 40 MB -10 MB/YEAR		
Velocity	EC_TOT: Hour		
Variety	CSV and language in IT		
Veracity	Medium		
Value	EC_TOT: PI_KPI01 - PI_KPI02 - PI_KPI03 – PI_KPI06 LLUC P-3b-01 Buildings Heating and Cooling consumption analysis and Forecast, LLUC P-3b-03 Lighting Consumption Estimation and Benchmarking		

Table 36 Energy Data Consumption on building and internal climate information

Data Source Title	**Building Systems (or System Registry)**	**Acronym**	**BS**
Type	Plants data	**Provided by**	Poste Italiane SPA
Description	Information on kind and characteristics of heating,cooling and lighting plants of all Buildings		
Volume	1,5 MB		
Velocity	One shot		
Variety	XLSX and language in IT		
Veracity	High		
Value	All KPIS in LLUC P-3b-01 Buildings Heating and Cooling consumption analysis and Forecast, LLUC P-3b-03 Lighting Consumption Estimation and Benchmarking		

Table 37 Datan Building Systems characteristics

Data Source Title	**Energy Data Consumption**	**Acronym**	**EC_SB**
Type	Consumption (LOAD) data	**Provided by**	Poste Italiane SPA
Description	Information on active energy consumption (kWh) both of line or type of system and internal temperature and humidity (for Smart Buildings) Line: cooling, heating, lighting		
Volume	500 MB /YEAR		
Velocity	Fifteen Minutes		
Variety	CSV and languages in IT		
Veracity	Medium		
Value	All KPIS in LLUC P-3b-01 Buildings Heating and Cooling consumption analysis and Forecast, LLUC P-3b-02 Predictive maintenance of cooling and heating plats, LLUC P-3b-03 Lighting Consumption Estimation and Benchmarking		

Table 38 Energy Data Consumption

Data Source Title	**System Anomalies**	**Acronym**	**ANOMALIES**
Type	Fault & Plant	**Provided by**	Poste Italiane SPA
Description	Information on anomalies occurred to the heating and cooling plants. In particular, they are alarms of abnormal behaviour of the systems		

Volume	400 MB/YEAR
Velocity	Daily
Variety	XLSX and language in IT
Veracity	High
Value	PI_KPI05* LLUC P-3b-02 Predictive maintenance (Anomaly Detection) of cooling and heating plants.

Table 39 System Anomalies

Pilot #3b_ROM

Data Source Title	Energy Meters Electrical Monthly Consumptions for ROM buildings	Acronym	EMEMC
Type	Consumption Data	Provided by	ROMA CAPITALE + RISORSE PER ROMA (TLP)
Description	Last month consumptions from all power meters (energy vendor)		
Volume	Data volume in MBs		
Velocity	30 days (data is delivered to ROM each month (previous complete month)		
Variety	CSV		
Veracity	Missing data (only 575 meters out of 6500 total meters)		
Value	Pilot 3b – ROM use cases		
	Downloadable from MYENEL (Vendor) portal		

Table 40 Energy Meters Electrical Monthly Consumptions

Data Source Title	Energy Meters Electrical Historical Consumptions for ROM buildings 1	Acronym	EMEHC1
Type	Consumption Data	Provided by	ROMA CAPITALE + RISORSE PER ROMA (TLP)
Description	Historical electric consumptions for ROM buildings of data from 1-1-2018 to 30-06-2020, that contains around 517,598 records for daily kwh; 96 columns for every 15 min consumption.		
Volume	428 MB, (total of 2.5-year data)		
Velocity	Monthly (updated monthly by vendor ENEL (based on ARETI data))		
Variety	CSV		
Veracity	missing data (only 575 meters out of 6500 total meters)		
Value	Pilot 3b – ROM use cases		
	Downloadable from MYENEL (Vendor) portal; ARETI can supply precious datasets back to 2015 for a period of 5.5 years (66 months)		

Table 41 Energy Meters Electrical Historical Consumptions for ROM buildings 1

Data Source Title	Energy Meters Electrical Historical Consumptions for ROM buildings 2	Acronym	EMEHC2

Type	Consumption Data	**Provided by**	ROMA CAPITALE + RISORSE PER ROMA (TLP)
Description	Historical electric consumptions for ROM buildings; 36 months data from 1-1-2015 to 31-12-2017 from GALA and ARETI vendors.		
Volume	Approximately ~250MB		
Velocity	static		
Variety	TXT files		
Veracity	Uncleaned Data		
Value	Pilot 3b - ROM all use cases		
	from GALA (previous vendor) and additional details are found from ARETI		

Table 42 Energy Meters Electrical Historical Consumptions for ROM buildings 2

Data Source Title	**Building master data for ROM buildings**	**Acronym**	**BMD**
Type	Building data	**Provided by**	ROMA CAPITALE (ROM)
Description	Building properties data from ROM asset management office and buildings energy audits database		
Volume	223KB		
Velocity	Static		
Variety	XLSX File		
Veracity	Uncleaned Data		
Value	Pilot 3b – ROM use cases		

Table 43 Building master data for ROM buildings

Data Source Title	**Energy Meter Gas Monthly Consumption RC Direct**	**Acronym**	**EMGMC**
Type	Consumption Data	**Provided by**	ROMA CAPITALE (ROM)
Description	Monthly consumption for RC direct Gas meters from ESTRA		
Volume	Data volume in MBs		
Velocity	Monthly		
Variety	XLSX File		
Veracity	Uncleaned Data		
Value	Pilot 3b - ROM use cases		

Table 44 Energy Meter Gas Monthly Consumption RC Direct

Data Source Title	**Energy Meter Gas Historical Consumption RC Direct**	**Acronym**	**EMGHC**
Type	Consumption Data	**Provided by**	ROMA CAPITALE

			(ROM)
Description	Historical consumption data for RC direct Gas meters from ESTRA		
Volume	19MB		
Velocity	static		
Variety	XLSX File		
Veracity	Uncleaned Data		
Value	Pilot 3b - ROM use cases		

Table 45 Energy Meter Gas Historical Consumption RC Direct

Data Source Title	Energy Meter Gas Thermal Consumption SIE3	Acronym	EMGTC
Type	Thermal Consumption Data	Provided by	ROMA CAPITALE (ROM)
Description	Thermal consumption for SIE3 Gas meters from CPL-EMF		
Volume	2.8GB		
Velocity	15min		
Variety	CSV		
Veracity	Uncleaned Data		
Value	Pilot 3b – ROM use cases		

Table 46 Energy Meter Gas Monthly Consumption SIE3

Data Source Title	Energy Meter Gas Historical Consumption SIE3	Acronym	EMGHC2
Type	Consumption Data	Provided by	ROMA CAPITALE (ROM)
Description	Historical gas consumption data for SIE3 Gas meters from CPL-EMF from November 2018 to April 2021		
Volume	~1MB		
Velocity	Static		
Variety	XLSXFile		
Veracity	-		
Value	Pilot 3b – ROM use cases		

Table 47 Energy Meter Gas Historical Consumption SIE3

Data Source Title	ROM PV production data	Acronym	RPVPD
Type	RES Production Data	Provided by	ROMA CAPITALE (ROM)
Description	Res data production from Lovato Electric system. This dataset contains the produced kWh of the installed PV plants in a set of ROM buildings divided by each district.		
Volume	Data volume in MBs.		
Velocity	15 min		
Variety	XLSX		
Veracity	Uncleaned Data		
Value	ROM - RES potentialities		

	Manual download from Lovato system

Table 48 ROM PV production data

Pilot 3c - Advanced Energy Management System and Efficiency and Predictive Maintenance In the Smart Tertiary Building Hubgrade

Data Source Title	SIMENS DESIGO 4.0	Acronym	
Type	SCADA	Provided by	GIR
Description	SCADA data: temperatures, electricity consumption, position of valves. Also, weather data and forecasts.		
Volume	1Gb approx.		
Velocity	1.5MB/day		
Variety	JSON		

Table 49 Simens Desigo 4.0

Appendix B. PLATOON Data Source Descriptions

Data Source Description Template

PLATOON Partner		Comment
Partner ID		ID of PLATOON Partner
Partner Name		Name of PLATOON Partner

Data Source		Comment
Title		Title of the data source
Alternate Title		Alternative title, if any
Acronym		Data source acronum, if exists
Description		Give short description of the dataset, e.g., purpose, type of data, etc
Temporal Coverage		If the dataset contains temporal information, provide which period it covers
Maintenance/Status		State if dataset is old or is update/maintained regularly
Other Comments		Give any additional comment, if any

Big Data Vs		Comment
Volume		Data Size (in MB, GB, TB)
Velocity		Data collection frequency or granularity of the observations (Longitudinal data). If the dataset increases regularly give information about this increase. State how often the data is collected.
Variety		Various formats, and/or management systems
Veracity		Type of quality problems
Value		Key Performance Indicators (KPI)
Variability		How the data evolves over time
Other comments		

Provider		Comment
Data Provider		Give the name of the data provider (e.g., organization, company, etc)
Provider URI		Data provider URI/URL
Protocol used to Access Data		Protocol used to access data or experimental strategy
Data Owner		Provide the ownership of the dataset along with any contact information
Data Administrator		Provide this information only if different from above (i.e., owner and administrator is different)

Permission Status		Is the dataset private, public, accessible under license or specific conditions?
Other Comment		

Use cases		**Comment**
Use case		The number of use cases this dataset relates to. Indicate if data can be integrated into the PLATOON knowledge base
Possible scenario coverage		Give examples about how this dataset is currently being used or will be used
Other comments		

Other Details		**Comment**
Data format(s)		Give the data format(s) in which the dataset is available. E.g. CSV, JSON, XML, RDF, ..
Data Language		Provide the language used for the data and metadata (e.g., EN, DE, IT, ..)
Assumptions		Are there any assumptions made with regard to the data? Procedure/Method followed to collect the data
Standard		Provide any standards that have been used for producing the data
Ontologies/ vocabularies used		Provide any ontologies or vocabularies used to describe the data
Accessibility, Permissions, Anonymization		Include access control and permission details. Should the data be anonymized? If so, which fields should be protected/anonymized?
Data collection frequency		State how often the data is collected
Other comments		

Raw data Sample (s) (or complete raw data, if possible)

Data Schema and Documentation

Data Source Descriptions by PLATOON Partners

Pilot 1a – Predictive Maintenance for Wind Turbine

PLATOON Partner	
Partner ID	VUB

Partner Name	Vrije Universiteit Brussel

Data Source	
Title	Wind turbine SCADA data
Alternate Title	SCADA
Acronym	-
Description	The data set contains sensors data of the Supervisory Control and Data Acquisition system of a wind turbine. This system contains sensors at the most important subcomponents of the wind turbine.
Temporal Coverage	not clear yet. Probably multiple years of data
Maintenance/Status	Updated regularly. Exchange of information is done by dedicated API. New turbine sensor data are extracted continuously from SCADA systems. The data is made available by ENGIE through an API, which is coupled to a centralized database that collects data of the turbines on a continuous basis.
Other Comments	A second dataset named wind turbine will be available. This dataset has the same structure, but no information is available on turbine condition and less parameters are available.

Big Data Vs	
Volume	Probably in the Gb range. Exact size not clear yet
Velocity	10-min averages and statistics
Variety	Typically, coming from multiple turbines. Each turbine brand has its own data structure and tag names
Veracity	Data is not clean. Unrealistic values can be present as well as missing data
Value	Data should contain turbines for periods where faults happened to the electrical subcomponents
Variability	turbines can be added/removed from the dataset over time
Provider	
Data Provider	ENGIE
Provider URI	not clear yet
Protocol used to Access Data	Data made available through dedicated API
Data Owner	ENGIE
Data Administrator	ENGIE
Permission Status	private
Use cases	
Use case	Pilot 1a all use cases
Possible scenario coverage	Will be used for fault diagnosis and tracking in Pilot 1a
Other Details	
Data format(s)	not clear yet. Probably .csv
Data Language	not clear yet.
Assumptions	-
Standard	-
Ontologies/ vocabularies used	Custom ontology for each wind turbine manufacturer/brand
Accessibility, Permissions, Anonymization	NDA

Data collection frequency	data is sampled at 10-min intervals. Dataset update is to be defined with ENGIE

Data Source	
Title	VUB
Alternate Title	Vrije Universiteit Brussel
Acronym	-
Description	High frequency electric measurements and a limited set of turbine operational parameters (e.g. wind speed). This data originates from a dedicated measurement campaign on onshore turbines
Temporal Coverage	1 year
Maintenance/Status	closed
Big Data Vs	
Volume	TB
Velocity	500Hz up to 5kHz
Variety	Data is sampled using same data acquisition system so all sensors are in the same file
Veracity	Uncleaned data.
Value	-
Variability	closed. No new data is collected
Provider	
Data Provider	ENGIE
Provider URI	Not clear yet
Protocol used to Access Data	Data are exchanged through file export and transfer.
Data Owner	ENGIE
Data Administrator	ENGIE
Permission Status	NDA
Use cases	
Use case	Pilot 1a use cases
Possible scenario coverage	Used for digital twin and model training
Other comments	-
Other Details	
Data format(s)	custom binary format of the measurement system and additionally data are available as .mat (Mathlab) after conversion done by VUB.
Data Language	ENGIE in house naming
Assumptions	-
Standard	Files are transformed into .mat files to make them independent of the proprietary binary format of the company that delivered the measurement system used for the monitoring campaign.
Ontologies/ vocabularies used	ENGIE in house naming
Accessibility, Permissions, Anonymization	NDA

Data collection frequency	sampling 500Hz to 5kHz. Measurement finished so not renewed

Data Source	
Title	Open wind speed data
Alternate Title	Flemish banks data
Acronym	
Description	The data set contains data of environmental measurements (wind speeds, wind directions, wave heights,…) across the Belgian North Sea.
Temporal Coverage	Multiple years
Maintenance/Status	Continuously updated

Big Data Vs	
Volume	In the Gb range
Velocity	10-min averages and statistics
Variety	Typically, coming from multiple measurement locations. Data is standardized to one format.
Veracity	Data is not clean. Unrealistic values can be present as well as missing data
Value	Floats
Variability	Measurement locations can be added/removed over time

Provider	
Data Provider	Vlaamse Meetbanken
Provider URI	https://meetnetvlaamsebanken.be/
Protocol used to Access Data	-
Data Owner	Flemish government Issued by the Agency Maritime Services and Coast (MDK)
Data Administrator	Vlaamse Meetbanken
Permission Status	public

Use cases	
Use case	Pilot 1a semantic reasoning: to assess typical changes in wind conditions to define environmental labels for use in knowledge graphs.
Other Details	
Data format(s)	.csv
Data Language	-
Assumptions	-
Standard	-
Ontologies/ vocabularies	Custom naming convention linked to measurement pole location.

used	
Accessibility, Permissions, Anonymization	Public
Data collection frequency	Continuously. Historically available for >5 years

Data Source	
Title	Offshore measurement campaign
Alternate Title	High frequency accelerations
Acronym	-
Description	The data set contains sensors data of accelerometers placed on the drivetrain of an offshore wind turbine.
Temporal Coverage	6 months
Maintenance/Status	consolidated

Big Data Vs	
Volume	In Tb range
Velocity	20kHz continuous measurements
Variety	Acceleration signals of 10 accelerometers and 2 encoders
Veracity	Data is not clean. Unrealistic values can be present as well as missing data
Value	Dataset gives an overview of normal behavior of acceleration signals when no failure is present
Variability	Fixed. Dataset is consolidated

Provider	
Data Provider	VUB
Provider URI	-
Protocol used to Access Data	Tdms custom binairy data format of National Instruments
Data Owner	VUB
Data Administrator	VUB
Permission Status	private
Use cases	
Use case	In the end not used since new current data is collected during dedicated measurements campaign which suits the project goals better
Possible scenario coverage	-
Data format(s)	tdms
Data Language	Binairy format of National Instruments
Assumptions	-
Standard	-
Ontologies/ vocabularies used	VUB in house naming conventions
Accessibility, Permissions, Anonymization	NDA
Data collection frequency	Data was collected for 6 months continuously at 20kHz

Data Source	
Title	Dedicated current measurement campaign
Alternate Title	ENGIE-VUB wind turbine monitoring campaign
Acronym	-
Description	Data of current sensors on one French onshore wind turbine is collected using the VUB data acquisition and edge processing hardware.
Temporal Coverage	Depending on success of measurement campaign; target is multiple months of data
Maintenance/Status	Campaign to start in November 2021
Big Data Vs	
Volume	Tb range
Velocity	20kHz/1Hz
Variety	Signals of current probes as well as turbine controller parameters (collected at lower frequency (1Hz))
Veracity	Data is not clean. Unrealistic values can be present as well as missing data
Value	Data will allow to test edge processing algorithms
Variability	1 turbine

Provider	
Data Provider	ENGIE/VUB
Provider URI	not clear yet
Protocol used to Access Data	Custom API to push data to ENGIE under design
Data Owner	VUB
Data Administrator	VUB
Permission Status	private
Use cases	
Use case	Pilot 1a edge case
Possible scenario coverage	Will be used for edge processing testing
Other Details	
Data format(s)	Custom API
Data Language	Custom API
Assumptions	-
Standard	-
Ontologies/ vocabularies used	Platoon ontology
Accessibility, Permissions, Anonymization	NDA
Data collection frequency	Several months one off campaign

Pilot 2a - Electricity Balance and Predictive Maintenance

PLATOON Partner	
Partner ID	IMP
Partner Name	Institute Mihajlo Pupin

Data Source	
Title	ENTSO-E Transparency Platform - Energy Identification Codes (EICs)
Alternate Title	ENTSO-E Transparency Platform
Acronym	ENTSO-E
Description	EIC a coding scheme has been developed, managed and maintained within ENTSO-E (under the Common Information Model Expert Group) to facilitate cross-border exchanges and to efficiently and reliably identify different objects and parties relating to the Internal Energy Market (IEM) and its operations. This is known as the Energy Identification Coding (EIC) scheme, approved by ENTSO-E for the harmonisation and implementation of standardised electronic data interchanges.
Temporal Coverage	NA
Maintenance/Status	Operational, occasionally, when database design changes needed
Other Comments	EIC will be used for modeling and integration of data
Provider	
Data Provider	ENTSO-E
Provider URI	https://www.entsoe.eu/data/energy-identification-codes-eic/#energy-identification-codes-eic-lists
Protocol used to Access Data	https
Data Owner	ENTSO-E
Data Administrator	ENTSO-E
Permission Status	free
Use cases	
Use case	#2a Electricity Balance and Predictive Maintenance
Possible scenario coverage	LLUC P-2a-03 Load Demand forecast on transmission level LLUC P-2a-04 Wind Production Forecast
Other Details	
Data format(s)	XML, CSV
Data Language	EN
Assumptions	Codes are up to date in case the Country has provided it to ENTSO-E
Standard	On 5 January 2015, in compliance with Regulation (EU) No 543/2013 on the submission and publication of data in electricity markets, ENTSO-E launched a new central transparency platform: the ENTSO-E Transparency Platform.
Ontologies/ vocabularies used	EIC scheme documentation is available here https://www.entsoe.eu/data/energy-identification-codes-eic/#energy-identification-codes-eic-documentation
Accessibility, Permissions, Anonymization	free
Data collection frequency	NA
Other comments	EIC code list is used to structure the domain (organization, market role)

Data Source	
Title	SLTF - Short Time Load Forecast
Alternate Title	SLTF - Short Time Load Forecast
Acronym	SLTF
Description	Short Term Load Forecast Datasets needed for LLUC P-2a-03 Load Demand forecast on transmission level
Temporal Coverage	2016-2021, hourly updates
Maintenance/Status	Data was transferred from SCADA to PLATOON server for training purposes.
Big Data Vs	
Volume	MB
Velocity	Hourly
Variety	XML, CSV, export from SCADA
Veracity	Missing values
Value	KPI-1 Cost efficient distribution and transmission; KPI-3 Better demand response
Variability	Constant
Provider	
Data Provider	IMP
Provider URI	-
Protocol used to Access Data	MySQL ODBC
Data Owner	Joint Stock Company EMS AD
Data Administrator	IMP
Permission Status	private
Use cases	
Use case	#2a Electricity Balance and Predictive Maintenance
Possible scenario coverage	Exploited for demand forecasting model training
Other Details	
Data format(s)	XML
Data Language	RS
Assumptions	-
Standard	-
Ontologies/ vocabularies used	cim: <http://www.iec.ch/TC57/CIM#> platoon: <https://w3id.org/platoon/> seas: <https://w3id.org/seas/> energy: <http://w3id.org/energy/> time: <http://www.w3.org/2006/time#>
Accessibility, Permissions, Anonymization	private
Data collection frequency	Hourly resolution
Raw data Sample (s) (or complete raw data, if possible)	

date	utc	load
12-31-2013	23:00:00	4431
1-1-2014	00:00:00	4375
1-1-2014	01:00:00	4131

1-1-2014	02:00:00	3898
1-1-2014	03:00:00	3675
1-1-2014	04:00:00	3548
1-1-2014	05:00:00	3483
1-1-2014	06:00:00	3430
1-1-2014	07:00:00	3539
1-1-2014	08:00:00	3771
1-1-2014	09:00:00	3973
1-1-2014	10:00:00	4091
1-1-2014	11:00:00	4130
1-1-2014	12:00:00	4067
1-1-2014	13:00:00	3995
1-1-2014	14:00:00	4010
1-1-2014	15:00:00	4370
1-1-2014	16:00:00	4535
1-1-2014	17:00:00	4557
1-1-2014	18:00:00	4555
1-1-2014	19:00:00	4550
1-1-2014	20:00:00	4488
1-1-2014	21:00:00	4394
1-1-2014	22:00:00	4359
1-1-2014	23:00:00	4291
1-2-2014	00:00:00	4072
1-2-2014	01:00:00	3787
1-2-2014	02:00:00	3582
1-2-2014	03:00:00	3486

Data Source	
Title	Meteorological data for RES production forecasting modelling
Alternate Title	-
Acronym	WeatherBit
Description	This dataset will be utilized for RES production forecasting models training proces as input data
Temporal Coverage	It will cover the same period as RES-PROD dataset
Maintenance/Status	Data is historical, so no update will be needed
Big Data Vs	
Volume	~20MB
Velocity	Weather parameters are obtained with the hourly time resolution
Variety	Data are organized within table
Veracity	Potential missing values
Value	-
Variability	Constant
Provider	
Data Provider	IMP
Provider URI	-

Protocol used to Access Data	-
Data Owner	Weather web service (e.g. WeatherBit)
Data Administrator	-
Permission Status	Private
Use cases	
Use case	LLUC P-2a-04; data will not be stored as a part of PLATOON knowledge base
Possible scenario coverage	This data will be used in RES production forecasting model training process
Other comments	-
Other Details	
Data format(s)	CSV
Data Language	EN
Assumptions	-
Standard	-
Ontologies/ vocabularies used	cim: http://www.iec.ch/TC57/CIM# platoon: https://w3id.org/platoon/ seas: https://w3id.org/seas/ wgs84_pos: <http://www.w3.org/2003/01/geo/wgs84_pos#> time: <http://www.w3.org/2006/time#>
Accessibility, Permissions, Anonymization	No
Data collection frequency	Hourly time resolution

Data Source	
Title	Historical Wind Power Production Measurements
Alternate Title	-
Acronym	RES-PROD
Description	This dataset contains measurements of the production from the wind power plant
Temporal Coverage	Depending on the necessity data could cover period of couple of previous years
Maintenance/Status	Data has been collected until 2018. Data was transferred from SCADA to PLATOON server for training purposes.
Big Data Vs	
Volume	~30MB
Velocity	Data is obtained with the hourly resolution and will probably not going to be updated any further
Variety	Data are organized within table
Veracity	Missing measurements
Value	-
Variability	Constant
Provider	
Data Provider	IMP
Provider URI	-
Protocol used to Access Data	-

Data Owner	IMP
Data Administrator	-
Permission Status	Private
Use cases	
Use case	LLUC P-2a-04; data will not be stored as a part of PLATOON knowlegde base
Possible scenario coverage	This dataset will be exploited for training of RES forecasting models
Other Details	
Data format(s)	CSV
Data Language	EN
Assumptions	-
Standard	-
Ontologies/ vocabularies used	cim: <http://www.iec.ch/TC57/CIM#> platoon: <https://w3id.org/platoon/> seas: <https://w3id.org/seas/> energy: <http://w3id.org/energy/> time: <http://www.w3.org/2006/time#>
Accessibility, Permissions, Anonymization	-
Data collection frequency	Hourly

Raw data Sample (s) (or complete raw data, if possible)	
date	**production**
5-5-2017 0:00	0
5-5-2017 1:00	0
5-5-2017 2:00	368.7
5-5-2017 3:00	989
5-5-2017 4:00	713.1
5-5-2017 5:00	650.8
5-5-2017 6:00	1182.2
5-5-2017 7:00	2433.9
5-5-2017 8:00	2537.9
5-5-2017 9:00	3468.5
5-5-2017 10:00	4778.8
5-5-2017 11:00	6783.3
5-5-2017 12:00	7329.6
5-5-2017 13:00	4796.5
5-5-2017 14:00	15.5
5-5-2017 15:00	3.2
5-5-2017 16:00	0
5-5-2017 17:00	2344.3
5-5-2017 18:00	4651.7
5-5-2017 19:00	3041.5

Data Source	
Title	Effects of Renewable Energy Sources on the Power System (distribution level)
Alternate Title	Effects of Renewable Energy Sources on the Power System (distribution level)
Acronym	RES Effects
Description	
Temporal Coverage	1 month without optimization / 1 month with optimization
Maintenance/Status	dataset will be created in PLATOON project
Big Data Vs	
Volume	10 kB
Velocity	each 15 min a report is generated by edge computing unit
Variety	XML
Veracity	missing data from meter
Value	no. of missing values from meter / from SCADA database
Variability	constant
Provider	
Data Provider	CS (measurements are coming from IMP SCADA system)
Provider URI	N.A.
Protocol used to Access Data	N.A.
Data Owner	IMP
Data Administrator	IMP
Permission Status	private
Use cases	
Use case	#2a Electricity Balance and Predictive Maintenance
Possible scenario coverage	LLUC P-2a-05 Effects of Renewable Energy Sources on the Power System (distribution level)
Other Details	
Data format(s)	XML
Data Language	EN, RS
Assumptions	CS is sending the data to PLATOON platform (1 st option - periodically)
Standard	the retrieval SCADA - > edge computing unit can be implemented via gateway
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	-
Data collection frequency	15 min resolution (minute resolution or higher if needed)
Raw data Sample (s) (or complete raw data, if possible)	

example P1 frame used for XML data generation

0-0:96.1.1(3332313037343635373930 34)
1-0:0.9.1(132514)
1-0:0.9.2(200713)
1-0:1.8.1(000099.951*kWh)
1-0:1.8.2(000000.000*kWh)
1-0:2.8.1(000000.000*kWh)
1-0:2.8.2(000000.000*kWh)
0-0:96.14.0(0001)
1-0:1.7.0(00.000*kW)
1-0:2.7.0(00.000*kW)
0-0:96.13.0()
!59C2

Data Source	
Title	RES PV Predictive maintenance
Alternate Title	RES PV Predictive maintenance
Acronym	-
Description	Dataset will be collected when the PMU is installed at IMP side
Temporal Coverage	From September 2021
Maintenance/Status	operational
Big Data Vs	
Volume	5 k
Velocity	1 s
Variety	JSON
Veracity	missing data
Value	no. Of missing values
Variability	constant
Provider	
Data Provider	CS (KPI, measurements are coming from PMU installed at IMP)
Provider URI	N.A.
Protocol used to Access Data	N.A.
Data Owner	IMP
Data Administrator	IMP
Permission Status	private
Use cases	
Use case	#2a Electricity Balance and Predictive Maintenance
Possible scenario coverage	LLUC P-2a- 07 Predictive maintenance in RES power plants
Other Details	
Data format(s)	JSON
Data Language	EN
Assumptions	one way data flow (CS is sending the data to PLATOON platform)
Standard	C37.118
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	-

Data collection frequency	1 s

Pilot 2b - Electricity Grid Stability, Connectivity, And Life Cycle

PLATOON Partner	
Partner ID	10
Partner Name	SAMPOL

Data Source	
Title	Power grid ZIV power meters
Alternate Title	-
Acronym	-
Description	Hourly measurements of active and reactive power delivered to the users, grouped by concentrator and identified by power meter.
Temporal Coverage	Data from Ocotober-2016 to now
Maintenance/Status	Manually updated each month
Big Data Vs	
Volume	300 MB
Velocity	1 value each hour for 77 power meters
Variety	-
Veracity	Not known
Value	-
Variability	Depends on building use. There are offices and educational institutions
Provider	
Data Provider	SAMPOL
Provider URI	-
Protocol used to Access Data	mySQL database
Data Owner	SAMPOL
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff and Platoon Parners can access to data
Use cases	
Use case	LLUC 2b-01 , LLUC 2b-02
Possible scenario coverage	-
Other Details	
Data format(s)	mySQL database
Data Language	EN / ES
Assumptions	PK comprises Datetime, Concentrator ID and Counter ID
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Only predefined IP addresses can access the database
Data collection frequency	stored hourly, collected monthly

Raw data Sample (s) (or complete raw data, if possible)

Cnc_ID	Cnt_ID	Tiempo1	AE	AI	R1	R2	R3	R4	Bc
ZIV0004409935	ZIV0036319014	2016-10-19 0:00	0	1432	450	0	0	0	48
ZIV0004409933	ZIV0039451041	2016-10-19 0:00	0	887	207	0	0	0	0
ZIV0004409933	ZIV0039451040	2016-10-19 0:00	0	616	0	0	0	372	0
ZIV0004409933	ZIV0039344928	2016-10-19 0:00	0	379	23	0	0	0	0
ZIV0004409935	ZIV0040337807	2016-10-19 0:00	0	369	0	0	0	226	48
ZIV0004409933	ZIV0036317828	2016-10-19 0:00	0	40	18	0	0	35	0
ZIV0004409935	ZIV0036319011	2016-10-19 0:00	0	598	44	0	0	77	48
ZIV0004409935	ZIV0039451052	2016-10-19 0:00	0	452	0	0	0	414	0
ZIV0004409933	ZIV0036319007	2016-10-19 0:00	0	8236	2015	0	0	0	0

Data Schema and Documentation	
Cnc_ID	Concentrator ID
Cnt_ID	Counter ID
tiempo	Datetime
AE	Energy Channel1
AI	Energy Channel2
R1	Energy Channel3
R2	Energy Channel4
R3	Energy Channel5
R4	Energy Channel6
Bc	Quality Control bts

Data Source	
Title	Transformer sensors
Alternate Title	Temperature transformer sensors
Acronym	TTEMP
Description	8 temperature sensors located at different positions of the transformers, 2 sensors from ambient temperature, humidity and pressure, 1 sensor for oil temperature
Temporal Coverage	Data from July 2021 to now
Maintenance/Status	Automatically uploaded daily
Big Data Vs	
Volume	60 MB
Velocity	Data is received each 5 minutes
Variety	MYSQL
Veracity	Hight
Value	Temperature data is important for the degradation of power transformers models
Variability	Depends on the weather data and the load of power transformers
Provider	
Data Provider	SAMPOL
Provider URI	N.A.
Protocol used to Access Data	mySQL database
Data Owner	SAMPOL
Data Administrator	SAMPOL

Permission Status	Private
Use cases	
Use case	LLUC 2b-01
Possible scenario coverage	LLUC 2b-01
Other Details	
Data format(s)	mySQL database
Data Language	English / Spanish
Assumptions	PK comprises Datetime, Concentrator ID and Counter ID
Standard	N.A.
Ontologies/ vocabularies used	N.A.
Accessibility, Permissions, Anonymization	Only predefined IP addresses can access the database
Data collection frequency	stored each 5 minutes, collected daily

Data Source	
Title	Medium voltage Network analyzer
Alternate Title	Medium voltage
Acronym	MVNA
Description	Electrical Network analyzer for current transformers, not yet installed
Temporal Coverage	Data from July 2021 to now
Maintenance/Status	Automatically uploaded daily
Other Comments	This device will be installed probably in Januay 2021
Big Data Vs	
Volume	60 MB
Velocity	Data is received each 5 minutes.
Variety	MYSQL
Veracity	Hight
Value	Useful to calculate loses in the power transformer, so on the degradation and RUL models and on the NTL detection
Variability	Depends on the consumption of the prosumers connected to the grid
Provider	
Data Provider	SAMPOL
Provider URI	N.A.
Protocol used to Access Data	mySQL database
Data Owner	SAMPOL
Data Administrator	SAMPOL
Permission Status	Private
Use cases	
Use case	LLUC 2b-01
Possible scenario coverage	LLUC 2b-01
Other Details	
Data format(s)	mySQL database

Data Language	-
Assumptions	PK comprises Datetime, Concentrator ID and Counter ID
Standard	N.A.
Ontologies/ vocabularies used	N.A.
Accessibility, Permissions, Anonymization	Only predefined IP addresses can access the database
Data collection frequency	stored each 5 minutes, collected daily

Pilot 3b - Advanced Energy Management System and Spatial (Multi-scale) Predictive Models in the Smart City

PLATOON Partner	
Partner ID	14
Partner Name	POSTE ITALIANE SPA

Data Source	
Title	Building Data
Alternate Title	Office Registry
Acronym	ANAG
Description	Detailed data about each building characteristics and general (ID Office, address, destination use, smq, climate zone, etc.). It will contain info regarding 'Hystorical Data Line Consumption Coefficient', i.e. the esteemed ratio of consumptions due to the specific lines (cooling, heating, lighting)
Temporal Coverage	No past temporal information
Maintenance/Status	There will be a dataset initialization, that will be final, save for unexpected relevant changes
Other Comments	Must keep trace of historical information
Big Data Vs	
Volume	(30) KB
Velocity	One shot, If changes occur. The database does not increase regularly.
Variety	Excel tables
Veracity	High
Value	ALL
Variability	Never/ Very Slow
Other comments	It could be updated if significant changes to building information occur.
Provider	
Data Provider	Poste Italiane
Provider URI	-
Protocol used to Access Data	-
Data Owner	Poste Italiane
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff at Poste Italiane and Platoon Parners can access to data
Use cases	

Use case	LLUC P-3b-01 Buildings Heating & Cooling consumption Analysis and Forecast LLUC P-3b-02 Predictive maintenance of cooling & heating plants LLUC P-3b-03 Lighting Consumption Estimation & Benchmarking
Possible scenario coverage	This information is used as a common base and reference for all use cases
Other Details	
Data format(s)	XLSx
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems.No anonymization required.
Data collection frequency	One shot
Other comments	It could be updated if significative changes on building information occurs.

Raw data Sample (s) (or complete raw data, if possible)

Funzione	Tipo di misurazione	Destinazione d'uso	Codice immobile	Frazionario	POD	Denominazione immobile	Indirizzo	Comune	Provincia	Zona climatica	Latidudine	Longitudine	kW disponibile	Superficie netta (m²)	Volume (m³)	Incidenza Consumi Heating (%)	Incidenza Consumi Cooling (%)	budget kWh
Gruppo Pilota	102	datacenter	RMLA0900	SIA08	IT002E4101381A	PALAZZO DEI CONGRESSI DATA CENTER	VIA DELLA PITTURA	ROMA	RM	D	41,83414	12,47608	2.500	2.942	5.130	0%	31%	14.010.580
	102	datacenter	RMLA0900	SIA08	IT002E4372779A													
	102	logistico	RMX13400	559622	IT001E00019945	CMP FIUMICINO - AEREOPORTO	VIA GINO CAPANNINI 2	FIUMICINO	RM	C	41,78784	12,26402	1.576	28.726	86.586	3%	15%	5.060.982
	multioraria	direzionale	RMP01600	55708	IT002E5453922A	ROMA TRULLO	VIA LENIN	ROMA	RM	D	41,70329	12,68308	490	6.915	10.410	1%	30%	375.714
	multioraria	retail	RMX10000	55977	IT002E3568472A	ROMA EUR	VIALE BEETHOVEN 36	ROMA	RM	D	41,83306	12,46638	630	4.967	22.818	21%	25%	635.527
	multioraria	direzionale	RMP02000	55712	IT002E4120893A	ROMA CINECITTA' EST	VIA TERZILIO CARDINALI S.N.C.	ROMA	RM	D	41,84991	12,59236	600	7.771	16.775	24%	26%	451.211
	multioraria	direzionale	RMX01800	55647	IT002E3931111A	ROMA NOMENTANO	PIAZZA BOLOGNA 39	ROMA	RM	D	41,91392	12,52015	490	5.817	14.703	18%	34%	494.009
	multioraria	direzionale	RMP01400	55979	IT002E4122697A	ROMA BELSITO	VIA SAPPADA	ROMA	RM	D	41,93848	12,43361	506	7.311	19.752	23%	29%	675.096
	SB	logistico	RML77731	55777	IT002E0014612A	CPD ROMA RECAPITO CASILINO	VIA CASILINA 1674	ROMA	RM	D	41,86188	12,64802	210	1.347	9.426	0%	35%	353.061
	SB	retail	RMX12400	55121	IT002E5658479A	ROMA AURELIO	VIA ACCURSIO 2	ROMA	RM	D	41,90003	12,42835	105	930	2.512	6%	19%	89.522
	SB	retail	RML83630	55836	IT002E4248928A	ROMA TRIGORIA	VIA ANTONINO GIUFFRE' 156	ROMA	RM	D	41,77113	12,47582	35	197	533	23%	25%	33.402
	SB	retail	RML28900	55289	IT002E2955080A	ROMA 40	PIAZZALE FLAMINIO 22	ROMA	RM	D	41,91227	12,47644	30	198	534	23%	25%	32.615
	SB	retail	RML89900	55899	IT002E4030829A	ROMA 132	VIA VAL PELLICE 34	ROMA	RM	D	41,94479	12,51993	43,75	197	533	23%	25%	40.737
	SB	retail	RML61900	55619	IT002E5463345A	ROMA CORVIALE	VIA DEGLI ADIMARI 22	ROMA	RM	D	41,85195	12,42314	25,6	195	527	23%	25%	43.404
	SB	retail	RML75200	55752	IT002E2988609A	ROMA 114	VIA ANTONINO LO SURDO 28	ROMA	RM	D	41,86319	12,4706	31,25	189	510	23%	25%	46.458
	SB	retail	RMP07900	55937	IT002E4039292A	ROMA 107	VIA ROSA RAIMONDI GARIBALDI 10	ROMA	RM	D	41,85686	12,49011	31,25	184	497	23%	25%	35.523
	SB	retail	RML81200	55812	IT002E4158491A	ROMA 162	VIA DI SELVA CANDIDA 467	ROMA	RM	D	41,94193	12,37987	25	177	477	23%	25%	32.821

Codice immobile	Presenza Caldaia	Cooling January	Cooling February	Cooling March	Cooling April	Cooling May
RMLA0900	no	0,0%	0,0%	0,0%	0,0%	0,0%
RMX13400	sì	0,0%	0,0%	0,0%	0,0%	0,0%
RMP01600	sì	0,0%	0,0%	0,0%	0,0%	0,0%
RMX10000	sì	0,0%	0,0%	0,0%	0,0%	0,0%
RMP02000	sì	0,0%	0,0%	0,0%	0,0%	0,0%
RMX01800	sì	0,0%	0,0%	0,0%	0,0%	0,0%
RMP01400	sì	0,0%	0,0%	0,0%	0,0%	0,0%

ZONA CLIMATICA	PERIODO DI ACCENSIONE	ORAIO DI FUNZIONAMENTO
A	1° DIC- 15 MARZO	6 ore giornaliere
B	1° DIC- 31 MARZO	8 ore giornaliere
C	15 NOV-31 MARZO	10 ore giornaliere
D	1° NOV- 15 APRILE	12 ore giornaliere
E	15 OTT- 15 APRILE	14 ore giornaliere
F	nessuna limitazione	nessuna limitazione

Data Schema and Documentation

Table 1 (Building Master Data)	
Tipo Misurazione	Building Category (Smart Building, DL-102, Multiorarie). Each category has a different data detail
Destinazione D'Uso	Building Type (Destination Use): Datacenter, Logistic, Staff, Retail
Codice Immobile	Building ID
Frazionario	Department ID
POD	Point of Distribution (of Energy).
Denominazione Immobile	Building Name
Indirizzo	Address
Comune	City
Provincia	Province
Zona climatica	Climatic zone
Latitudine	Building Latitude
Longitudine	Building Longitude
KW disponibile	Available KW

Superfice Netta (m^2)	Building Area (m^2)
Volume (m^3)	Building Volume (m^3)
Incidenza Consumi Heating (%)	Heating Consumption Rate (%)
Incidenza Consumi Cooling (%)	Cooling Consumption Rate (%)
budget kWh	KWh budget
Table 2 (Percentuali_H_C)	
Codice Immobile	Building ID
Presenza Caldaia	Presence of gas boiler
Cooling January	percentage of energy consumption from cooling - January
Cooling February	percentage of energy consumption from cooling - February
Cooling March	percentage of energy consumption from cooling - March
Cooling April	percentage of energy consumption from cooling - April
Cooling May	percentage of energy consumption from cooling - May
Cooling June	percentage of energy consumption from cooling - June
Cooling July	percentage of energy consumption from cooling - July
Cooling August	percentage of energy consumption from cooling - August
Cooling September	percentage of energy consumption from cooling - September
Cooling October	percentage of energy consumption from cooling - October
Cooling November	percentage of energy consumption from cooling - November
Cooling December	percentage of energy consumption from cooling - December
Heating January	percentage of energy consumption from cooling - January
Heating February	percentage of energy consumption from cooling - February
Heating March	percentage of energy consumption from cooling - March
Heating April	percentage of energy consumption from cooling - April
Heating May	percentage of energy consumption from cooling - May
Heating June	percentage of energy consumption from cooling - June
Heating July	percentage of energy consumption from cooling - July
Heating August	percentage of energy consumption from cooling - August
Heating September	percentage of energy consumption from cooling - January
Heating October	percentage of energy consumption from cooling - October
Heating November	percentage of energy consumption from cooling - November
Heating December	percentage of energy consumption from cooling - December
Table 3 (Zone Climatiche)	
Zona Climatica	Climate Zone
Periodo di Accensione	Heatings power-on period
Orario di Funzionamento	Heatings Operating Hours
Table 4 (Temperature_Rif)	
Inverno	20
Resto dell'Anno	26
Tolleranza Inverno	1
Tolleranza Resto dell'Anno	1

Data Source	
Title	Calendar

Alternate Title	-
Acronym	CALE
Description	Information on office openings and shifts
Temporal Coverage	From 01/01/2018
Maintenance/Status	Monthly
Other Comments	Must keep trace of historical information
Big Data Vs	
Volume	1.2 MB/Year
Velocity	Daily
Variety	CSV
Veracity	High
Value	ALL
Variability	Structure is the same, values change
Data Provider	Poste Italiane
Provider URI	-
Protocol used to Access Data	-
Data Owner	Poste Italiane
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff at Poste Italiane and Platoon Parners can access to data
Use cases	
Use case	LLUC P-3b-01 Buildings Heating & Cooling consumption Analysis and Forecast LLUC P-3b-02 Predictive maintenance of cooling & heating plants LLUC P-3b-03 Lighting Consumption Estimation & Benchmarking
Possible scenario coverage	This information is used as a common base and reference for all use cases
Other Details	
Data format(s)	CSV
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems. No anonymization required.
Data collection frequency	Monthly
Raw data Sample (s) (or complete raw data, if possible)	

FRAZIONARIO	DATA	ID_SETTIMANA	APERTURA	CHIUSURA
55121	2020-10-13 00:00:00	41	08:20	19:05
55121	2020-10-14 00:00:00	41	08:20	19:05
55121	2020-10-15 00:00:00	41	08:20	19:05
55121	2020-10-16 00:00:00	41	08:20	19:05
55121	2020-10-17 00:00:00	41	08:20	12:35
55121	2020-10-19 00:00:00	42	08:20	19:05
55121	2020-10-20 00:00:00	42	08:20	19:05
55121	2020-10-21 00:00:00	42	08:20	19:05
55121	2020-10-22 00:00:00	42	08:20	19:05
55121	2020-10-23 00:00:00	42	08:20	19:05
55121	2020-10-24 00:00:00	42	08:20	12:35
55121	2020-10-26 00:00:00	43	08:20	19:05

Data Schema and Documentation	
Frazionario	Is a department ID
Data	Date of observation
Apertura	Office Opening Time
Chiusura	Office Closing Time

Data Source	
Title	Occupancy
Alternate Title	Customers Occupancy
Acronym	OCCU_C
Description	Information on numbers of customers in the building
Temporal Coverage	From 01/01/2018
Maintenance/Status	Info will be updated and checked monthly
Big Data Vs	
Volume	2 MB / year
Velocity	Daily
Variety	CSV
Veracity	High
Value	PI_KPI01
Variability	Structure is the same, values change daily
Provider	
Data Provider	Poste Italiane
Provider URI	-
Protocol used to Access Data	-
Data Owner	Poste Italiane
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff at Poste Italiane and Platoon Parners can access to data
Use cases	
Use case	LLUC P-3b-01 Buildings Heating & Cooling consumption Analysis and Forecast
Possible scenario coverage	This information is used to evaluate the correlation between consumption and people inside the building
Other comments	

Other Details	
Data format(s)	CSV
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems. No anonymization required.
Data collection frequency	Monthly

Raw data Sample (s) (or complete raw data, if possible)

cd_fraz	Nr_Clienti	Data
55121	624	02/11/2020
55121	612	03/11/2020
55121	585	04/11/2020

Data Schema and Documentation

Cd_fraz	Department ID
Nr Clienti	Number of Customers
Data	Date of observation

Data Source	
Title	Occupancy
Alternate Title	Employees Occupancy
Acronym	OCCU_E
Description	Information on numbers of employees in the building
Temporal Coverage	From 01/01/2018
Maintenance/Status	Info will be updated and checked monthly
Big Data Vs	
Volume	2 MB / year
Velocity	Daily
Variety	CSV
Veracity	High
Value	PI_KPI01 - PI_KPI07
Variability	Structure is the same, values change daily
Provider	
Data Provider	Poste Italiane
Provider URI	-
Protocol used to Access Data	-
Data Owner	Poste Italiane
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff at Poste Italiane and Platoon Parners can access to data
Use cases	

Use case	LLUC P-3b-01 Buildings Heating & Cooling consumption Analysis and Forecast LLUC P-3b-03 Lighting Consumption Estimation & Benchmarking
Possible scenario coverage	This information is used to evaluate correlation between consumption and people inside the building
Other Details	
Data format(s)	CSV
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems.No anonymization required.
Data collection frequency	Monthly
Raw data Sample (s) (or complete raw data, if possible)	

Data	Ufficio	C.I.D.	Timbratura	
01/11/2020	559622	XXXXXXXX	S	
01/11/2020	559622	XXXXXXXX	S	
01/11/2020	559622	XXXXXXXX	S	
01/11/2020	559622	XXXXXXXX	S	
01/11/2020	559622	XXXXXXXX	S	

Data Schema and Documentation	
Data	Date
Ufficio	Department ID
Timbratura	Clocking In

Data Source	
Title	Total Energy Consumption
Alternate Title	
Acronym	EC_TOT
Description	Information on building (total) active energy consumption (kWh) of Multi Distr Buildings, DL 102 Buildings and Smart Buildings
Temporal Coverage	From 01/01/2018
Maintenance/Status	Information will be updated monthly
Big Data Vs	
Volume	Start up 40 MB -10 MB/YEAR
Velocity	Hour
Variety	CSV
Veracity	Medium
Value	PI_KPI01 - PI_KPI02 - PI_KPI03 - PI_KPI06
Variability	Structure is the same, values change
Provider	
Data Provider	Poste Italiane
Provider URI	-
Protocol used to Access Data	-
Data Owner	Poste Italiane

Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff at Poste Italiane and Platoon Developers can access data
Use cases	
Use case	LLUC P-3b-01, LLUC P-3b-03
Possible scenario coverage	Energy data consumption will be used for many purposes, such as consumption prediction, consumption benchmarking, and lighting consumption esteem
Other Details	
Data format(s)	CSV
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems.No anonymization required.
Data collection frequency	Montly
Raw data Sample (s) (or complete raw data, if possible)	

POD	Data/ora	Misura kW
IT001E00019945	01.01.18 00:00	524
IT001E00019945	01.01.18 01:00	528
IT001E00019945	01.01.18 02:00	521

Data Schema and Documentation	
POD	Point of Distribution (of Energy)
Data/ora	Timestamp
Misura KW	KW Hourly Measure

Data Source	
Title	Energy Data Consumption
Alternate Title	Detailed Energy consumption
Acronym	EC_SB
Description	Information on active energy consumption (kWh) both of line or type of system and internal temperature and humidity (for Smart Buildings) Line: cooling, heating, lighting
Temporal Coverage	By March 2021
Maintenance/Status	Information will be updated monthly
Big Data Vs	
Volume	500 MB /YEAR
Velocity	Fifteen Minutes
Variety	CSV
Veracity	Medium
Value	PI_KPI01 - PI_KPI02 - PI_KPI03 - PI_KPI06
Variability	Structure is the same, values change
Provider	
Data Provider	Poste Italiane
Provider URI	-
Protocol used to Access Data	-

Data Owner	Poste Italiane
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff at Poste Italiane and Platoon Developers can access data
Use cases	
Use case	LLUC P-3b-01, LLUC P-3b-02, LLUC P-3b-03
Possible scenario coverage	Energy data consumption will be used for many purposes, such as consumption prediction, consumption benchmarking, and lighting consumption esteem. Climate Sensors Info will be used for many purposes, such as consumption predictions which guarantee given comfort level, proper consumption benchmarking (eventually normalizing the comfort level)
Other Details	
Data format(s)	CSV
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems.No anonymization required.
Data collection frequency	Table 1 - Montly / Table 2- Daily
Raw data Sample (s) (or complete raw data, if possible)	

TABLE 1 – DL_102

```
POD,Apparato,Macro Categoria,Data,Ora,kWh
IT001E00019945,IES0006558#,CDZ,03-09-2021,00:00:00,1.2
IT001E00019945,IES0006558#,CDZ,03-09-2021,00:15:00,1.3
```

TABLE 2 –SB

```
Codice Immobile,POD,Apparato,Descrizione,Identificativo impianto,UnitÃ  di Misura,Metrica,DataOra
RMP01600,IT002E5453922A,EMD21 D1,Energia Attiva Positiva,RMP01600_Energia Attiva Positiva,kWh,0.3,2021-03-07 23:58:50
```

Data Schema and Documentation

TABLE 1-DL_102	
POD	Point of Distribution (of Energy)
Apparato	System Code
Macro Categoria	System Line category (CDZ, Lighting,)
Data	Date of observation
Ora	Hour of observation
KWh	

TABLE 2-SB	
POD	Point of Distribution (of Energy)
Codice Immobile	Building ID
Apparato	System Code
Descrizione	Type of measurement description
Identificativo Impianto	System ID
Unità di Misura	Unit (Kwh, C°, etc.)
Metrica	Measurement (Value)
DataOra	DateTime

Data Source

Title	Building Systems
Alternate Title	System Registry
Acronym	BS
Description	Information on kind and characteristics of heating, cooling and lighting plants of all Buildings
Temporal Coverage	No temporal information
Maintenance/Status	Is updated if changes occur
Other Comments	Must keep trace of historical information
Big Data Vs	
Volume	1.5 MB
Velocity	One shot
Variety	Excel tables
Veracity	High
Value	ALL
Variability	Never/ Very Slow
Other comments	It's a static data source. It could be updated only if significant changes to information occur.
Provider	
Data Provider	Poste Italiane
Provider URI	-
Protocol used to Access Data	-
Data Owner	Poste Italiane
Data Administrator	-
Permission Status	Private - only for Platoon purposes
Other Comment	Only authorized staff at Poste Italiane and Platoon Developers can access to data
Use cases	
Use case	LLUC P-3b-01, LLUC P-3b-02, LLUC P-3b-03
Possible scenario coverage	Building Systems Info will be used for many purposes, such as consumption prediction, consumption benchmarking, plant anomalies prediction. Building Lighting Plants Info will be used for many purposes, such as lighting consumption benchmarking, and lighting consumption esteem
Other Details	
Data format(s)	XLS
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems.No anonymization required.
Data collection frequency	One shot
Other comments	It could be updated only if significant changes to information occur
Raw data Sample (s) (or complete raw data, if possible)	

Codice Immobile	Categoria Impianto	ID Impianto	Tipo impianto	Tecnologia Impianto	Tipo apparato	Nr Impianti	Potenza unitaria W	Potenza tot	h/g	h/sett	h/anno	fattore di aggiustamento 1 (es rendimento)	fattore di aggiustamento 2 (es caric	Consumo MWh
RMP07900	Lighting Esterno			Faretto Alogeno		1	70							
RMP07900	Lighting Interno			Faretto Led		29	16							
RMP07900	Lighting Interno			Faretto Led		4	16							
RMP07900	Lighting Interno			Faretto Led		29	16							
RMP07900	Lighting Interno			Plafoniera Neon		6	21							
RMP07900	Lighting Interno			Plafoniera Neon		3	43							

Data Schema and Documentation

Codice Immobile	Building ID
Categoria impianto	System Line category (CDZ, Lighting,)
ID impianto	System ID
Tipo impianto	Name/Type of System
Tecnologia impianto	System Technology (for lighting only)
Nr Impianti	Number of Systems
Potenza Unitaria W (nominale)	System Unitary Rated Power

Data Source	
Title	Systems Anomalies
Alternate Title	-
Acronym	FAULT
Description	Information on anomaly behavior of heating and cooling plants
Temporal Coverage	by April 2021
Maintenance/Status	Information will be updated monthly
Big Data Vs	
Volume	400 MB/YEAR
Velocity	Daily
Variety	Excel tables
Veracity	High
Value	PI_KPI05
Variability	Structure is the same, values change
Provider	
Data Provider	Poste Italiane
Provider URI	-
Protocol used to Access Data	-
Data Owner	Poste Italiane
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff at Poste Italiane and Platoon Developers can access data
Use cases	
Use case	LLUC P-3b-02
Possible scenario coverage	System anomalies will be used for anomalies detection
Other Details	
Data format(s)	CSV
Data Language	IT
Assumptions	-
Standard	-

Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems. No anonymization required.
Data collection frequency	Daily

Raw data Sample (s) (or complete raw data, if possible)

POD,DescrizioneAllarme,Severity,DataOra

IT002E5453922A, Dispositivo diverso da KET-THL-200, high,2021-03-07 23:58:50

Data Schema and Documentation

POD	Point of Distribution (of Energy).
DescrizioneAllarme	Alarm Description
Severity	Severity
DataOra	DateTime

PLATOON Partner	
Partner ID	ROM
Partner Name	ROMA CAPITALE + RISORSE PER ROMA (TLP)

Data Source	
Title	energy meters electrical Monthly consumptions
Alternate Title	energy meters electrical Monthly Current consumptions for ROM buildings
Acronym	EMEMC
Description	Last month consumptions from all power meters (energy vendor)
Temporal Coverage	30 days
Maintenance/Status	delivered to ROM each month - data concerning the previous complete month
Other Comments	CSV file, downloadable from MYENEL (vendor) portal

Data Source	
Title	energy meters Electric Historical consumptions
Alternate Title	energy meters Electric Historical consumptions for ROM buildings
Acronym	EMEHC1
Description	517598 records for daily kwh; 96 columns for 15minutes consumptions
Temporal Coverage	30 months from 1-1-2018 to 30-6-2020
Maintenance/Status	UPDATED MONTHLY BY vendor ENEL (based on ARETI data)
Other Comments	CSV file, downloadable from MYENEL (vendor) portal; ARETI can supply the previous datasets back to 2015 for a period of 5,5 years (66 months)
Big Data Vs	
Volume	428 Mb
Velocity	monthly
Variety	CSV file, downloadable from MYENEL (vendor) portal

Veracity	available for ONLY 575 meters on 6500 of the total meters
Value	Pilot 3b - ROM
Variability	monthly increased
Provider	
Data Provider	Enel
Provider URI	-
Protocol used to Access Data	Download CSV from MyEnel portal
Data Owner	SIMU - Energy Manager Office
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff can access data
Use cases	
Use case	Pilot 3b – ROM use case
Possible scenario coverage	Historical consumption can be used to train forecast models
Other Details	
Data format(s)	CSV
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems. No anonymization required.
Data collection frequency	Static

Nr Utente	Grandezza	Consumo Giorno (Kwh)	Q1 (kWh)	Q2 (kWh)	Q3 (kWh)	Q4-Q95	Q96 (kWh)	PIVA_CF | POD	Data Consumo
114164501	ATTIVA	101	0.781	0.656	0.738	...	1	02438750586 | IT002E0177251A	9/10/2018
114164501	ATTIVA	85	1.069	0.831	0.763	...	0.725	02438750586 | IT002E0177251A	9/11/2018
114164501	ATTIVA	117	0.806	0.806	0.713	...	1.175	02438750586 | IT002E0177251A	9/12/2018
114164501	ATTIVA	123	1.206	1.094	1.169	...	0.856	02438750586 | IT002E0177251A	9/13/2018
114164501	ATTIVA	137	0.8	0.863	0.713	...	0.231	02438750586 | IT002E0177251A	9/14/2018
114164501	ATTIVA	24	0.275	0.231	0.231	...	0.238	02438750586 | IT002E0177251A	9/15/2018

114164501	ATTIVA	25	0.231	0.281	0.231	...	0.231	02438750586 | IT002E0177251A	9/16/2018
114164501	ATTIVA	132	0.275	0.238	0.231	...	0.8	02438750586 | IT002E0177251A	9/17/2018
114164501	ATTIVA	141	0.65	0.519	0.55	...	0.588	02438750586 | IT002E0177251A	9/18/2018

Data Source	
Title	Buildings Master Data
Alternate Title	-
Acronym	-
Description	from ROM Asset Management Office and buildings Energy Audits DB
Temporal Coverage	Static file
Maintenance/Status	This dataset would not change
Other Comments	-
Big Data Vs	
Volume	223 KB
Velocity	Static
Variety	XLSX file
Veracity	Uncleaned Data (address and location of buildings)
Value	Pilot 3b – ROM
Variability	-
Provider	
Data Provider	CPL-EMF
Provider URI	-
Protocol used to Access Data	-
Data Owner	SIMU - Energy Manager Office
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff can access data
Use cases	
Use case	Pilot 3b – ROM use case
Possible scenario coverage	This dataset is used as the building registry
Other Details	
Data format(s)	XLSX
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies	-

used	
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems. No anonymization required.
Data collection frequency	Static

Codice cliente	Desc.cliente	Codice sito	Nome sito	Codice edificio	Nome edificio	Indirizzo 2	N.civico	C.A.P.	Longitudine	Latitudine	Dest.d'uso
00004	ROMA CAPITALE	01314	VIA DELLE FRAGOLE 30	0001577	05153 - TERESA GULLACE	VIA DELLE FRAGOLE, 30 - Roma (RM)	30	00100	12,58195130	41,88002250	E.7 - Edificio adibito ad attività
00004	ROMA CAPITALE	01500	VIA GADOLA 25	0001578	05154 - LUIGI GADOLA - CT	VIA GADOLA, 25 - Roma (RM)	25	00100	12,58590770	41,88334920	E.7 - Edificio adibito ad attività scolastiche a tutti i livelli ed
00004	ROMA CAPITALE	01524	VIA GIOVANNI BATTISTA VALENTE 142	0001579	05167 - IL PETTIROSSO - CT	VIA GIOVANNI BATTISTA VALENTE, 142 - Roma (RM)	142	00100	12,56652530	41,89886930	E.7 - Edificio adibito ad attività scolastiche a tutti i livelli ed
00004	ROMA CAPITALE	01523	VIA GIOVANNI BATTISTA	0001580	05168 - I. C. . GIOVANNI BATTISTA	VIA GIOVANNI BATTISTA VALENTE, 100 -	100	00100	12,56654790	41,89777480	E.7 - Edificio adibito ad attività
00004	ROMA CAPITALE	01521	VIA GIORGIO PERLASCA 33	0001583	05176 - MUNICIPIO 7 UOT	VIA GIORGIO PERLASCA, 33 - Roma (RM)	33	00100	12,57310690	41,90059730	E.2 - Edificio adibito ad ufficio ed assimilabili
00004	ROMA CAPITALE	01522	VIA GIORGIO PERLASCA	0001584	05177 - PERLASCA - MATERNA - CT	VIA GIORGIO PERLASCA, 59 - Roma (RM)	59	00100	12,57045450	41,90061230	E.7 - Edificio adibito ad attività
00004	ROMA CAPITALE	01522	VIA GIORGIO PERLASCA 59	0001585	05178 - ASILO NIDO PERLASCA - CT	VIA GIORGIO PERLASCA, 59 - Roma (RM)	59	00100	12,57045450	41,90061230	E.7 - Edificio adibito ad attività scolastiche a tutti i livelli ed
00004	ROMA CAPITALE	02024	VIA TORRE ANNUNZIATA	0001586	05185 - PALESTRA	VIA TORRE ANNUNZIATA, 15 - Roma (RM)	15	00100	12,54891050	41,89097250	E.6 (2) - Edificio adibito a
00004	ROMA CAPITALE	01378	VIA EMILIO MACRO 25	0001587	06001 - I. C. VIA E. MACRO - PLESSO	VIA EMILIO MACRO, 25 - Roma (RM)	25	00100	12,57468980	41,87145170	CENTRALE TERMICA
00004	ROMA CAPITALE	01908	VIA RUGANTINO 99	0001591	06005 - RUGANTINO - CT	VIA RUGANTINO, 99 - Roma (RM)	99	00100	12,57893280	41,86452810	CENTRALE TERMICA
00004	ROMA CAPITALE	01496	VIA G.BERNERI 7/9	0001593	06007 - I. C. VIA E. MACRO -	VIA G.BERNERI, 7/9 - Roma (RM)	7/9	00100	12,58609720	41,86823000	E.7 - Edificio adibito ad attività
00004	ROMA CAPITALE	00920	VIA A.GIAQUINTO 24	0001594	06008 - I. C. VIA E. MACRO - SUCCURSALE - CT	VIA A.GIAQUINTO, 24 - Roma (RM)	24	00100	12,57429720	41,86765150	CENTRALE TERMICA
00004	ROMA CAPITALE	00919	VIA A.GIAQUINTO 12	0001595	06009 - I. C. VIA E. MACRO - SUCCURSALE - PLESSO F. DE	VIA A.GIAQUINTO, 12 - Roma (RM)	12	00100	12,57554440	41,86795560	E.7 - Edificio adibito ad attività scolastiche a tutti i livelli ed
00004	ROMA CAPITALE	01666	VIA MARCIO RUTILIO 10	0001596	06010 - SERVIZIO GIARDINI	VIA MARCIO RUTILIO, 10 - Roma (RM)	10	00100	12,57227000	41,87066000	E.2 - Edificio adibito ad ufficio ed assimilabili
00004	ROMA CAPITALE	01335	VIA DI LUNGHEZZA 1	0001597	06011 - FRANCO MARTELLI - CT	VIA DI LUNGHEZZA, 1 - Roma (RM)	1	00100	12,67237000	41,92307000	CENTRALE TERMICA
00004	ROMA CAPITALE	01136	VIA CATIGNANO 4	0001598	06012 - I. C. CASTELVERDE -	VIA CATIGNANO, 4 - Roma (RM)	4	00100	12,69214390	41,90587130	E.7 - Edificio adibito ad attività
00004	ROMA CAPITALE	00887	V.FOSSO DELL'OSA 501/503	0001599	06013 - I. C. VILLAGGIO PRENESTINO - SUCCURSALE - CT	V.FOSSO DELL'OSA, 501/503 - Roma (RM)	501/503	00100	12,68178000	41,90829990	E.7 - Edificio adibito ad attività scolastiche a tutti i livelli ed
00004	ROMA CAPITALE	01451	VIA FOSSO DELL'OSA 507	0001600	06014 - I. C. VILLAGGIO PRENESTINO - CT	VIA FOSSO DELL'OSA, 507 - Roma (RM)	507	00100	12,68101300	41,90933960	E.7 - Edificio adibito ad attività scolastiche a tutti i livelli ed
00004	ROMA CAPITALE	01150	VIA CITTA'S.ANGELO 31	0001602	06018 - I. C. CASTELVERDE - CT	VIA CITTA'S.ANGELO, 31 - Roma (RM)	31	00100	12,69294000	41,90745000	E.7 - Edificio adibito ad attività scolastiche a tutti i livelli ed
00004	ROMA CAPITALE	01150	VIA CITTA'S.ANGELO 31	0001603	06019 - I. C. CASTELVERDE - CT	VIA CITTA'S.ANGELO, 31 - Roma (RM)	31	00100	12,69294000	41,90745000	E.7 - Edificio adibito ad attività scolastiche a tutti i livelli ed
00004	ROMA CAPITALE	01080	VIA CANTIANO 131	0001605	06027 - I. C. S. VITTORINO - CORCOLLE	VIA CANTIANO, 131 - Roma (RM)	131	00100	12,72799440	41,91228630	E.7 - Edificio adibito ad attività
00004	ROMA CAPITALE	01749	VIA NUSCO	0001608	06032 - I. C. M. G. CUTUL -	VIA NUSCO, snc - Roma (RM)	snc	00100	12,63048900	41,89173880	E.7 - Edificio adibito ad attività
00004	ROMA CAPITALE	01227	VIA DEI TORDI 38	0001609	06033 - ELE - NON IN CONTRATTO	VIA DEI TORDI, 38 - Roma (RM)	38	00100	12,59138100	41,87120950	E.7 - Edificio adibito ad attività
00004	ROMA CAPITALE	01243	VIA DEL FRINGUELLO 19	0001610	06034 - I. C. VIA DELLE ALZAVOLE 21 - SUCC. V. BACHELET - CT	VIA DEL FRINGUELLO, 19 - Roma (RM)	19	00100	12,59606800	41,86797920	CENTRALE TERMICA

Data Source	
Title	energy meters Electric Historical consumptions
Alternate Title	energy meters Electric Historical consumptions for ROM buildings
Acronym	EMEHC2
Description	----- records for daily kwh;-- columns for --- consumptions
Temporal Coverage	36 months from 1-1-2015 to 31-12-2017
Maintenance/Status	old dataset
Other Comments	from GALA (previous vendor) AND (more detailed) from ARETI
Big Data Vs	
Volume	~250 MB
Velocity	Static
Variety	TXT file
Veracity	Uncleaned Data
Value	Pilot 3b – ROM
Variability	-
Provider	
Data Provider	GALA
Provider URI	-
Protocol used to Access Data	-
Data Owner	SIMU - Energy Manager Office
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff can access data
Use cases	
Use case	Pilot 3b – ROM use case

Possible scenario coverage	Use for benchmarking and forecast analysis
Other Details	
Data format(s)	XLSX
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems. No anonymization required.
Data collection frequency	Static

IDANAG	RAGIONESOCIALE	POD	IDSEDE	DATA	TOTATTIVA	H00	H...	H23	REA00	REA...	REA23	POT00	POT...	POT23	COSPHI
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	01/01/2016 00:00	128,171	4,088	4,101	4,239	0,239	0,239	0,338	4,2	4,3	4,4	0,959245
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	02/01/2016 00:00	118,611	4,063	4,114	4,389	0,263	0,264	0,338	4,252	4,5	4,6	0,976316
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	03/01/2016 00:00	105,873	4,039	4,251	4,376	0,276	0,289	0,575	4,152	4,252	4,5	0,99597
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	04/01/2016 00:00	159,246	4,464	4,338	4,101	0,563	0,588	0,801	4,7	4,552	4,252	0,940402
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	05/01/2016 00:00	144,881	4,025	4,014	4,089	0,764	0,776	0,738	4,2	4,4	4,452	0,938203
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	06/01/2016 00:00	125,568	3,938	3,864	3,876	0,775	0,825	0,662	4	4,252	4,052	0,929111
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	07/01/2016 00:00	553,969	4,014	3,951	4,239	0,764	0,763	1,064	4,3	4,052	4,452	0,987566
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	08/01/2016 00:00	534,366	4,263	4,264	4,575	0,913	1,201	1,113	4,6	4,4	4,7	0,98341
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	09/01/2016 00:00	204,129	4,514	4,551	5,738	1,051	1,152	1,901	4,652	4,9	5,948	0,960381
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	10/01/2016 00:00	132,276	5,901	5,675	4,876	1,889	1,738	1,213	6,2	5,752	5	0,963642
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	11/01/2016 00:00	532,218	4,988	4,763	4,826	1,251	1,177	0,188	5,352	5,1	5,148	0,981784
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	12/01/2016 00:00	525,523	4,752	4,727	5,688	0,163	0,089	1,176	5,252	4,9	6,1	0,9805
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	13/01/2016 00:00	541,88	5,563	5,477	5,651	1,163	1,126	1,088	5,7	5,9	6	0,979282
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	14/01/2016 00:00	579,884	5,576	5,701	5,288	1,014	0,976	1,088	5,752	5,552	5,4	0,985862
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	15/01/2016 00:00	551,583	5,201	5,313	5,464	1,089	1,026	0,95	5,452	5,3	5,752	0,986191
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	16/01/2016 00:00	356,442	5,2	5,188	4,85	0,988	0,864	0,6	5,352	5,652	5,2	0,98507
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	17/01/2016 00:00	122,233	4,676	4,788	5,226	0,562	0,526	0,437	4,9	4,9	5,5	0,996784
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	18/01/2016 00:00	522,615	5,063	5,114	4,588	0,363	0,476	0,587	5,4	5,352	5,148	0,987526
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	19/01/2016 00:00	503,027	4,289	4,225	4,613	0,525	0,638	0,176	4,4	4,752	4,852	0,985873
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	20/01/2016 00:00	525,716	4,352	4,425	4,513	0,163	0,2	0,113	4,452	4,752	4,7	0,990328
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	21/01/2016 00:00	467,983	4,313	4,476	4,439	0,138	0,213	0,586	4,6	5,1	4,652	0,98511
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	22/01/2016 00:00	493,02	4,401	4,414	4,638	0,475	0,613	0,601	4,6	4,6	4,852	0,986539
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	23/01/2016 00:00	197,321	4,35	4,588	4,826	0,551	0,549	0,375	4,4	4,8	5,2	0,983005
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	24/01/2016 00:00	114,86	4,725	4,789	4,839	0,4	0,351	0,45	5	4,952	5,2	0,997664
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	25/01/2016 00:00	523,019	4,589	4,713	4,589	0,376	0,326	0,575	5,052	4,8	4,752	0,985616
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	26/01/2016 00:00	482,998	4,539	4,689	4,651	0,525	0,475	0,564	4,752	4,652	4,952	0,982686
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	27/01/2016 00:00	482,677	4,514	4,551	4,763	0,688	0,6	0,425	4,752	4,852	4,9	0,987022
181900	ROMA CAPITALE - COMUNE DI ROMA DIPARTIME	IT002E0057553A	474257	28/01/2016 00:00	513,919	4,525	4,588	4,738	0,439	0,375	0,688	4,8	4,952	4,9	0,988238

Data Source	
Title	Energy Meter Gas Monthly Consumption RC Direct
Alternate Title	-
Acronym	-
Description	Monthly consumption for RC direct Gas meters from ESTRA
Temporal Coverage	-
Maintenance/Status	Information will be updated monthly
Big Data Vs	
Volume	Data volume in MBs
Velocity	Montly
Variety	XLSX
Veracity	Uncleaned Data
Value	Pilot 3b - ROM
Variability	-
Provider	
Data Provider	CPL-EFM
Provider URI	-
Protocol used to Access Data	-
Data Owner	SIMU - Energy Manager Office
Data Administrator	-
Permission Status	Private

Other Comment	Only authorized staff can access to data
Use cases	
Use case	Pilot 3b - ROM
Possible scenario coverage	Gas consumption data can be used to perform benchmarking analysis
Other Details	
Data format(s)	XLSX
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems.No anonymization required.
Data collection frequency	Montly
Data Source	
Title	Energy Meter Gas Historical Consumption RC Direct
Alternate Title	-
Acronym	-
Description	Monthly consumption for RC direct Gas meters from ESTRA
Temporal Coverage	From 07/2016 to 01/2021
Maintenance/Status	static
Big Data Vs	
Volume	19MB
Velocity	Montly
Variety	XLSX
Veracity	Uncleaned Data
Value	Pilot 3b - ROM
Variability	-
Provider	
Data Provider	ESTRA
Provider URI	-
Protocol used to Access Data	-
Data Owner	SIMU - Energy Manager Office
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff can access data
Use cases	
Use case	Pilot 3b - ROM
Possible scenario coverage	Historical gas consumption data can be used to train forecast models
Other Details	
Data format(s)	XLSX
Data Language	IT
Assumptions	-

Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems.No anonymization required.
Data collection frequency	Monthly

Raw data Sample (s) (or complete raw data, if possible)

anno	Anno calendario/mese	Numero impianto	pdr	indirizzo Fornitura	Località	Ricavi Vendita (Doc. Calcolo)	Quantità GAS
2016	LUG 2016	1371376	00881106273723	PIAZZA FRANCESCO BORGONGINI DUCA	ROMA	-110,17 EUR	52 m3
2016	LUG 2016	1371377	00881106298407	VIA DOMENICO SILVERI	ROMA	32,19 EUR	47 m3
2016	LUG 2016	1371378	00880000079224	VIA DELL' ARA PACIS	ROMA	545,51 EUR	786 m3
2016	LUG 2016	1371379	00880000869036	VIALE GIOVANNI BATTISTA VALENTE	ROMA	-29,47 EUR	0 m3
2016	LUG 2016	1371380	00881106431032	VIA DEI LAMPUGNANI	ROMA	16,83 EUR	24 m3
2016	LUG 2016	1371381	00881109514107	VIA CUTIGLIANO	ROMA	7,18 EUR	14 m3
2016	LUG 2016	1371382	00881109626323	VIA VINCENZO BRUNACCI	ROMA	46,16 EUR	30 m3
2016	LUG 2016	1371383	00880001277418	VIA MONTAGANO	ROMA	2,77 EUR	0 m3
2016	LUG 2016	1371384	00881109828325	VIA OSTIENSE	ROMA	100,25 EUR	0 m3
2016	LUG 2016	1371385	00881110537733	VIA DELL' ARCHITETTURA	ROMA	30,03 EUR	46 m3

Data Source	
Title	Energy Meter Gas Thermal Consumption SIE3
Alternate Title	
Acronym	
Description	Thermal consumption for SIE3 Gas meters from CPL-EMF
Temporal Coverage	From 09/2018 to 11/2021
Maintenance/Status	static
Big Data Vs	
Volume	2.8GB
Velocity	15min
Variety	CSV
Veracity	Uncleaned Data
Value	Pilot 3b - ROM
Variability	-
Provider	
Data Provider	CPL-EMF
Provider URI	-
Protocol used to Access Data	-
Data Owner	SIMU – Thermal Plants Office
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff can access to data
Use cases	
Use case	Pilot 3b - ROM
Possible scenario coverage	Thermal consumption data can be used for benchmarking
Other Details	

Data format(s)	CSV
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems.No anonymization required.
Data collection frequency	Montly

Raw data Sample (s) (or complete raw data, if possible)

AUTO_NUMBER	DATA_POINT_ID	NAME	TO_CHAR(DATE_TIME_MEASURED,'DD/MM/YYYYHH24:MI')	VALUE_REPORTED
203356349	145330	VL_ENERGIA_02.004_VVU	02/08/2020 04:30	1572995456
203356350	145330	VL_ENERGIA_02.004_VVU	02/08/2020 05:00	1572995456
203356351	145330	VL_ENERGIA_02.004_VVU	02/08/2020 05:30	1572995456
203356352	145330	VL_ENERGIA_02.004_VVU	02/08/2020 06:00	1572995456
203356353	145330	VL_ENERGIA_02.004_VVU	02/08/2020 06:30	1572995456
203356354	145330	VL_ENERGIA_02.004_VVU	02/08/2020 07:00	1572995456
203356355	145330	VL_ENERGIA_02.004_VVU	02/08/2020 07:30	1572995456
203356356	145330	VL_ENERGIA_02.004_VVU	02/08/2020 08:00	1572995456
203356357	145330	VL_ENERGIA_02.004_VVU	02/08/2020 08:30	1572995456
203356358	145330	VL_ENERGIA_02.004_VVU	02/08/2020 09:00	1572995456
203356359	145330	VL_ENERGIA_02.004_VVU	02/08/2020 09:30	1572995456
203356360	145330	VL_ENERGIA_02.004_VVU	02/08/2020 10:00	1572995456
203356361	145330	VL_ENERGIA_02.004_VVU	02/08/2020 10:30	1572995456
203356362	145330	VL_ENERGIA_02.004_VVU	02/08/2020 11:00	1572995456
203356363	145330	VL_ENERGIA_02.004_VVU	02/08/2020 11:30	1572995456

Data Source	
Title	Energy Meter Gas Historical Consumption SIE3
Alternate Title	-
Acronym	-
Description	Historical gas consumption data for SIE3 Gas meters from CPL from November 2018 to April 2021
Temporal Coverage	From 09/2018 to 04/2021
Maintenance/Status	static
Big Data Vs	
Volume	1MB
Velocity	Month
Variety	XLSX
Veracity	Uncleaned Data
Value	Pilot 3b - ROM
Variability	-
Provider	
Data Provider	CPL-EMF
Provider URI	-
Protocol used to Access Data	-
Data Owner	SIMU – Thermal Plants Office
Data Administrator	-
Permission Status	Private
Other Comment	Only authorized staff can access to data
Use cases	

Use case	Pilot 3b - ROM
Possible scenario coverage	Historical thermal consumption data can be used to train forecast models
Other Details	
Data format(s)	XLSX
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems.No anonymization required.
Data collection frequency	-

Raw data Sample (s) (or complete raw data, if possible)

POD/PDR	Tipo misurazione	ID rilevatore dati	Codice cliente	Descrizione	Codice sito	Nome sito	Codice edificio	Nome edificio	Indirizzo	Data misurazione	Mese	Anno	Consumo
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/12/2018	Dicembre	2018	2.076,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/11/2018	Novembre	2018	1.482,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/04/2019	Aprile	2019	67,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/12/2019	Dicembre	2019	1.491,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/02/2019	Febbraio	2019	2.439,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/01/2019	Gennaio	2019	2.561,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/03/2019	Marzo	2019	1.980,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/11/2019	Novembre	2019	1.045,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/04/2020	Aprile	2020	1,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/12/2020	Dicembre	2020	1.873,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/02/2020	Febbraio	2020	2.661,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/01/2020	Gennaio	2020	2.912,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/03/2020	Marzo	2020	1.399,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/11/2020	Novembre	2020	1.477,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/04/2021	Aprile	2021	724,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/02/2021	Febbraio	2021	2.254,00
00880000026792	GAS - NATURAL	1558	00004	ROMA CAPITALE	01519	VIA GIARRE, 41 - 45	0001676	06141 - GIARRE - CT	VIA GIARRE, 41 - 45	01/01/2021	Gennaio	2021	2.477,00

Data Source	
Title	ROM PV production data
Alternate Title	-
Acronym	-
Description	Res data production from Lovato Electric system. This dataset contains the produced kWh of the installed PV plants in a set of ROM buildings divided by each district
Temporal Coverage	From 01/2020
Maintenance/Status	Monthly
Big Data Vs	
Volume	Data volume in MBs
Velocity	15 min
Variety	XLSX
Veracity	Uncleaned Data
Value	Pilot 3b - ROM
Variability	-
Provider	
Data Provider	Lovato Electric
Provider URI	-
Protocol used to Access Data	Manual download from Lovato Electric "Synergy" platform
Data Owner	SIMU
Data Administrator	-
Permission Status	Private

Other Comment	Only authorized staff can access to data
Use cases	
Use case	RES Potentialities
Possible scenario coverage	This dataset is used as the reference for PV plants production
Other Details	
Data format(s)	XLSX
Data Language	IT
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Free for upload on Platoon Systems.No anonymization required.
Data collection frequency	-

Raw data Sample (s) (or complete raw data, if possible)

Produzione_Totale_Municipio_XIV - 2022-01-16 18:51:37

Data	Via Andrea Verga - kWh-	Via Andrea Verga - kWh	Via Andrea Verga - Delta kWh	Energia_Totale_Prodotta
10/02/2020 15:00:00	0	0,28		0,28
10/02/2020 15:15:00	0	0,29		0,29
10/02/2020 15:30:00	0	0,34		0,34
10/02/2020 15:45:00	0	0,37		0,37
10/02/2020 16:00:00	0	0,39		0,39
10/02/2020 16:15:00	0,01	0,42		0,42
10/02/2020 16:30:00	0,02	0,43		0,43
10/02/2020 16:45:00	0,03	0,44		0,44
10/02/2020 17:00:00	0,05	0,45		0,45
10/02/2020 17:15:00	0,06	0,46		0,46
10/02/2020 17:30:00	0,06	0,46		0,46
10/02/2020 17:45:00	0,06	0,46		0,46
10/02/2020 18:00:00	0,06	0,47		0,47
10/02/2020 18:15:00	0,06	0,47		0,47
10/02/2020 18:30:00	0,06	0,48		0,48
10/02/2020 18:45:00	0,06	0,48		0,48
10/02/2020 19:00:00	0,06	0,49		0,49

Pilot 3c Advance Energy Management and Energy Efficiency and Predictive Maintenance in the Smart Tertiary Building Hubgrade

PLATOON Partner	
Partner ID	GIR
Partner Name	Giroa Sociedad Anonima

Data Source	
Title	SIEMENS DESIGO 4.0
Alternate Title	-
Acronym	SCADA
Description	SCADA data: temperatures, electricity consumption, position of valves. Also, weather data and forecasts.
Temporal Coverage	2019-Now
Maintenance/Status	ACTIVE
Big Data Vs	

Volume	1GB
Velocity	1.5 MB/Day
Variety	JSON
Veracity	Available in all the data
Value	Experimental Strategy: Predictive maintenance and energy management.
Variability	-
Provider	
Data Provider	GIR
Provider URI	www.veolia.com
Protocol used to Access Data	SQL
Data Owner	GIR
Data Administrator	GIR
Permission Status	Private - only for Platoon purposes
Use cases	
Use case	**Smart tertiary Building**
Possible scenario coverage	Monitoring signals from chillers to predict possible failures or analyze electrical consumption from pumps to reduce it.
Other Details	
Data format(s)	JSON
Data Language	SQL
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Allowed Partner access via VPN client with r/w permission
Data collection frequency	1 min
Raw data Sample (s) (or complete raw data, if possible)	

2021-08-11 01:00:00.0000000 NAN_Edi_P1_Grupo_kW_Des_Fue Grupo: Potencia P1 Despachos Enchufes 0,0264109298586845

Pilot 4a Energy Management in Microgrids

PLATOON Partner	
Partner ID	MPD
Partner Name	POLITECNICO DI MILANO
Data Source	
Title	Microgrid PV Power
Alternate Title	-
Acronym	MicroGridPVPower
Description	-
Temporal Coverage	-
Maintenance/Status	ACTIVE
Big Data Vs	
Volume	Approx.. in the range of Gb
Velocity	Updates per minute

Variety	-
Veracity	Available in all the data
Value	-
Variability	-
Provider	
Data Provider	MDP
Provider URI	-
Protocol used to Access Data	SQL
Data Owner	MDP
Data Administrator	MDP
Permission Status	Private - only for Platoon purposes
Use cases	
Use case	-
Possible scenario coverage	-
Other Details	
Data format(s)	-
Data Language	-
Assumptions	-
Standard	-
Ontologies/ vocabularies used	-
Accessibility, Permissions, Anonymization	Allowed Partner with r/w permission
Data collection frequency	-
Raw data Sample (s) (or complete raw data, if possible)	