Grant Agreement N° 872592

PLATOON

Digital platform and analytics tools for energy

Deliverable D5.3
Harmonization and Knowledge Extraction Service

Contractual delivery date:
M27

Actual delivery date:
31/03/2022

Responsible partner:
P11: TIB, Germany

Project Title	PLATOON – Digital platform and analytic tools for energy
Deliverable number	D5.3
Deliverable title	Harmonization and Knowledge Extraction Service
Author(s):	Maria-Esther Vidal, Ahmad Sakor, Enrique Iglesias, Mazen Bechara
Responsible Partner:	P11 – TIB
Date:	31.03.2022
Nature	R
Distribution level (CO, PU):	PU
Work package number	WP5 – Big data sharing and analysis reference implementations

Work package leader	UBO
Abstract:	The PLATOON pilots make available data sources that, even heterogeneous, share similar concepts from the energy domain. Task T5.3 aims at uncovering the commonalities across the PLATOON data sources in terms of semantic types from the PLATOON Semantic Data Models. The Data Catalog (DCAT) vocabulary provides the basis for the harmonized data source description. At the same time, semantic types enable their annotation with concepts from the energy domain whose meaning is commonly accepted by the energy sector community. An RDF graph is created with the harmonized description of the PLATOON data sources; it provides machine-readable documentation of the PLATOON, which PLATOON services like the metadata registry can consume. Moreover, network analysis over the RDF graph suggests relatedness among PLATOON datasets based on the annotations from the PLATOON Semantic Data Models.
Keyword List:	Data Harmonization, Semantic Data Models, DCAT, IDS

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

Editor(s):	Maria-Esther Vidal (TIB), Ahmad Sakor (TIB), Enrique Iglesias (TIB), Mazen Bechara (TIB), Gabriela Ydler (TIB)
Contributor(s):	Valentina Janev (IMP)
Reviewer(s):	Carsten Draschner (UBO) Farshad Bakhshandegan (UBO) Philippe Calvez (ENGIE) Erik Maqueda (TECN)
Approved by:	
Recommended/mandatory readers:	WP3, WP4, WP5, and WP6

Document Description

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v0.1	25/01/2022	Draft of Table of Content	Ahmad Sakor (TIB)
v0.2	31/01/2022	Updated Table of Content	Ahmad Sakor (TIB)
v0.3	23/02/2022	Preliminaries section	Ahmad Sakor and Enrique Iglesias (TIB)
v1.0	03/03/2022	First version for internal review	Ahmad Sakor (TIB), Enrique Iglesias (TIB), Mazen Bechara (TIB), Maria-Esther Vidal (TIB)
v.1.1	08/03/2022	Internal Review	Carsten Draschner (UBO)
v.1.2	13/03/2022	Version addressing comments from internal review	Maria-Esther Vidal (TIB) Gabriela Ydler (TIB)
v.1.2	22/03/2022	Internal Review	Valentina Janev (IMP)
v.1.3	22/03/2022	Internal Review	Erik Maqueda (TECN)
v2.0	22/03/2022	Version addressing comments from internal review	Maria-Esther Vidal (TIB)
v3.0	30/03/2022	Final version for Submission	Ahmad Sakor and Maria-Esther Vidal (TIB)

Table of Contents

List of Figures

List of Tables

Terms and Abbreviations

API	Application Programming Interface
BI	Business Intelligence
BOT	Building Ontology
CA	Consortium Agreement

CIM	Common Information Model
CO	Confidential
CSV	Comma-separated Values
DCAT	Data Catalog Vocabulary
DM	Dissemination Manager
DQV	The Data Quality Vocabulary
EC	European Commission
EM	Exploitation Manager
EU	European Union
GA	Grant Agreement
GAM	General Assembly Meeting
H2020	Horizon 2020
IDS	International Data Spaces
JSON	JavaScript Object Notation
KPI	Key Performance Indicators
ODRL	The Open Digital Rights Language
OEMA	Ontology for Energy Management Applications
OJM	Object Join Map
ORG	The Organization Ontology
ORM	Object Reference Map
OWL	Web Ontology Language
PJTT	Predicate Join Tuple Table
PM	Project Manager
PROV	The Provenance Ontology
PTT	Predicate Tuple Table
PU	Public
QA	Quality Assurance
R2RML	Relational Database to RDF
RDF	Resource Description Framework
RML	RDF Mapping Language
RPC	Remote Procedure Call
SAREF	Smart Appliances REFerence ontology
SEAS	Smart Energy Aware Systems
SHACL	Shapes Constraint Language
SKOS	Simple Knowledge Organisation System
SOM	Simple Object Map
SPARQL	SPARQL Protocol and RDF Query Language
SSN	Semantic Sensor Network
SWG	Sub Working Group
T	Task
TSV	Tab Separated Values
URI	Uniform Resource Identifier
VoID	Vocabulary of Interlinked Datasets
W3C	World Wide Web Consortium
WP	Work Package
WPL	Work Package Leader
XML	Extensible Markup Language

Executive Summary

This document reports on performing task T5.3 of WP5 - Data Collection and Harmonization. It presents a methodology and pipeline for ensuring the interoperability of the PLATOON data sources. Existing vocabularies from the International Data Space (e.g., Data Catalog-DCAT) are utilized to provide the unified description of the data sources. Moreover, annotations from the PLATOON data models facilitate the definition of the data sources in terms of energy concepts like wind power systems, solar power systems, conventional power plants, cooling, heating, lighting systems, and smart grids. A pipeline that resorts to PLATOON semantic connectors performs this data harmonization task. The description the PLATOON data sources put in perspective the role of recommended vocabularies and domain data models in resolving data heterogeneity. In the future, the harmonized descriptions of the PLATOON data sources will be used to populate the PLATOON data marketplace catalog via the metadata registry developed in WP3. Moreover, the pipeline described in this document will be applied to integrate a unified description of the PLATOON analytical tools.

1. Introduction

1.1 Purpose and Scope of the Document

This document describes strategies for creating a harmonized description of the PLATOON data sources. Methodological strategies complement the work done in T2.4 to analyze the PLATOON data sources in terms of the concepts from the energy domain that better characterize a dataset. A questionnaire facilitates the participation of the data providers in linking the PLATOON data sources with concepts in the semantic data models. Knowledge extracted from the questionnaires is used to populate a relational database; a set of mapping rules declaratively define the harmonized description of the data sources based on the Data Catalog Vocabulary (DCAT). As a result, an RDF graph representing a unified description of the PLATOON data sources has been created. Semantic connectors are under development to integrate the harmonized descriptions with other services of the PLATOON platform (e.g., the metadata registry from WP3 and the analytical toolbox for WP4). The reported outcomes of task 5.3 show the relevance of the PLATOON semantic data models defined in T2.3 and the role they play in establishing a common understanding of the meaning of concepts that characterize the energy sector. Moreover, the observed results provide evidence of the potential that harmonizing schemas using recommended vocabularies and semantic data models brings into interoperability resolution.

Six sections compose this document. Section 2 presents preliminaries on concepts; it includes concepts like data integration systems, the international data space, mapping languages, and semantic data models. Section 3 sketches a methodology for data harmonization and defines the pipeline of generating a unified definition of the PLATOON data sources using DCAT and the PLATOON data models. Section 4 reports on the results of executing the pipeline and the harmonized description of the pilots' data sources is reported in Section 5. Finally, the conclusions and next steps are outlined in Section 6.

1.2 Relationship with Other Documents

This document is related to two deliverables in WP2: i) D2.1 where the PLATOON reference architecture is defined; and ii) D2.3 where the PLATOON common data models for energy are presented, ii) D2.4 where the PLATOON data sources are reported; and D2.8 where the data semantic data adapters are presented.

2. Preliminaries

2.1 Data Integration Systems

Data integration is the process of combining data from various sources into a single, coherent view. The first phase in an integration process is ingestion, which is followed by cleansing, mapping, and transformation. Data integration allows analytics technologies to give relevant, actionable business intelligence into the end. There is no such thing as a one-size-fits-all solution when it comes to data integration.

The importance of data integration comes from the need to integrate big data for enterprises. With all of its benefits and challenges, big data is being embraced by enterprises that want to stay competitive and relevant. Data integration enables querying in these massive databases, with benefits ranging from corporate intelligence and consumer data analytics to data enrichment and real-time data delivery. The management of company and customer data is one of the most common use cases for data integration services and solutions. To provide corporate reporting, business intelligence (BI data integration), and advanced analytics, enterprise data integration feeds integrated data into data warehouses or virtual data integration architecture. Customer data integration gives a holistic picture of key performance indicators (KPIs), financial risks, customers, manufacturing and supply chain operations, regulatory compliance activities, and other areas of business processes to business managers and data analysts. In the energy sector industry, data integration is extremely vital. By arranging data from several systems into a single perspective of relevant information from which helpful insights can be gained, integrated data from various equipment and actors across the value chain (from generation to transport/distribution and end use of energy) allows to optimise the global operation of the system. Interoperability is the term used to describe the exchange of data across various systems. This is key for a successful energy transition where the energy system is pivoting towards a more complex decentralised system formed of heterogeneous systems and actors.

A data integration system (DIS) integrates two or more data sources [1] [2]. A DIS comprises a unified schema to provide a reconciled view of all data available in integrated different data sources. Mappings between the unified schema and data source schemas need to be defined to combine data in the data sources. Formally, a DIS is defined as a triple, i.e., DIS=<O,S,M>, where O stands for a unified ontology, and S and M represent sets of sources and mapping rules, respectively. In Task T5.3, a DIS is utilized to generate a uniform description of the project catalogs and datasets. The controlled vocabularies from the International Data Space (IDS) and the Data Catalog (DCAT) vocabulary correspond to the unified schema O. In contrast, data sources in S correspond to a relational database that collects the metadata that describes the project catalogs and datasets. R2RML[1] is the language recommended by World Wide Web Consortium (W3C) to describe mapping rules; it is utilized to declaratively specify how the datasets and catalogs are described in DCAT [3].

2.2 The International Data Space and its Information Model

The Information Model [4] is an RDFS/OWL-ontology that covers the essential ideas of International Data Spaces (IDS), i.e., the forms of digital content transferred by participants via IDS infrastructure components. In IDS, the Information Model principally aims at

[1] https://www.w3.org/TR/r2rml

characterizing, publishing, and detecting data products (Data Assets) and reusable data processing software (Data Apps). IDS primary resources, Data Assets and Data Apps, are referred to as resources in this document. It is ensured that only relevant materials are presented by using a systematic semantic annotation (i.e., resources appropriate to meet the requirements of the Data Consumer). Once the resources have been discovered, they can be traded and consumed in an automated manner using semantically specified service interfaces and protocol bindings. Aside from those fundamental commodities, the Information Model specifies the entities, players, infrastructure components, and processes that make up IDS.

The IDS Information Model [4] is a generic model that is not tied to a specific area. Domain modelling is delegated to domain-specific communities of the International Data Space, which provide shared vocabularies and data schemata. Figure 1 illustrates the three levels of the IDS information model; they enable a conceptual and declarative definition of the process of data exchange while data sovereignty for data providers is guaranteed. IDS propose a message-based infrastructure to enable the communication of the different nodes and components; it resorts to the Semantic Web standards to describe shared data source.

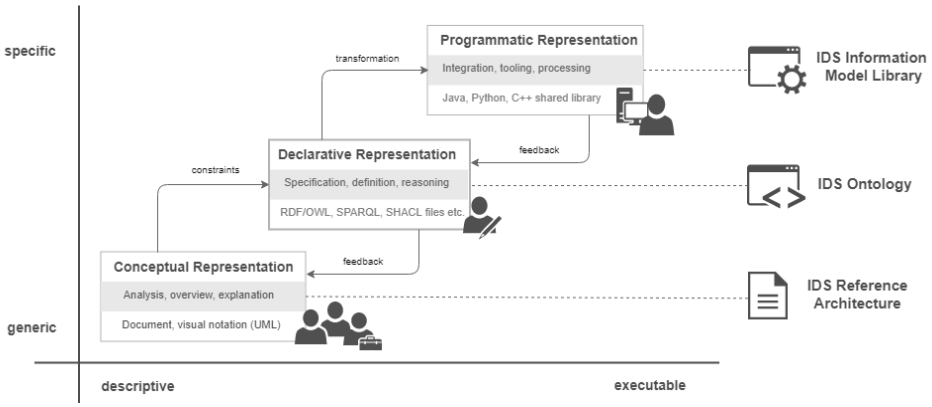

Figure 1 : Representations of the Information Model [4]. Figure taken from https://international-data-spaces-association.github.io/InformationModel/docs/index.html

2.3 Mapping Languages

A mapping language enables the definition of concepts in a unified schema in terms of data sources. R2RML and RDF Mapping Language (RML) [5] are exemplary mapping languages to transform data into RDF. R2RML is the World Wide Web Consortium (W3C) recommendation to represent the transformations from relational databases into RDF. RML is a mapping language that extends what is established in R2RML to include not only relational databases but also file data sources like CSV, XML, JSON, and TSV.

R2RML allows to express mapping rules only from data in relational databases to RDF data models. An R2RML mapping is comprised of one or more triples maps and occur on a Logical Table iterating row by row. A triples map is composed of three main parts:

- ***Logical Table*** is the table from the relational database from which the data will be extracted from.
- ***Subject Map*** defines the rule generates the unique identifiers (URIs) that is used for the subject value for all the triples created from a specific triples map.
- ***Predicate Object Map*** comprised of two parts:

- ***Predicate Map*** describes the rule that generates the predicate for the triple.
- ***Object Map*** defines the rule that creates the object for the triples.

A triple can have zero or more Predicate Object Maps.

The RDF mapping language (RML) [5] is a generic mapping language. RML is designed to express customized mapping rules from heterogeneous data structures and serializations to the RDF data model. This mapping language is defined as a superset of the mapping language R2RML, with the purpose of expanding its capabilities.

RML follows the rules established by R2RML. The main difference is that RML does not restrict itself to only relational databases but it is also capable of converting data from multiple data sources like CSV, JSON, XML, and TSV. For that reason, RML uses a logical source and not a logical table. The logical source indicates which data file or table from a relational database to use as a source for that triples map.

2.4 The PLATOON Semantic Data Models

The PLATOON semantic data models provide a formal specification of the meaning of energy domain concepts and their properties, helping, thus, to develop a common understanding of the domain. They comprise a semantically rich set of classes that extend existing ontologies in the energy sector, with concepts required to model requirements specific to the PLATOON pilots. Figure 2 (taken from [6]) illustrates these concepts. [7]

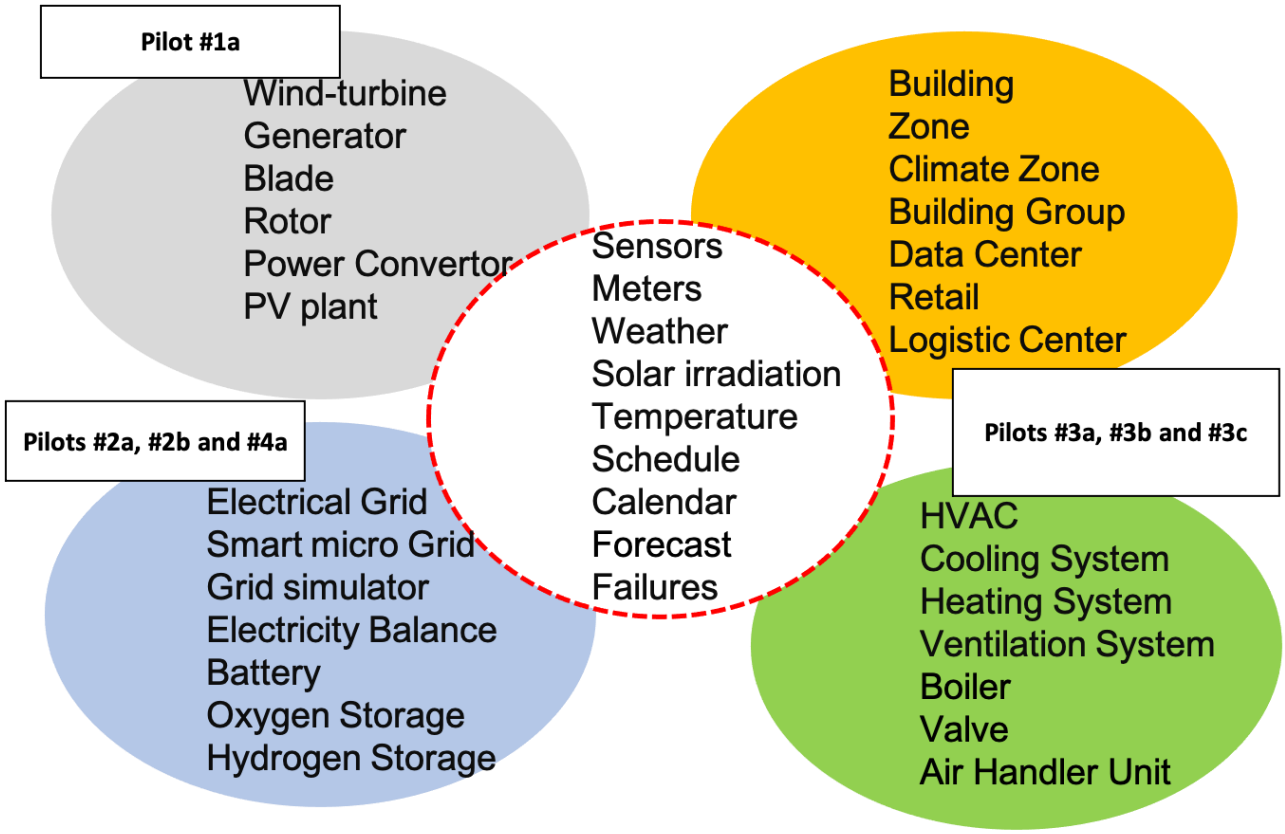

Figure 2: Main concepts in the PLATOON Semantic Data Models (Figure from D2.3). Classes and properties modeled in the PLATOON Semantic Data Models provide a common understanding of the energy domain.

The PLATOON semantic data models also comprise existing related ontologies:

a) **Smart Appliances REFerence ontology (SAREF[2]):** a modular ontology for the internet of things domain; it integrates a family of vocabularies to represent smart cities, buildings, energy, agriculture, food, and environmental.
b) **Smart Energy Aware Systems (SEAS[3]):** an ontology composed of concepts to model energy systems and their interactions. SEAS covers the following concepts: features of interest and their properties, evaluation of features, smart and microgrids, smart homes, electrical cars, electrical market, and weather forecast.
c) **Common Information Model (CIM[4]):** an ontology composed of concepts to describe power grids. CIM includes concepts to describe: Grid-related domains transmission, distribution, micro-grids, resource connected-to-the-grid domain, electrical transportation, smart metering, and asset management.
d) **Other ontologies:** Several other domain ontologies of the energy sector have been proposed such as Semantic Sensor Network (SSN[5]), Ontology for Energy Management Applications (OEMA[6]), Building Ontology (BOT[7]), and Semanco[8].

2.5 Mapping Rule Engines

Knowledge graphs are generated from data sources by following the rules established by a mapping. One of the more commonly used mapping formats is RML. Knowledge graphs are generated from an RML mapping when using a mapping rule engine with an RML interpreter. The SDM-RDFizer [7] is the mapping rule engine used for this report.

2.5.1 SDM-RDFizer

The SDM-RDFizer is an RML engine capable of creating a knowledge graph from multiple data sources formats like CSV, TSV, JSON, and XML and relational databases like MySQL and Postgres following the rules established in an RML mapping. The SDM-RDFizer implements the data structures Predicate Tuple Table (PTT) and Predicate Join Tuple Table (PJTT) to execute duplicated removal and joins more efficiently. The PTT stores for each predicate P the triples that have been generated so far. The key encodes the subject and object of the triple. The PJTT stores the subjects of the triples generated by a join. PJTT is an index hash table where the key encodes each value of the attributes in the join condition. The value is a set of subject values in the second source associated with the values of the attributes in the hash key. Additionally, the physical operators Simple Object Map (SOM), Object Reference Map (ORM), and Object Join Map (OJM) have been implemented in the SDM-RDFizer. SOM generates an RDF triple from the execution of a simple Predicate Object Map. Each generated triple is checked against the associated PTT. If the triple was already generated before, it is discarded. If not, it will be added to the knowledge graph and the PTT will be updated accordingly. ORM extends SOM by using the subject of one Triples Map as the object of another Triples Map. The condition for this operator to work is that both Triples

[2] https://saref.etsi.org/
[3] https://ci.mines-stetienne.fr/seas/index.html
[4] https://ontology.tno.nl/cerise/cim-profile/
[5] https://www.w3.org/TR/vocab-ssn/
[6] https://innoweb.mondragon.edu/ontologies/oema/index-en.html
[7] https://w3c-lbd-cg.github.io/bot/
[8] http://www.semanco-tools.eu/urban-enery-ontology

Maps must use the same data source. Afterward, the same process as with SOM is followed, i.e., the triples are checked against to the corresponding PTT to avoid duplicate triples. The OJM is an extension of ORM with the difference where the Triples Maps can be defined over different data sources and there exists a join condition between them. The corresponding PJTT is used in an index join where the outer table corresponds to the child map and the inner table to the PJTT. If an entry *e* exists with the same hash key, all the subjects in *e* are used to generate the resulting RDF triples. Finally, a similar procedure as before is followed to avoid the generation of duplicate triples. Iglesias et al. provide a more detailed description of the operators implemented by SDM-RDFizer.

3. Methodology and Pipeline for Data Harmonization

This section describes the methodology defined for data harmonization and the pipeline developed by the TIB team to generate a unified description of the PLATOON data sources. This section will also illustrate the expressivity power of the IDS information model in the context of PLATOON data harmonization.

3.1 Methodology for Data Harmonization

Figure 3 depicts the steps of the methodology followed to generate the descriptions of the PLATOON data sources; it involves three types of users: a) **data providers** who correspond to the partners of the PLATOON pilots; b) **knowledge engineers** and **domain experts** who are the experts in knowledge representation, and the PLATOON semantic data models and IDS information model; and **software developers** who implement the programs required to perform the data harmonization and access the PLATOON data source descriptions. This methodology is iteratively applied because some sources may change over time.

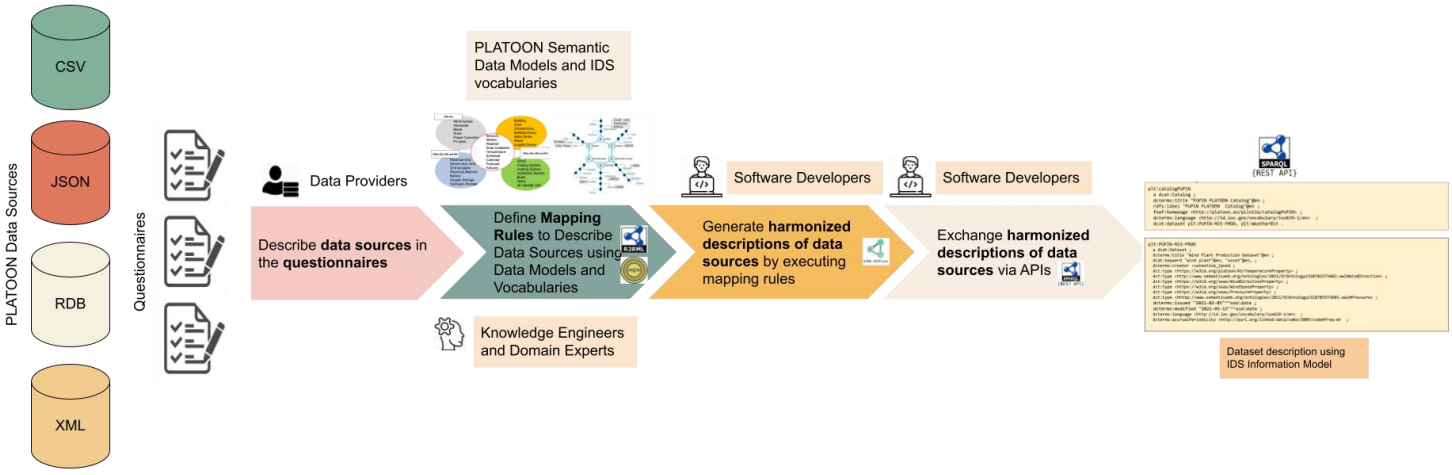

Figure 3: Methodology followed to generate Harmonized Descriptions of the PLATOON Data Sources

The integration methodology is composed of the following steps:

1. **Description of the PLATOON data sources**.

A questionnaire allows partners of the project who are data providers to describe their data sources. These questionnaires are composed of the following five parts: **Overview:** collects a general description of a dataset; **Big data Vs**: the dataset is defined in terms of 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; **Dataset detailed features**: outlines the main characteristics of the data in a data 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 dataset can be utilized and the coverage of the data set. In total, a questionnaire comprises 30 questions. Deliverable D2.4 presents a detailed description of the questionnaires and the answers provided by the data providers.

2. **Definition of the Mapping Rules**.

A second questionnaire enables the definition of the main concepts that characterize the data included in a data source. These concepts are mapped to classes and relationships in the PLATOON data sources. A relational database is created with metadata about data sources extracted from these questionnaires. Knowledge engineers (e.g., from TIB and ENGIE) rely on these questionnaires to define mapping rules using a declarative mapping language (e.g., R2RML, RML, or SPARQL-Generate). A detailed description of the questionnaires and the answers provided by the partners is presented in D2.4.

3. **Generation of Harmonized Data Source Descriptions.**

The metadata describing the PLATOON data sources is represented as a structured database (e.g., relational), and the execution of the mapping rules transforms this metadata into an RDF knowledge base. Vocabularies from the IDS information models (e.g., DCAT) and the PLATOON semantic data modes are used to describe this metadata. A semantic connector (e.g., implemented using SDM-RDFizer) executes the mapping rules, creates the RDF knowledge base, and uploads the knowledge base into a Virtuoso SPARQL endpoint.

4. **Exchange of Harmonized Data Source Descriptions**.

The RDF knowledge base created by the execution of the mapping rules is traversed via the execution of the SPARQL queries posted against the Virtuoso SPARQL endpoint. Rest APIs are implemented (e.g., by software developers) to allow for the retrieval of a harmonized description of the PLATOON data sources. These APIs allow for the population of the metadata registry (implemented in WP3) with metadata collected from the partners.

3.2 Data Source Characterization Via Questionnaires

As previously described, metadata about the PLATOON data sources have been collected from the data providers via questionnaires. As a result, a relational database has been created. It is composed of nine tables that the PLATOON catalogs, the data sources provided by each partner, and the classes from the PLATOON semantic data models that annotate these data sources. Figure 4 depicts a *snapshot* of this database.

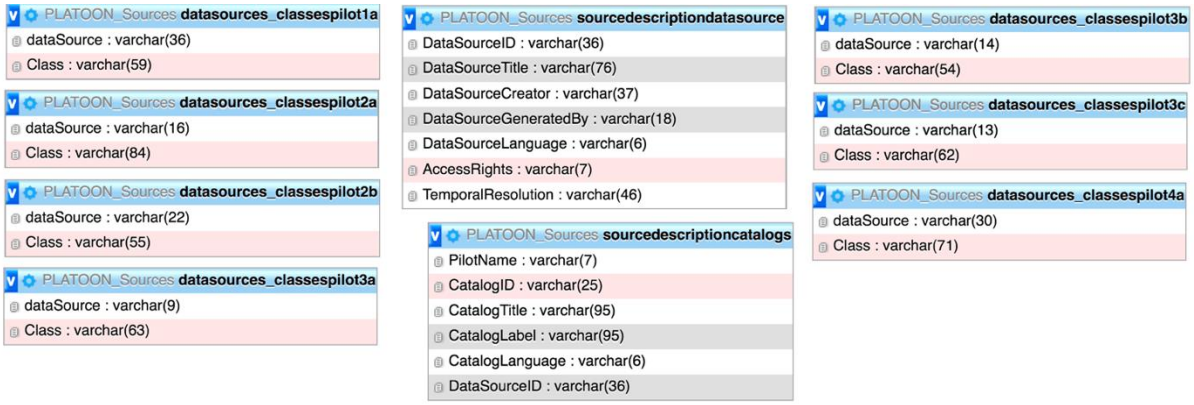

Figure 4: Structured Representation in a Relational Database of the Metadata collected via Questionnaires

3.3 Metadata and controlled vocabularies

Figure 4 depicts the proposed W3C standards that are used for expressing the meaning and content of data in the international data spaces. The figure presents a dataspace in terms of Content, Concept, Community of trust, Commodity, Communication, and Context. Different vocabularies are used to describe each dimension based on its nature:

- The content is represented using DCAT [3], VoID [8], DataCube [9], and SHACL [10]. DCAT standardizes the classes and properties that characterize a catalog of datasets and data services. VoID is an RDF schema vocabulary that expresses metadata about RDF datasets; it aims at serving as a link between RDF data publishers and users. The DataCube vocabulary allows multi-dimensional data, such as statistics, to be published in a way that may be linked to other datasets and concepts. SHACL is a language for checking RDF graphs against a set of rules. These rules are expressed in the form of an RDF graph as shapes and other constructs.
- SKOS [11] is used to express the concept dimension; it is an RDF vocabulary for representing the underlying structure and content of controlled vocabulary schemes such as taxonomies and thesauri.
- The ORG vocabulary [12] is used to represent the community of trust dimension. ORG is an ontology for organizational structures with the goal of enabling linked data publishing of organizational data across multiple domains. It is built to accommodate domain-specific extensions that provide organization and job classification, as well as extensions that support nearby information like organizational activities.
- PROV [13], ODRL [14], and DQV [15] vocabularies are used to describe the commodity covering the provenance, policy, and quality dimensions. The PROV Ontology (PROV-O) uses the OWL2 Web Ontology Language to represent the PROV Data Model (OWL2). It defines classes, characteristics, and constraints used to represent and exchange provenance information created in various systems and settings. It can also be specialized in order to generate new classes and features to model provenance information for other applications and domains. The Open Digital Rights Language (ODRL) is a policy expression language that offers a dynamic and interoperable information model, vocabulary, and encoding techniques for representing statements concerning material and service usage. DQV is a method for describing the quality of a dataset, whether by the dataset provider or by a wider community of users. It does not provide a formal, exhaustive definition of quality; rather, it establishes a standard mechanism of providing information so that a potential user of a dataset can make his or her own judgment about its suitability for purpose.
- WGS84 [16] and OWL-Time [17] vocabularies are used to describe the context covering the space and time dimensions. WGS84 is a vocabulary used in the WGS84 geodetic reference to represent latitude, longitude, and altitude information. OWL-Time is an OWL-2 DL ontology of temporal concepts used to represent the temporal aspects of resources in the real world or in Web pages. The ontology defines a language for describing facts about topological (ordering) relationships between instants and intervals, as well as information about periods and temporal location, including date-time information. Time positions and durations can be described using the standard (Gregorian) calendar and clock, or another chronological reference system such as Unix-time, geologic time, or alternative calendars.

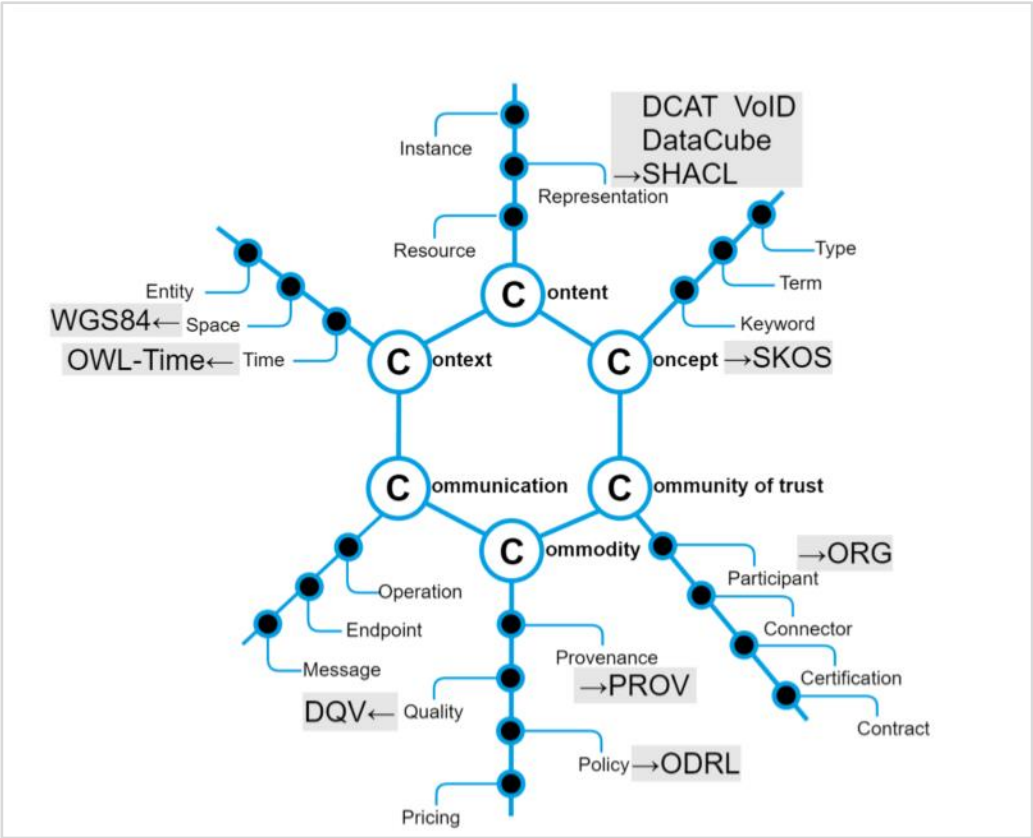

Figure 5: Proposed W3C standards to express meaning and content in International Data Spaces. The figure is taken from Bader et al [4]: Standards like SHACL, SKOS, and PROV provide a unified way to describe DEs in terms of content, concepts, and provenance.

The Data Catalog Vocabulary (DCAT)[9] provides a common understanding of the classes and properties that describe a catalog of datasets and data services. DCAT is expressed in RDF (Resource Description Language) and facilitates the unified representation of a catalog main properties in the way that it can be understandable by humans and also by machines. Figure 5 depicts the main classes and properties from DCAT; they include classes from other vocabularies, e.g., foaf:Agent, skos:Concept, or skos:ConceptSchema.

[9] https://www.w3.org/TR/vocab-dcat-2/

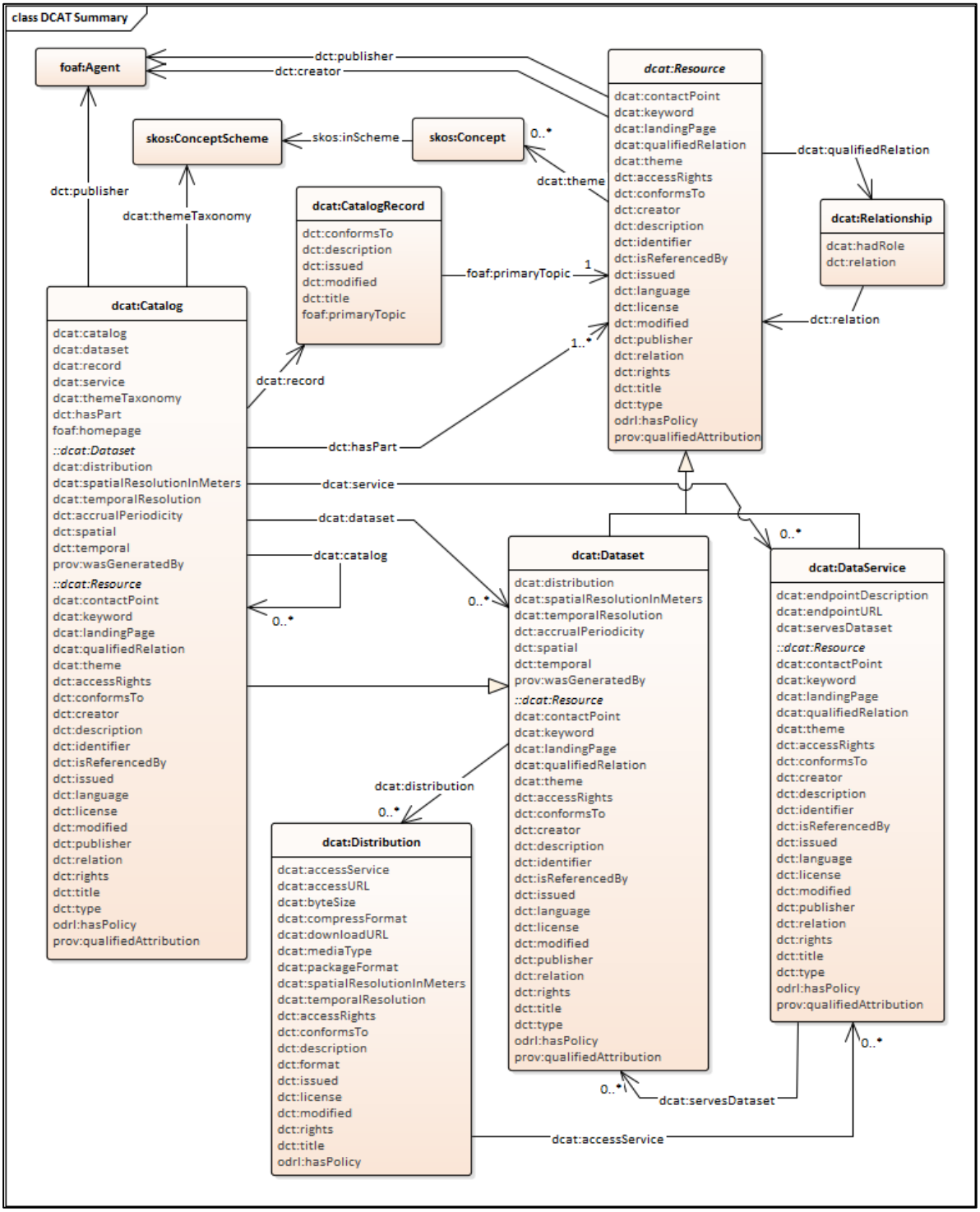

Figure 6: Concepts in the DCAT vocabulary. Figure taken from https://www.w3.org/TR/vocab-dcat-2/

The main classes from DCAT are as follows:

- **dcat:Catalog** is a set of datasets; each item is a metadata entry describing a resource. The scope of dcat:Catalog is collections of metadata about datasets or data services.
- **dcat:Resource** represents a dataset, a data service, or any resource that can be described by a metadata record in a catalog.
- **dcat:Dataset** describes a dataset. A dataset is a set of data that a single person has published. Numbers, words, pixels, imagery, sound, other multi-media, and potentially other types of data are all examples of data collected into a dataset.

- **dcat:Distribution** defines a dataset in an accessible format (downloaded file).
- **dcat:DataService** represents a data service. A data service is a set of operations that allow access to one or more datasets or data processing services via an interface.
- **dcat:CatalogRecord** represents a catalog's metadata item, primarily relating to registration information, such as by whom the item is created and when.

Classes **dcat:Catalog** and **dcat:Dataset** model catalogs and datasets of the PLATOON project. For each catalog, the partners filled in a questionnaire describing the title, identifier, and language used in the catalog. Similarly, partners described data sources by the means of title and identifier of the data source, the creator of the data source, language, access rights (public/private), and the temporal resolution of the data Values of these attributes correspond to the extracted metadata; they are described with the following properties:

- **dcat:dataset** is used to list all the datasets related to a catalog.
- **dct:title** represents the literal title (name) of a catalog.
- **dct:language** refers to language used for textual metadata (i.e., titles) of a cataloged resource. It can also be used to represent the language of the textual values of a dataset distribution. The language is represented using *ISO 639-1* two-letter code.
- **dct:creator** models the entity that creates the resource (i.e., ENGIE).
- **dct:accessRights** includes information about who can access the resource. It can also be used as an indicator of its security status.
- **dcat:temporalResolution** describes the minimum spacing of items within the dataset. The value is given using the standard datatype **xsd:decimal**.
- **dcat:keyword** is used to give a resource a tag or keyword.
- **dct:type** is used to indicate the type or genre of a dataset. In PLATOON, this property is used to annotate the dataset with a class from the PLATOON semantic data model, e.g., <https://w3id.org/platoon/WindFarm>, <http://www.iec.ch/TC57/CIM#WindPlantIEC>, or <https://w3id.org/platoon/OffshoreWindTurbine>.

Figure 7 illustrates a description of the dataset **PUPIN-RES-PROD** using the DCAT vocabulary. The DCAT properties **dcat:dataset, dct:title, dct:language, dct:creator, dct:accessRights,** and **dcat:keyword** express the basic properties of the dataset**.** On the other hand, the property <http://purl.org/dc/terms/type> provides a harmonized description of this dataset by the means of the PLATOON semantic data models.


```
<https://w3id.org/platoon/entity/EBPM> <http://www.w3.org/ns/dcat#dataset> <https://w3id.org/platoon/entity/PUPIN-RES-PROD>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/ns/dcat#Datas
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/title> "Historical Wind Power Production Measurements".
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://www.w3.org/ns/prov#wasGeneratedBy> "IMP".
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/creator> "IMP".
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://www.w3.org/ns/dcat#temporalResolution> "MultipleYears".
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/accessRights> "Private".
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/language> <http://id.loc.gov/vocabulary/iso639-1/ENG>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://www.w3.org/ns/dcat#keyword> "Wind Power Production"@en
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/language> <http://id.loc.gov/vocabulary/iso639-1/ENG>

<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/platoon/WindFarm>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <http://www.w3.org/2006/time#Instant>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <http://www.w3.org/2006/time#TemporalEntity>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/seas/FeatureOfInterest>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/seas/ElectricPowerSystem>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/seas/ElectricPowerProducer>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <http://www.semanticweb.org/ontologies/2011/9/Onto
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <http://www.iec.ch/TC57/CIM#WindGeneratingUnit>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <http://www.iec.ch/TC57/CIM#WindPlantIEC>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/platoon/OffshoreWindTurbine>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <http://www.iec.ch/TC57/CIM#GeneratingUnit>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <http://www.iec.ch/TC57/CIM#Plant>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/platoon/AirTemperatureProperty>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/platoon/AirTemperatureEvaluation
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/seas/WindDirectionProperty>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <http://www.semanticweb.org/ontologies/2011/9/Onto
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/seas/WindDirectionEvaluation>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/platoon/WeatherStation>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/platoon/ForecastOfElectricProduc
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/platoon/ForecastOfActivePower>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/seas/ElectricPowerProperty>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://saref.etsi.org/core/Power>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <http://www.iec.ch/TC57/CIM#ActivePower>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/platoon/ForecastOfActivePowerEva
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/platoon/ForecastOfElectricPowerE
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/seas/ElectricPowerEvaluation>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/platoon/ActivePowerEvaluation>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <http://www.w3.org/2006/time#Interval>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/platoon/SolarInsolationProperty>
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/seas/PressureProperty>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <http://www.semanticweb.org/ontologies/2011/9/Onto
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/seas/PressureEvaluation>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/seas/WindSpeedProperty>.
<https://w3id.org/platoon/entity/PUPIN-RES-PROD> <http://purl.org/dc/terms/type> <https://w3id.org/seas/WindSpeedEvaluation>.
```

Figure 7: Description of a Data Source Using DCAT. Annotations are represented using the property <http://purl.org/dc/terms/type> and classes from the PLATOON Semantic Data Models

The DCAT descriptions of the catalogs and datasets shared by the PLATOON partners, compose a knowledge base, which can be accessible via an RDF triple store (e.g., Virtuoso). For example, the execution of the following SPARQL query retrieves pilot 2a datasets and their annotations from the PLATOON semantic data models.

```
SELECT DISTINCT ?pilot ?dataset ?annotation
WHERE { ?catalog      a  <http://www.w3.org/ns/dcat#Catalog> .
?catalog      <http://www.w3.org/ns/dcat#dataset>   ?dataset.
?catalog      <http://www.w3.org/ns/dcat#keyword> ?pilot  .
?dataset     <http://purl.org/dc/terms/type>              ?annotation .
FILTER  regex(?pilot,"2a" ) }
```

pilot	dataset	annotation
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	http://www.w3.org/2006/time#Interval
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	http://www.w3.org/2006/time#TemporalEntity
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	http://www.w3.org/2006/time#Instant
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/seas/FeatureOfInterest
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/seas/ElectricPowerSystem
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/seas/ElectricPowerProducer
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	http://www.iec.ch/TC57/CIM#GeneratingUnit
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	http://www.iec.ch/TC57/CIM#Plant
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/seas/Zone
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/seas/SolarArray
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/platoon/AirTemperatureProperty
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/platoon/AirTemperatureEvaluation
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/seas/WindDirectionProperty
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	http://www.semanticweb.org/ontologies/2011/9/Ontology1318785573683.owl#WindDirection
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/seas/WindDirectionEvaluation
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/platoon/SolarInverter
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/platoon/WeatherStation
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/seas/ElectricPowerProperty
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://saref.etsi.org/core/Power
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	http://www.iec.ch/TC57/CIM#ActivePower
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/platoon/SolarInsolationProperty
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/platoon/SolarInsolationEvaluation
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/seas/SolarPanel
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	http://www.iec.ch/TC57/CIM#SolarGeneratingUnit
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/seas/WindSpeedProperty
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PV	https://w3id.org/seas/WindSpeedEvaluation
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	http://www.w3.org/2006/time#Interval
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	http://www.w3.org/2006/time#TemporalEntity
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	http://www.w3.org/2006/time#Instant
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	https://w3id.org/seas/FeatureOfInterest
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	https://w3id.org/platoon/WindFarm
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	https://w3id.org/seas/ElectricPowerSystem
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	https://w3id.org/seas/ElectricPowerProducer
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	http://www.semanticweb.org/ontologies/2011/9/Ontology1318785573683.owl#WindTurbine
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	http://www.iec.ch/TC57/CIM#WindGeneratingUnit
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	http://www.iec.ch/TC57/CIM#WindPlantIEC
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	https://w3id.org/platoon/OffshoreWindTurbine
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	http://www.iec.ch/TC57/CIM#GeneratingUnit
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	http://www.iec.ch/TC57/CIM#Plant
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	https://w3id.org/platoon/AirTemperatureProperty
"Pilot2a"@en	https://w3id.org/platoon/entity/PUPIN-RES-PROD	https://w3id.org/platoon/AirTemperatureEvaluation

Figure 8: Result of a SPARQL query over the DCAT description of PLATOON catalogs and datasets

3.4 Generic Pipeline for Data Harmonization

The pipeline presented in Figure 9 is generic and can be implemented by different semantic connectors. However, the one reported in this document is developed with the SDM-RDFizer. The pipeline makes use of metadata describing the schema of PLATOON schema of the data sources and guided by mapping rules, generates a harmonized description in terms of the vocabularies DCAT and IDS, as well as the PLATOON semantic data models. Then, the pipeline uploads the generated RDF knowledge base into a Virtuoso SPARQL endpoint.

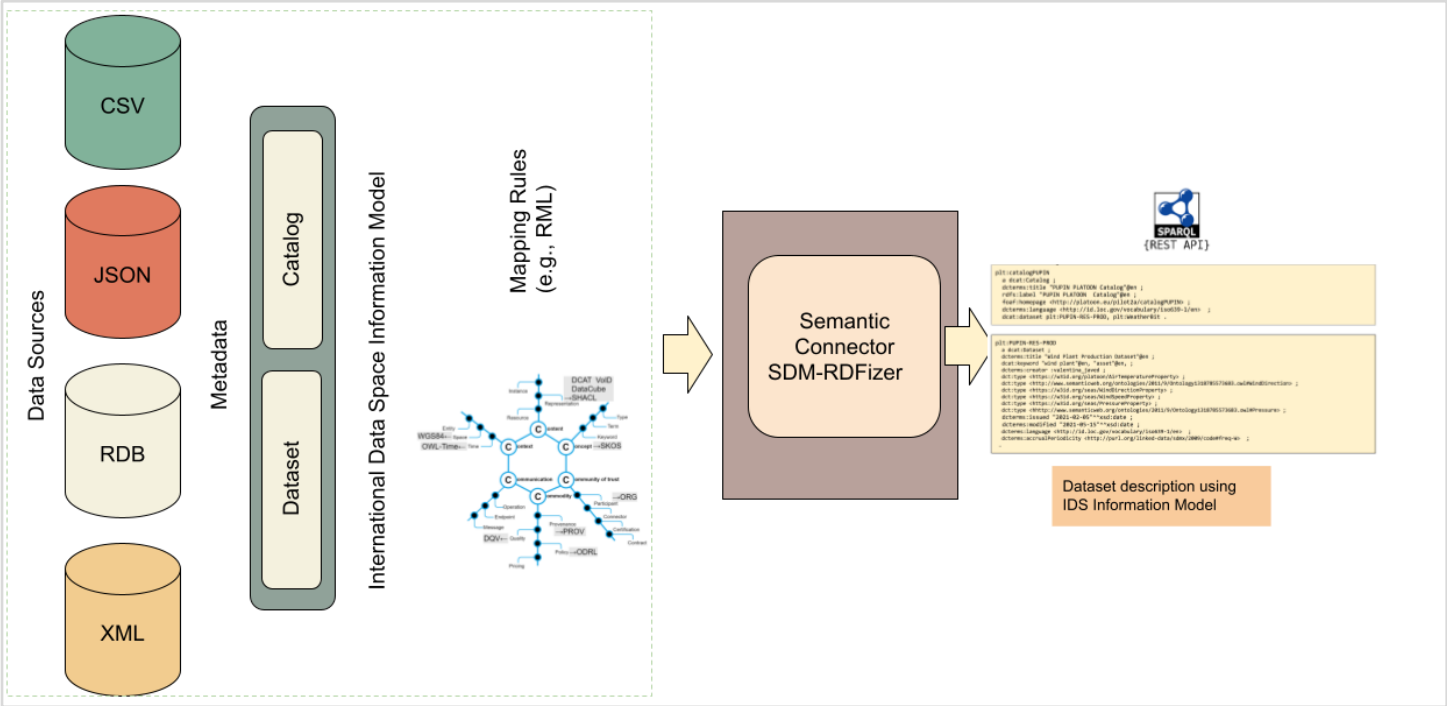

Figure 9: Pipeline for generating a harmonized description of the PLATOON data sources

The pipeline is executed as a bash script that performs a series of docker images; each implements a different component of the pipeline. As a result, the pipeline is comprised of two docker images, one for SDM-RDFizer and another for the Virtuoso triple store. This pipeline is available as open source in a GitHub repository[10].

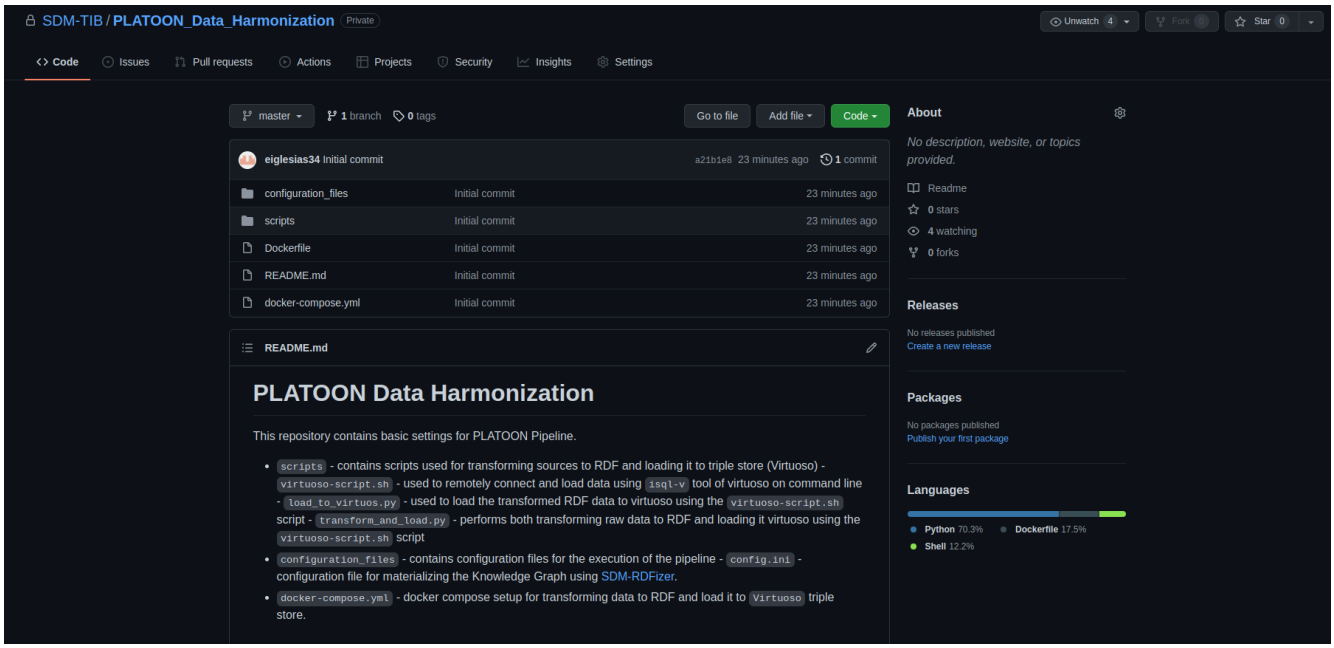

Figure 10: GitHub Repository of the PLATOON Data Harmonization - This repository contains all necessary files and scripts for the execution of the PLATOON Data Harmonization pipeline. Included are the configuration file for the SDM-RDFizer, the docker-compose.yml

[10] https://github.com/SDM-TIB/PLATOON_Data_Harmonization

The aforementioned GitHub repository contains required files and scripts for the execution of the PLATOON pipeline. These components are:

- ***Scripts:*** is a folder containing the scripts that transform mapping files and their corresponding data sources into RDF data, which is then loaded into the triple store (Virtuoso). These scripts are:
  - *Virtuoso-script.sh:* used to remotely connect and load data using *isql-v* tool of virtuoso.
  - *Load_to_virtuoso.py:* uses the *virtuoso-script.sh* to upload transformed RDF data to virtuoso triples store.
  - *Transform_and_load.py:* performs both the transformation of raw data into RDF data by invoking the SDM-RDFizer and uploads the data into a triples store by using *virtuoso-script.sh.* Figure 11 presents the code of the script.

```
#!/usr/bin/env python3

import os
import sys
import logging
import traceback
import getopt
from configparser import ConfigParser, ExtendedInterpolation

from rdfizer.semantify import semantify

logFormatter = logging.Formatter("%(asctime)s [%(threadName)-12.12s] [%(levelname)-5.5s]  %(message)s")
logger = logging.getLogger()

def transform(configfile, script="/data/scripts/virtuoso-script.sh"):
    config = ConfigParser(interpolation=ExtendedInterpolation())
    config.read(configfile)
    try:
        logger.info("Transforming data using " + str(configfile) + " configuration...")
        outputfolder = config["datasets"]["output_folder"]
        semantify(configfile)
        status = os.path.exists(outputfolder)
        if status:
            virtuosoIP = os.environ["SPARQL_ENDPOINT_IP"]
            virtuosoUser = os.environ["SPARQL_ENDPOINT_USER"]
            virtuosoPass = os.environ["SPARQL_ENDPOINT_PASSWD"]
            virtuosoPort = os.environ["SPARQL_ENDPOINT_PORT"]
            virtuosoGraph = os.environ["SPARQL_ENDPOINT_GRAPH"]
            outputfolder = os.environ["RDF_DUMP_FOLDER_PATH"]
            try:
                os.system( str(script) + " " + virtuosoIP + " " + virtuosoUser + " " + virtuosoPass + " " + virtuosoPort + " " + virtuosoGraph + " " + outputfolder)
                logger.info("Semantification sucessful!")
            except Exception as ex:
                logger.error("ERROR while loading data to viruoso! " + str(ex))
                exc_type, exc_value, exc_traceback = sys.exc_info()
                emsg = repr(traceback.format_exception(exc_type, exc_value, exc_traceback))
                logger.error("Exception while semantifying ... " + str(emsg))

        else:
            logger.error("Error during semantification of data. Please check your configureation file.")

    except Exception as e:
        exc_type, exc_value, exc_traceback = sys.exc_info()
        emsg = repr(traceback.format_exception(exc_type, exc_value,
                                               exc_traceback))
        logger.error("Exception while semantifying ... " + str(emsg))
```

Figure 11: Portion of the transform_and_load.py script – This figure illustrates the portion of the transform_and_load.py script that uploads the transformed RDF data into the SPARQL endpoint.

- ***Configuration_files:*** contains the configuration files that are necessary for the execution of the SDM-RDFizer; Figure 12 presents an example of a configuration.

```
[default]
main_directory: /data

[datasets]
number_of_datasets: 1
output_folder: ${default:main_directory}/rdf-dump
all_in_one_file: yes
remove_duplicate: yes
name: test
enrichment: yes
dbtype: mysql
ordered: yes
large_file: false

[dataset1]
name: test
mapping: ${default:main_directory}/mappings/data.ttl
```

Figure 12 Example of configuration file for the SDM-RDFizer –This file includes the location of the mapping files and output folder, credentials for accessing relational databases.

- ***Docker-compose.yml:*** docker compose set up file for the creation of the docker image of the SDM-RDFizer, and Virtuoso. Figure 13 presents a portion of the compose file.

```yaml
version: "3.3"
services:
  sdmrdfizer:
    image: asakor/sdmrdfizier:4.0.1
    hostname: sdmrdfizer
    container_name: sdmrdfizer
    domainname: platoon
    volumes:
      - .:/data
    networks:
      - platoon
    depends_on:
      - pilot2akg
    environment:
      - SPARQL_ENDPOINT_IP=dataharmonization
      - SPARQL_ENDPOINT_USER=dba
      - SPARQL_ENDPOINT_PASSWD=dba
      - SPARQL_ENDPOINT_PORT=1111
      - SPARQL_ENDPOINT_GRAPH=http://platoon.eu/DataHarmonization/KG
      - RDF_DUMP_FOLDER_PATH=/data

  pilot2akg:
    image: kemele/virtuoso:6-stable
    hostname: dataharmonization
    container_name: dataharmonization
    domainname: platoon
    volumes:
      - ./rdf-dump:/data
    ports:
      - "8891:8890"
      - "1116:1111"
    networks:
      - platoon

networks:
  platoon:
    external: false
```

Figure 13: Screenshot of the docker-compose.yml file – This figure illustrates the docker compose script used to generate the docker image components for the pipeline.

4. Harmonized Description of the PLATOON Data Sources

This section reports on analyzing the harmonized description of the PLATOON data sources generated by the pipelined depicted in Figure 9. By the day of submitting this deliverable, there are seven catalogs (one per pilot) and 34 datasets (on average, 4.71 datasets per catalog). Moreover, there are 333 annotations from the PLATOON semantic data models, which makes, on average, 12.81 annotations per dataset; Table 1 summarizes the results.

Table 1: Results of the Data Harmonization Process on the PLATOON Data Sources.

Type of Resource	Value
Catalog	Seven (One per Pilot)
Dataset	34 (In average 4.71 datasets per catalog)
Annotations of classes from the PLATOON Semantic Data Models	333 (in average 12.81 annotations per dataset)

In this current version, there are datasets that share several annotations (Figure 14), e.g., six datasets comprise data of the type **https://w3id.org/seas/WindDirectionEvaluation.**

Figure 14: Number of PLATOON datasets annotated with classes of the PLATOON Data Models.

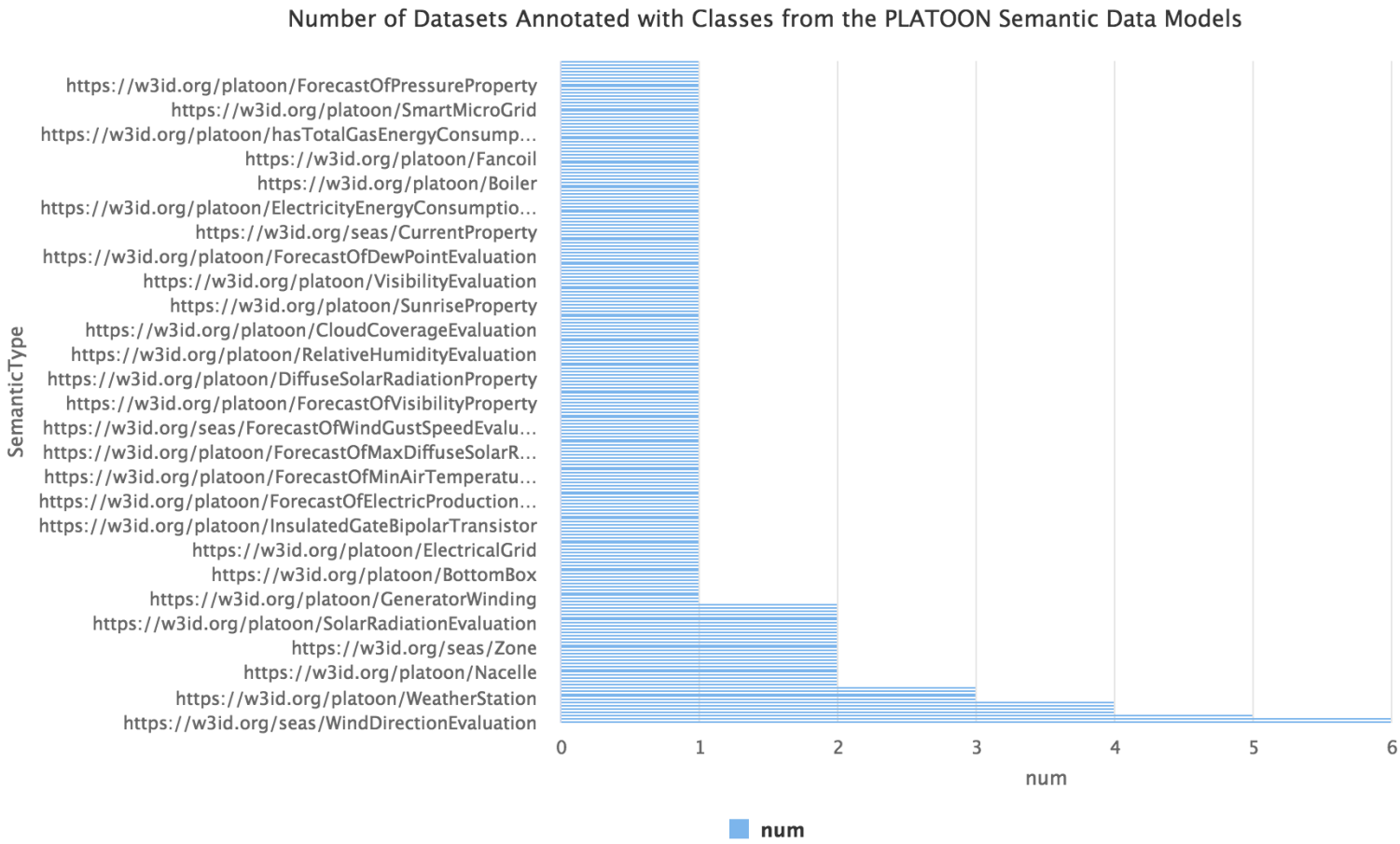

Table 2 presents the ten classes from the PLATOON semantic data models that annotate the highest number of PLATOON datasets.

Table 2: The top-10 most frequent annotations of the PLATOON datasets.

Class from the PLATOON Semantic Data Models	Annotated PLATOON Dataset
https://w3id.org/seas/WindDirectionEvaluation	PUPIN-RES-PV PUPIN-RES-PROD PUPIN-WeatherBit VUB-Pilot1a Flemish-banks-data-Pilot1a MicroGridWeatherStationPilot4a
https://w3id.org/platoon/AirTemperatureEvaluation	PUPIN-RES-PV PUPIN-RES-PROD PUPIN-WeatherBit LLUC3a-02 SCADA-Pilot3c MicroGridWeatherStationPilot4a
https://w3id.org/seas/WindSpeedProperty	PUPIN-RES-PV PUPIN-RES-PROD PUPIN-WeatherBit VUB-Pilot1a Flemish-banks-data-Pilot1a
http://www.w3.org/2006/time#Instant	PUPIN-RES-PV PUPIN-RES-PROD PUPIN-WeatherBit PUPIN-ENTSO-E ENGIE-VUB-Pilot1a
https://w3id.org/seas/WindDirectionProperty	PUPIN-RES-PV PUPIN-RES-PROD PUPIN-WeatherBit VUB-Pilot1a
https://w3id.org/seas/ElectricPowerProperty	PUPIN-RES-PV PUPIN-RES-PROD ENGIE-VUB-Pilot1a SCADA-Pilot3c
http://www.w3.org/2006/time#TemporalEntity	PUPIN-RES-PV PUPIN-RES-PROD PUPIN-WeatherBit PUPIN-ENTSO-E
https://w3id.org/platoon/AirTemperatureProperty	PUPIN-RES-PV PUPIN-RES-PROD PUPIN-WeatherBit SCADA-Pilot3c
https://w3id.org/seas/WindSpeedEvaluation	PUPIN-RES-PV PUPIN-RES-PROD PUPIN-WeatherBit LLUC3a-02 SCADA-Pilot3c MicroGridWeatherStationPilot4a
https://w3id.org/platoon/WindFarm	PUPIN-RES-PROD VUB-Pilot1a Flemish-banks-data-Pilot1a

Moreover, traditional network analysis enables to uncover the relatedness about the data sources according to the classes from the PLATOON data models that describe these data sources. Data source relatedness is measured in terms of the level of connective reached by enhancing the description of the PLATOON data sources using classes in the PLATOON data models. Cytoscape[11] is used to conduct in this assessment.

An RDF graph node represents a resource (e.g., <https://w3id.org/platoon/entity/PUPIN-RES-PROD>) or a literal (e.g., "Historical Wind Power Production Measurements"). On the other hand, an edge connecting two nodes corresponds to a property (e.g., <http://purl.org/dc/terms/title>). Graph1 and Graph2 correspond to two RDF graphs comprising the DCAT descriptions of the PLATOON datasets;
Figure 15 and Figure 16 depict Graph1 and Graph2, respectively. Both RDF graphs are composed of the same catalog and dataset properties, and they only differ in the annotations from the PLATOON semantic data models using <http://purl.org/dc/terms/type>, i.e., Graph2 does include the annotations from the PLATOON semantic data models, while Graph1 does not comprise any RDF triple representing annotations.. As expected, Graph1 is smaller than Graph2 and comprises less edges. Moreover, Graph2 expresses connections among datasets not represented in Graph1. For example, the datasets PUPIN-RES-PROD, VUB-Pilot1a, and Flemish-banks-data-Pilot1a are connected via several graphs; these connections are because they comprise data of the same types: wind farms (i.e., **https://w3id.org/platoon/WindFarm**), wind speed properties (i.e., **https://w3id.org/seas/WindSpeedProperty**), and wind direction evaluation (i.e., **https://w3id.org/seas/WindDirectionEvaluation**). These annotations provide the basis for a semantic search based on classes from the PLATOON data models, and enable a common understanding of the data that compose the PLATOON datasets.

[11] https://cytoscape.org/

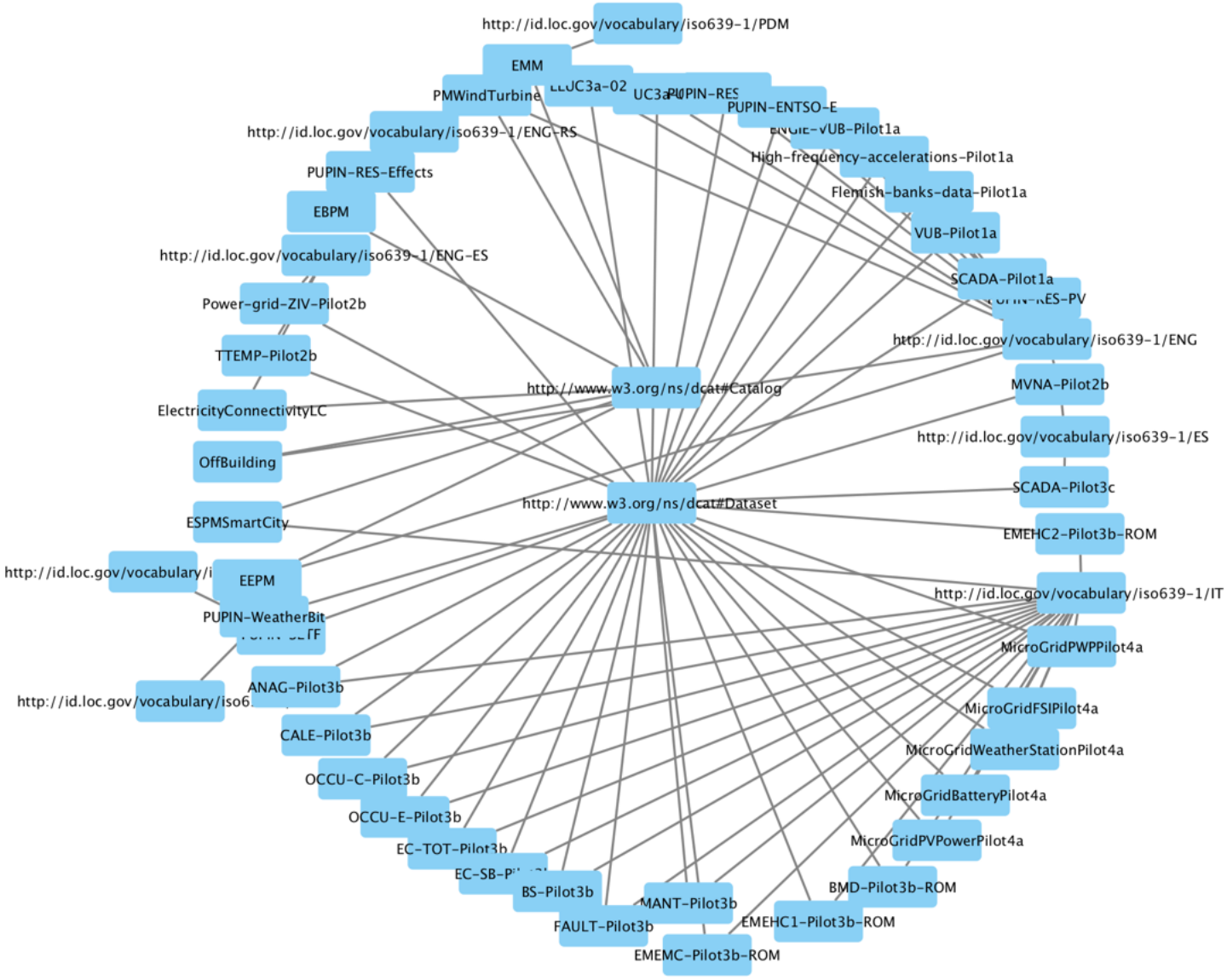

Figure 15: Graph1 where DCAT is used for data source description. Visualization powered by Cystoscape.

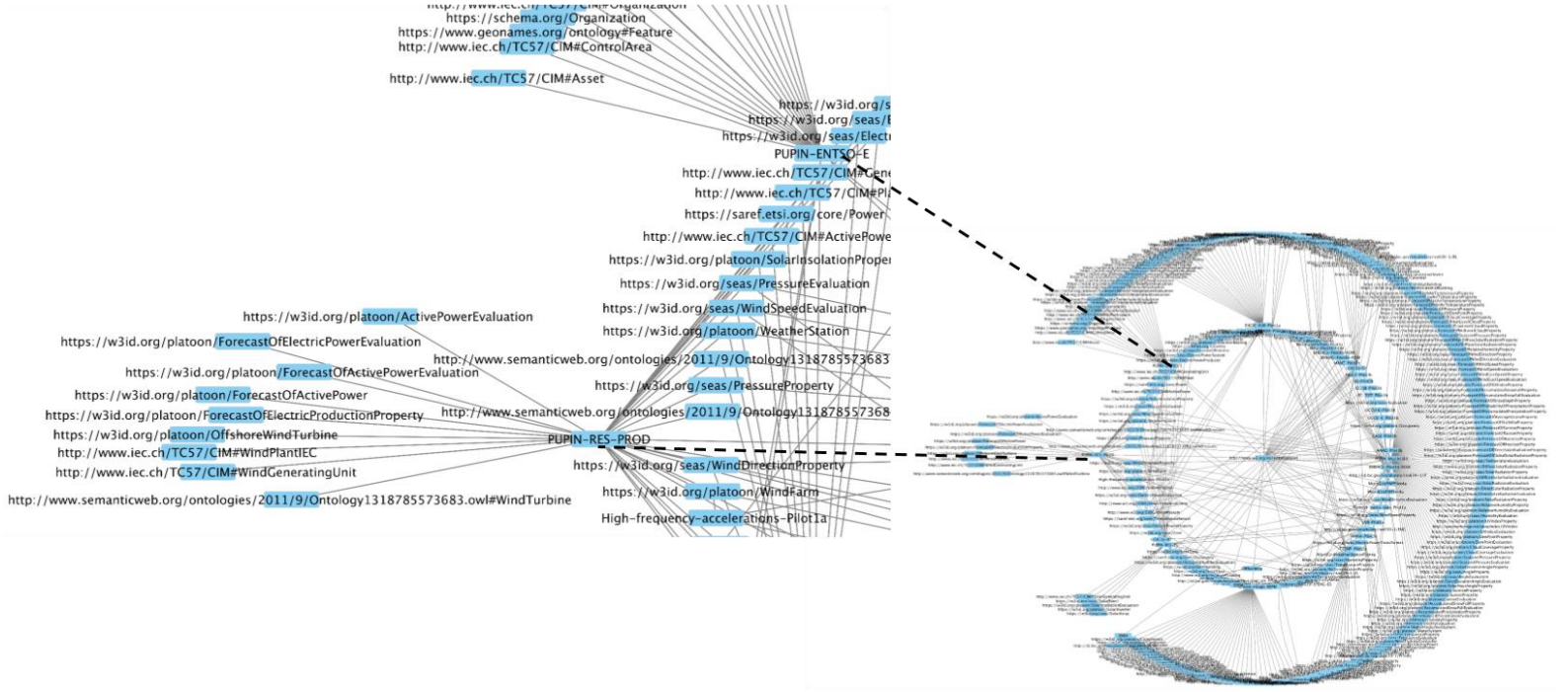

Figure 16: Graph2 where DCAT used for data source description and descriptions also include annotations of the PLATOON Semantic Data Models. Visualization powered by Cystoscape.

By using Cytoscape[12], the main properties of Graph1 and Graph2 are analyzed in terms of graph measures. Table 3 reports on the results of these measures computed by the network analysis tool of Cytoscape. Average number of neighbors indicates the average connectivity of a vertex or node in a graph. Network diameter measures the shortest path that connects the two most distant nodes in a graph. Network density measures the portion of potential edges in a graph that are actually edges; a value close to 1.0 indicates that the graph is fully connected. Lastly, the number of connected components indicates the number of subgraphs composed of vertices connected by at least one path. A number of connected components greater than 1 indicates that portions in a graph are disconnected. The results in Table 3 indicate that Graph1 size has increased considerably in Graph2; it is measured in the number of nodes and edges. Additionally, the connectivity of Graph1 has also changed. Thus, the average number of neighbors indicates that each entity in Graph2 has an average of 3.23 related entities instead of 2.74 in Graph1.

Table 3: Graph Metrics for Graph1 and Graph2 powered by Cystoscape.

Metric	Graph1	Graph2
Number Nodes	52	308
Number Edges	84	423
Avg. Number of Neighbors	2.74	3.23
Network diameter	1	1

[12] https://cytoscape.org/

Network density	0.032	0.004
Number of Connected Components	1	1

5. Analysis of the Datasets from the PLATOON Pilots

This section presents the main characteristics of the generated harmonized descriptions. Each pilot is defined in terms of the datasets, the RML mapping rules executed to create these descriptions, and an illustration of the portion of the RDF graph that comprises them.

5.1 Pilot 1a, Predictive Maintenance of Wind Farms

5.1.1 Datasets

Pilot 1a aims to predict the maintenance status of wind turbine electrical drivetrain components, such as generators and power converters. It examines onshore and offshore wind turbines powered by a doubly-fed induction generator and provides five datasets:

- **La Haute-Lys dataset- Wind turbine SCADA data (SCADA-Pilot1a):** This dataset is collected from various wind turbines situated in multiple wind farms. There is a unique data structure and tag name for every turbine brand. A wind turbine's Supervisory Control and Data Acquisition system contains sensor data at the most important subcomponents of the wind turbine; the collected data are analysed at 10-minute intervals. Turbines during a period where the electrical subcomponents had faults are also included.
- **High-frequency Data (VUB-Pilot1a):** This data is derived from a dedicated measurement campaign on onshore wind turbines and consists of a limited set of electric measurements and operational parameters (e.g., wind speed).
- **Open wind speed dataset (Flemish-banks-data-Pilot1a):** includes environmental measurements (e.g., wind speeds, wind directions) collected along the Belgian North Sea. As a basis for defining semantic labels describing wind conditions, the dataset in LLUC 1a-01 shows the typical range of wind measurements occurring in the field.
- **Offshore measurement campaign data (High-frequency-accelerations-Pilot1a):** This dataset includes acceleration measurements, that were taken of the drivetrain of an offshore wind turbine.
- **Dedicated current measurement campaign data (ENGIE-VUB-Pilot1a):** This dataset 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.

5.1.2 Mapping Rules and Annotations Using the PLATOON Semantic Data Models

Metadata describing the Pilot 1a datasets is stored in relational database. The following RML mapping rules (Figure 17) allow for the creation of the instances of class dcat:Dataset and the annotation of these entities with types from the PLATOON semantic data models.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix plt: <https://w3id.org/platoon/>.
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dctype: <http://purl.org/dc/dcmitype/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix ql: <http://semweb.mmlab.be/ns/ql#> .
@base <https://w3id.org/platoon/>.

<Pilot1a>
    rml:logicalSource [
      rml:source <RDB_source>;
      rr:sqlVersion rr:SQL2008;
      rml:query """SELECT distinct dataSource, Class FROM datasources_classespilot1a""";
    ];
    rr:subjectMap [
        rr:template "https://w3id.org/platoon/entity/{dataSource}";
        rr:class dcat:Dataset
    ];
    rr:predicateObjectMap [
        rr:predicate dct:type
        rr:objectMap [
                rr:template "{Class}"
        ]
    ].


#################################################################
```

Figure 17: Portion of the RML Mapping Rule to Define the Annotations of the Pilot1a Data Sources.

Additional relational tables store metadata about the properties of the datasets and catalogs. Figure 18 depicts the RML mapping rules that define the values of the properties dcat:keyword, dct:title, rdfs:label, dct:language, dct:creator, prov:wasGeneratedBy, dct:accessRight, and dcat:temporalResolution. The tables accessed by these mapping rules have been populated with metadata extracted from the questionnaires reported in D2.4.

```
<Source_Description_Catalogs>
    rml:logicalSource [
      rml:source <RDB_source>;
      rr:sqlVersion rr:SQL2008;
      rml:query """SELECT distinct PilotName, CatalogID, CatalogTitle, CatalogLabel, CatalogLanguage, DataSourceID FROM sourcedescriptioncatalogs""";
    ];
    rr:subjectMap [
        rr:template "https://w3id.org/platoon/entity/{CatalogID}";
        rr:class dcat:Catalog
    ];
    rr:predicateObjectMap [
        rr:predicate dcat:dataset
        rr:objectMap [
                rr:template "http://www.w3.org/ns/dcat#{DataSourceID}"
        ]
    ];
    rr:predicateObjectMap [
        rr:predicate dcat:keyword
        rr:objectMap [
            rml:reference "PilotName";
            rr:language "en"
        ]
    ];
    rr:predicateObjectMap [
        rr:predicate dct:title
        rr:objectMap [
                rml:reference "CatalogTitle"
        ]
    ];
    rr:predicateObjectMap [
        rr:predicate rdfs:label
        rr:objectMap [
                rml:reference "CatalogLabel"
        ]
    ];
    rr:predicateObjectMap [
        rr:predicate dct:language
        rr:objectMap [
                rr:template "http://id.loc.gov/vocabulary/iso639-1/{CatalogLanguage}"
        ]
    ].

###################################################################


<Source_Description_Data_Source>
    rml:logicalSource [
      rml:source <RDB_source>;
      rr:sqlVersion rr:SQL2008;
      rml:query """SELECT distinct DataSourceID, DataSourceTitle, DataSourceLanguage, DataSourceCreator, DataSourceGeneratedBy, AccessRights, TemporalResolutio
    ];
    rr:subjectMap [
        rr:template "https://w3id.org/platoon/entity/{DataSourceID}";
        rr:class dcat:Dataset
    ];
    rr:predicateObjectMap [
        rr:predicate dct:title
        rr:objectMap [
                rml:reference "DataSourceTitle"
        ]
    ];
    rr:predicateObjectMap [
        rr:predicate dct:language
        rr:objectMap [
                rr:template "http://id.loc.gov/vocabulary/iso639-1/{DataSourceLanguage}"
        ]
    ];
    rr:predicateObjectMap [
        rr:predicate dct:creator
        rr:objectMap [
                rml:reference "DataSourceCreator"
        ]
    ];
    rr:predicateObjectMap [
        rr:predicate prov:wasGeneratedBy
        rr:objectMap [
                rml:reference "DataSourceGeneratedBy"
        ]
    ];
    rr:predicateObjectMap [
        rr:predicate dct:accessRights
        rr:objectMap [
                rml:reference "AccessRights"
        ]
    ];
    rr:predicateObjectMap [
        rr:predicate dcat:temporalResolution
        rr:objectMap [
                rml:reference "TemporalResolution"
        ]
    ].
```

Figure 18: Portion of the RML Mapping Rule to Define the Catalog and the Data Source Descriptions.

5.1.3 Pilot 1a Data Harmonization in Numbers

As shown in Table 4, Pilot 1a datasets are annotated in averaged by 11.4 (median 8) classes.

Table 4: Pilot 1a datasets and their annotations

Pilot1a Dataset	Number of Annotations from the PLATOON Semantic Data Models
ENGIE-VUB-Pilot1a	35
SCADA-Pilot1a	11
VUB-Pilot1a	8
Flemish-banks-data-Pilot1a	2
High-frequency-accelerations-Pilot1a	1

Figure 19 illustrates a portion of the RDF that comprises the descriptions of the Pilot 1a datasets. The red and blue nodes represent catalog and datasets, respectively, while yellow nodes depict annotations from the PLATOON data models. ENGIE-VUB-Pilot1a and SCADA-Pilot1a are similar in terms of annotations, which is consistent with the description of these datasets. VUB-Pilot1a, Flemish-banks-data-Pilot1a, and High-frequency-accelerations-Pilot1a composed wind measurements and their annotations indicate that they are also related.

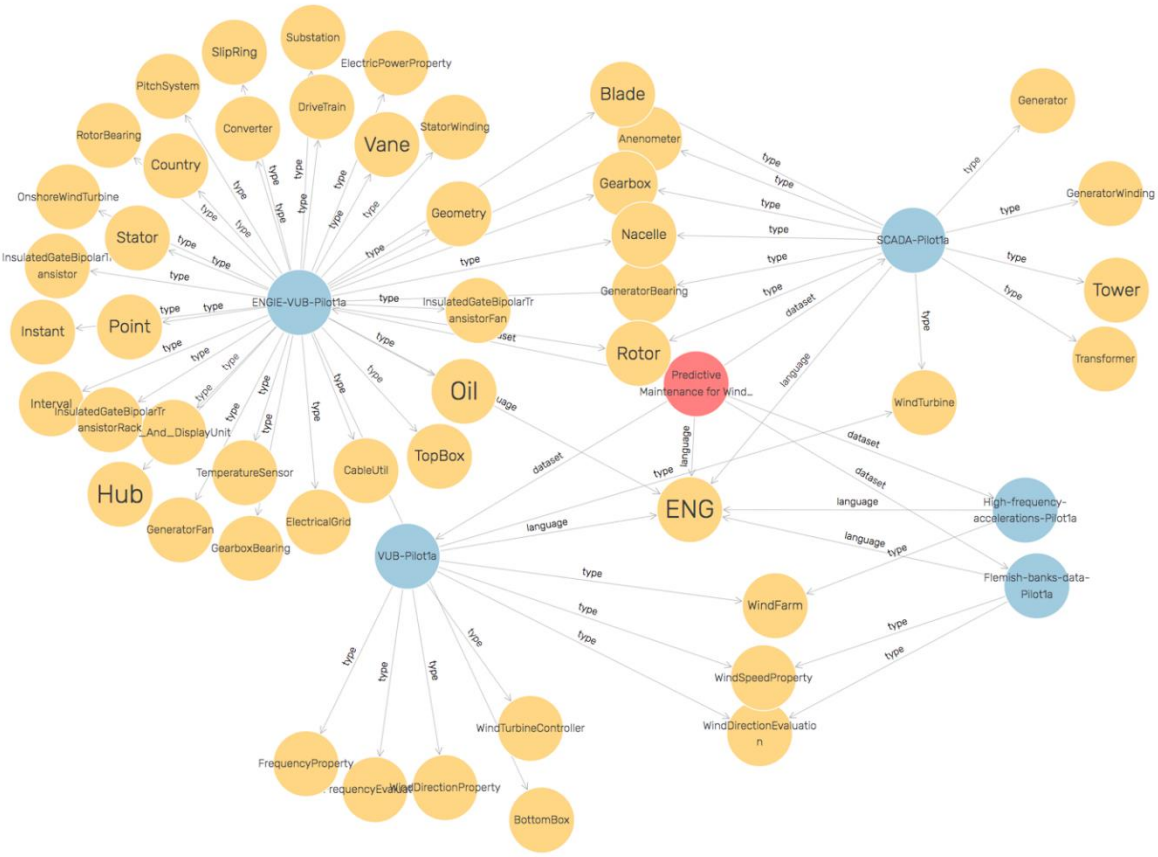

Figure 19: RDF Knowledge Base with the Description of Pilot1a Catalog and the Data Source Descriptions.

5.2 Pilot 2a, Electricity Balance and Predictive Maintenance

5.2.1 Datasets

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. 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. These data sources are described as follows:

PUPIN-RES-PROD: Historical Wind Power Production Measurements; it contains measurements of the production from the wind **power plant,** as well as **topology data**.

PUPIN-RES-PV (Predictive Maintenance): Data is collected by the Phasor Measurement Unit installed at Institute Mihajlo Pupin.

PUPIN-WeatherBit: Meteorological Data for RES Production (Generation) Forecasting Modelling Data. Meteorological dataset is utilized for RES production forecasting models training process as input data. Data is historical observational data.

PUPIN-RES-Effects: Power System calculated based on the input by Phasor Measurement Unit installed at PUPIN.

PUPIN-ENTSO-E: Transparency Platform-Energy Identification Codes (EICs); it maintains data about 39 electricity transmission system operators (TSOs) from 35 countries in Europe.

These four data sources composed the catalog EBPM (Electricity Balance and Predictive Maintenance); they provide data in English (ENG) and Serbian (RS). More details in D2.4.

5.2.2 Mapping Rules and Annotations Using the PLATOON Semantic Data Models

A relational database stores the metadata describing the Pilot 2a datasets. Figure 20 illustrates the mapping rules whose evaluation creates instances of dcat:Dataset and the annotation from the PLATOON semantic data models. Additionally, mapping rules in Figure 18 generates the entries for the rest of the DCAT predicates.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix plt: <https://w3id.org/platoon/>.
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dctype: <http://purl.org/dc/dcmitype/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix ql: <http://semweb.mmlab.be/ns/ql#> .
@base <https://w3id.org/platoon/>.

<Pilot2a>
    rml:logicalSource [
      rml:source <RDB_source>;
      rr:sqlVersion rr:SQL2008;
      rml:query """SELECT distinct dataSource, Class FROM datasources_classespilot2a""";
    ];
    rr:subjectMap [
        rr:template "https://w3id.org/platoon/entity/{dataSource}";
        rr:class dcat:Dataset
    ];
    rr:predicateObjectMap [
        rr:predicate dct:type
        rr:objectMap [
            rr:template "{Class}"
        ]
    ].
```

Figure 20: Portion of the RML Mapping Rule to Define the Annotations of the Pilot1a Data Sources.

5.2.3 Pilot 2a Data Harmonization in Numbers

As shown in Table 5, Pilot 2a datasets are annotated in averaged by 34.6 (median 26) classes.

Table 5: Pilot 2a datasets and their annotations.

Pilot2a Dataset	Number of Annotations from the PLATOON Semantic Data Models
PUPIN-WeatherBit	87
PUPIN-RES-PROD	34
PUPIN-RES-PV	26
PUPIN-ENTSO-E	22
PUPIN-RES-Effects	4

Figure 21 presents the RDF graph with the descriptions of the Pilot 2a datasets. Datasets are represented in blue nodes, while the red node models the catalog. Yellow nodes illustrate annotations from the PLATOON data models. PUPIN-RES-PROD, PUPIN-RES-PV and PUPIN-RES-Effects share a large number of annotations. PUPIN-WeatherBit is rich in terms of the different types of meteorological data that comprises, while PUPIN-ENTSO-E is described by generic types (e.g., <https://schema.org/Country>, <https://schema.org/Organization>), assets (e.g., <http://www.iec.ch/TC57/CIM#Asset>), and forecasted measures (e.g., <https://w3id.org/platoon/ForecastOfLowLevelCloudEvaluation>).

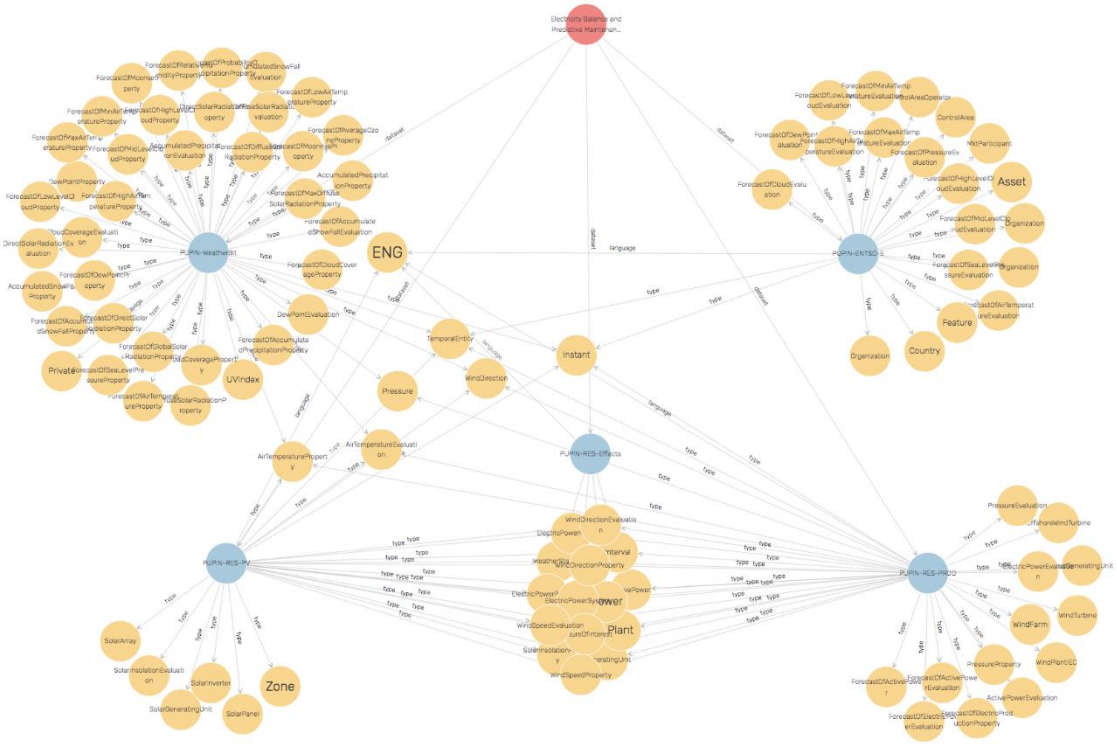

Figure 21: RDF Knowledge Base with the Description of Pilot2a Catalog and the Data Source Descriptions.

5.3 Pilot 2b, Electricity grid stability, connectivity and Life Extension

5.3.1 Datasets

Two use cases are demonstrated in ParcBit's technological park in Palma de Mallorca, Spain. ParcBit's grid consists of a 5 km long mid-voltage network and 5 km long low-voltage network. Pilot 2B uses three datasets:

- **Power grid ZIV Power Meters (Power-grid-ZIV-Pilot2b)**: dataset consists of hourly measurements of active and reactive power conveyed to users (measured by Smart Meters), gathered by concentrator and recognized by power meter.
- **Transformer Sensors data (TTEMP-Pilot2b)**: The data is collected from eight temperature sensors installed in various parts of the transformers, two sensors for ambient temperature, humidity, and pressure, and one sensor for oil temperature.
- **Medium-voltage Network Analyzer (MVNA-Pilot2b)**: contains an Electrical Network analyzer used for current transformers.

5.3.2 Mapping Rules and Annotations Using the PLATOON Semantic Data Models

Similarly, a relational database stores the metadata describing the Pilot 2b datasets. Figure 22 presents the mapping rules that create dcat:Dataset and annotations; mapping rules in Figure 18 generates the entries for the rest of the DCAT predicates.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix plt: <https://w3id.org/platoon/>.
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dctype: <http://purl.org/dc/dcmitype/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix ql: <http://semweb.mmlab.be/ns/ql#> .
@base <https://w3id.org/platoon/>.

<Pilot2b>
    rml:logicalSource [
      rml:source <RDB_source>;
      rr:sqlVersion rr:SQL2008;
      rml:query """SELECT distinct dataSource, Class FROM datasources_classespilot2b""";
    ];
    rr:subjectMap [
        rr:template "https://w3id.org/platoon/entity/{dataSource}";
        rr:class dcat:Dataset
    ];
    rr:predicateObjectMap [
        rr:predicate dct:type
        rr:objectMap [
            rr:template "{Class}"
        ]
    ].

```

Figure 22: Portion of the RML Mapping Rule to Define the Annotations of the Pilot2b Data Sources.

5.3.3 Pilot 2b Data Harmonization in Numbers

As shown in Table 6, Pilot 2b datasets are annotated in averaged by 3.6 (median 4) classes.

Table 6: Pilot 2b datasets and their annotations.

Pilot2b Dataset	Number of Annotations from the PLATOON Semantic Data Models
MVNA-Pilot2b	5
TTEMP-Pilot2b	4
Power-grid-ZIV-Pilot2b	2

Figure 23 presents the descriptions of the Pilot 2b datasets; they include electric power transformers data and share the annotation seas:ElectricPowerTransformer.

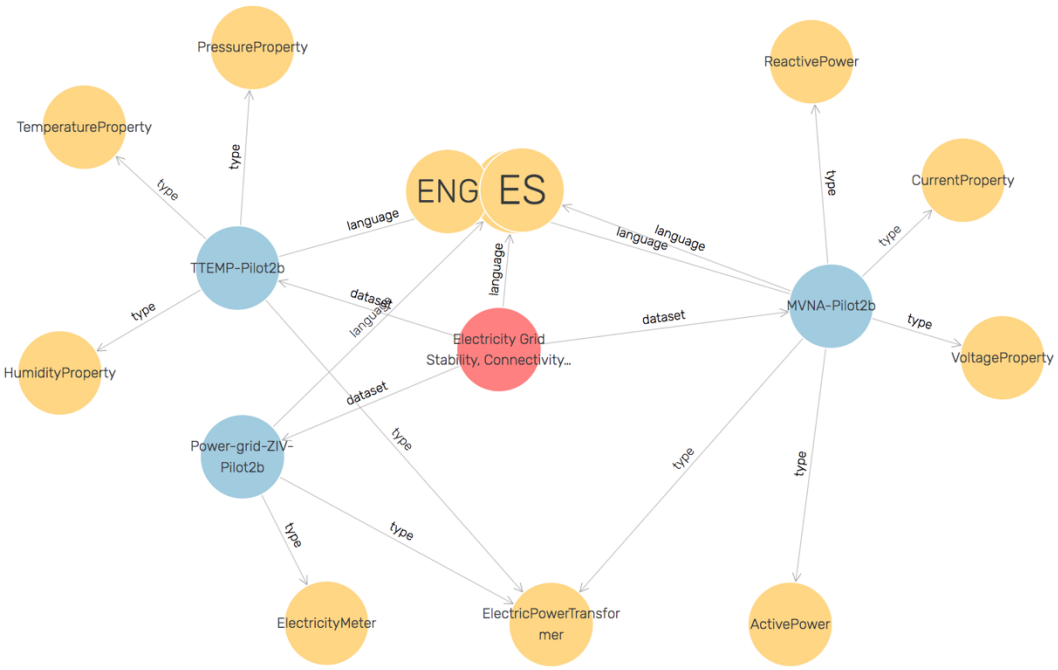

Figure 23: RDF Knowledge Base with the Description of Pilot3b Catalog and the Data Source Descriptions.

5.4 Pilot 3a, Office building: Operation performance thanks to physical models and IA algorithms

5.4.1 Datasets

Pilot 3a is about an office building equipped with a building management system (BMS) that controls HVAC and comfort in multiple zones of the building. This pilot includes LLUC 3a-01 - Optimizing HVAC control regarding occupancy, and LLUC 3a-02 - Providing Demand Response Service through HVAC control.

- **LLUC 3a-01:** This dataset offers a smart module that will optimize HVAC operation based on real-time occupancy data for an office building. Dedicated sensors provide occupancy data, and the building's BMS provides the comfort and HVAC controls. Learning algorithms can be applied to predict occupancy and estimate the heating and cooling periods of the building and its various zones.

- **LLUC 3a-02:** This dataset delivers a smart module to ensure that Demand Response services at an office building are implemented by using HVAC control and building inertia. Using the building parameters and weather forecast data, the module estimates the HVAC load and the potential flexibility of the building.

5.4.2 Mapping Rules and Annotations Using the PLATOON Semantic Data Models

Metadata describing the Pilot 3a datasets is stored in a relational database. Figure 24 presents the mapping rules that create dcat:Dataset and annotations; mapping rules in Figure 18 generates the entries for the rest of the DCAT predicates.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix plt: <https://w3id.org/platoon/>.
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dctype: <http://purl.org/dc/dcmitype/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix ql: <http://semweb.mmlab.be/ns/ql#> .
@base <https://w3id.org/platoon/>.

<Pilot3a>
    rml:logicalSource [
      rml:source <RDB_source>;
      rr:sqlVersion rr:SQL2008;
      rml:query """SELECT distinct dataSource, Class FROM datasources_classespilot3a""";
    ];
    rr:subjectMap [
        rr:template "https://w3id.org/platoon/entity/{dataSource}";
        rr:class dcat:Dataset
    ];
    rr:predicateObjectMap [
        rr:predicate dct:type
        rr:objectMap [
            rr:template "{Class}"
        ]
    ].
```

Figure 24: Portion of the RML Mapping Rule to Define the Annotations of the Pilot3a Data Sources.

5.4.3 Pilot 3a Data Harmonization in Numbers

Pilot 3a datasets are annotated in averaged by 11.5 (median 11.5) classes (in Table 7).

Table 7: Pilot 3a datasets and their annotations

Pilot3a Dataset	Number of Annotations from the PLATOON Semantic Data Models
LLUC3a-02	19
LLUC3a-01	4

Figure 25 presents the pilot 3a datasets; they comprise office building data (e.g., <https://saref.etsi.org/core/Occupancy>, <https://w3id.org/bot#Zone>, and <https://w3id.org/bot#Building>). Additionally, LLUC3a-02 is annotated with classes denoting building- and weather-related measurements.

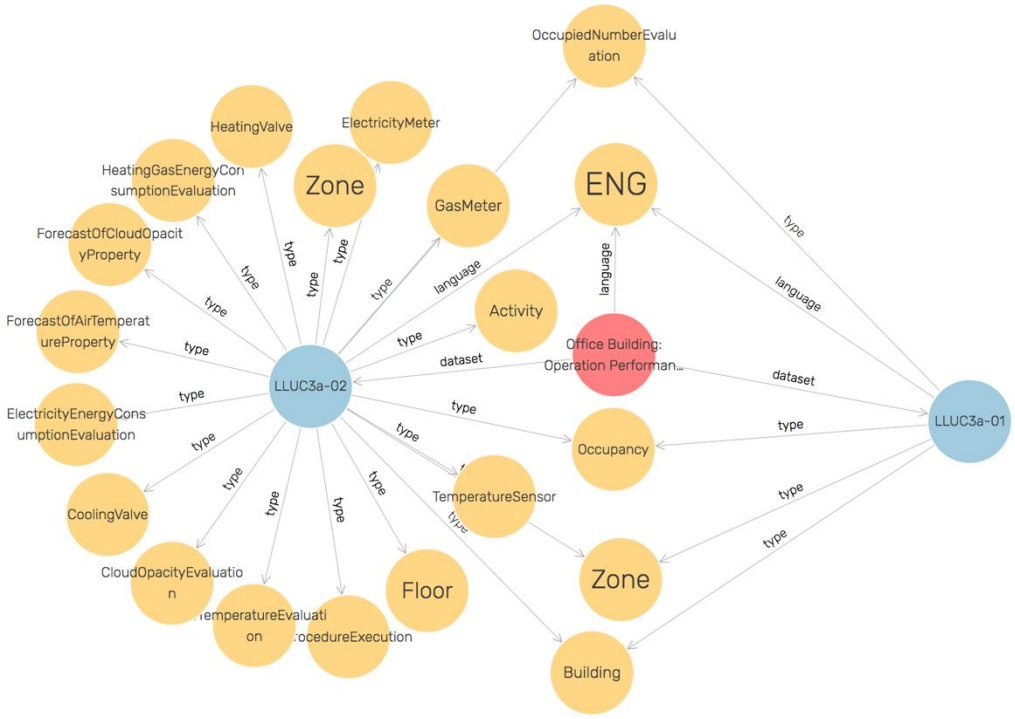

Figure 25: RDF Knowledge Base with the Description of Pilot3a Catalog and the Data Source Descriptions.

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

5.5.1 Datasets

Pilot 3b aims at acquiring, aggregating, and processing data of energy consumption and related properties of various buildings for making energy domain-specific analyses, e.g., consumption forecasting, predictive maintenance, benchmarking. Pilot 3b is formed of 2 subpilots: 3b-PI and 3b-ROM. Each of the subpilots have different datasets as explained below:

Pilot #3b_PI

There are four possible destinations for the building spaces in Poste Italiane buildings located in the Rome Municipality Area: Datacenter, Logistics distribution, and cross-docking (mail & parcels), Retail and Office (Directional), with a total of 16 buildings.

- **Building Data (ANAG-Pilot3b):** This dataset includes information about each building's features and general characteristics (ID office, address, destination use, square feet, climate zone, etc.).
- **Building Occupancy dataset (OCCU_C & OCCU_E):** In each building of the pilot, the number of employees and customers is recorded daily.
- **Calendar (CALE-Pilot3b):** keeps information regarding office openings and shifts.
- **Consumption on Building (EC_TOT & EC_SB):** An overview on the temperature and humidity inside a building or line as well as information about active energy consumption (kWh). Among its many uses will be consumption prediction and

benchmarking, identifying anomalies, and assessing lighting consumption. The information from the climate sensors serves various purposes, e.g., predicting consumption to maintain a certain comfort level, and appropriate consumption.

- **Building Energy Systems (BS-Pilot3b):** This dataset provides a description of the heating, cooling, and lighting systems in all buildings. It will be possible to use building HVAC plant information for a variety of purposes, including consumption prediction and consumption benchmarking. In addition, building lighting plants information will be utilized to calculate lighting consumption, benchmarks and estimate lighting energy consumption.
- **Systems Anomalies (Fault-Pilot3b):** The data is derived from monitoring the temperature within the building. In addition, alerts are generated when temperatures exceed a given threshold.

Pilot #3b_ROM

These are the available datasets:

- **Energy Meters Electrical Monthly Consumptions (EMEMC-Pilot3b-ROM):** All power consumption from last month's meters (energy vendor).
- **Energy Meters Electrical Historical Consumptions (EMEHC2):** A historical daily record (kwh) of the electric consumption for ROM buildings; divided in rows for each 15-minute consumption period.
- **Building master data (BMD-Pilot3b-ROM):** A database from ROM Asset Management Office and buildings Energy Audits.
- **Energy Meter Gas (EMGMC, EMGHC, EMGTC and EMGHC2):** These datasets include information about the monthly and historical gas consumption for RC direct meters, also about thermal and historical gas consumption data for SIE3 meters. These Gas consumptions can be used to perform benchmarking and to train forecast models.
- **ROM PV production data (RPVPD-Pilot3b-ROM):** This dataset comes from the Lovato Electric system. It contains the number of kWh produced by the installed PV plants within a set of ROM buildings divided by district.

5.5.2 Mapping Rules and Annotations Using the PLATOON Semantic Data Models

A relational database stores metadata about the Pilot 3b. The following RML mapping rules (Figure 26) enable for the generation of the class dcat:Dataset instances, as well as the annotation of these entities with types from the PLATOON semantic data models.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix plt: <https://w3id.org/platoon/>.
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dctype: <http://purl.org/dc/dcmitype/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix ql: <http://semweb.mmlab.be/ns/ql#> .
@base <https://w3id.org/platoon/>.

<Pilot3b>
    rml:logicalSource [
      rml:source <RDB_source>;
      rr:sqlVersion rr:SQL2008;
      rml:query """SELECT distinct dataSource, Class FROM datasources_classespilot3b""";
    ];
    rr:subjectMap [
        rr:template "https://w3id.org/platoon/entity/{dataSource}";
        rr:class dcat:Dataset
    ];
    rr:predicateObjectMap [
        rr:predicate dct:type
        rr:objectMap [
            rr:template "{Class}"
        ]
    ].
```

Figure 26: Portion of the RML Mapping Rule to Define the Annotations of the Pilot3b Data Sources.

5.5.3 Pilot 3b Data Harmonization in Numbers

As shown in Table 8, Pilot 3b datasets are annotated in averaged by 3.29 (median 2.5) classes.

Table 8: Pilot 3b datasets and their annotations.

Pilot3b Dataset	Number of Annotations from the PLATOON Semantic Data Models
BS-Pilot3b	2
ANAG-Pilot3b	2
CALE-Pilot3b	1
OCCU-C-Pilot3b	3
OCCU-E-Pilot3b	3
EC-TOT-Pilot3b	1
EC-SB-Pilot3b	1
FAULT-Pilot3b	1
BS-Pilot3b	2
EMEMC-Pilot3b-ROM	8
BMD-Pilot3b-ROM	5
EMEHC2-Pilot3b-ROM	9
RPVPD-Pilot3b-ROM	3
EMGMC-EMGHC-EMGTC-EMGHC2-Pilot3b-ROM	5

Figure 27 presents the descriptions of the Pilot 3b datasets; they include advanced energy management system data describing mainly buildings and structures. All the datasets are described in the Italian language (IT) and are related to one catalog. The blue nodes represent the pilot datasets. The annotations of the dataset are represented in yellow nodes. The annotations of the datasets differ according to the corresponding dataset describing data

related to structures (Building, Site, Place), measurements for power consumptions and their related prices, and time related measurement (Calendar, Interval).

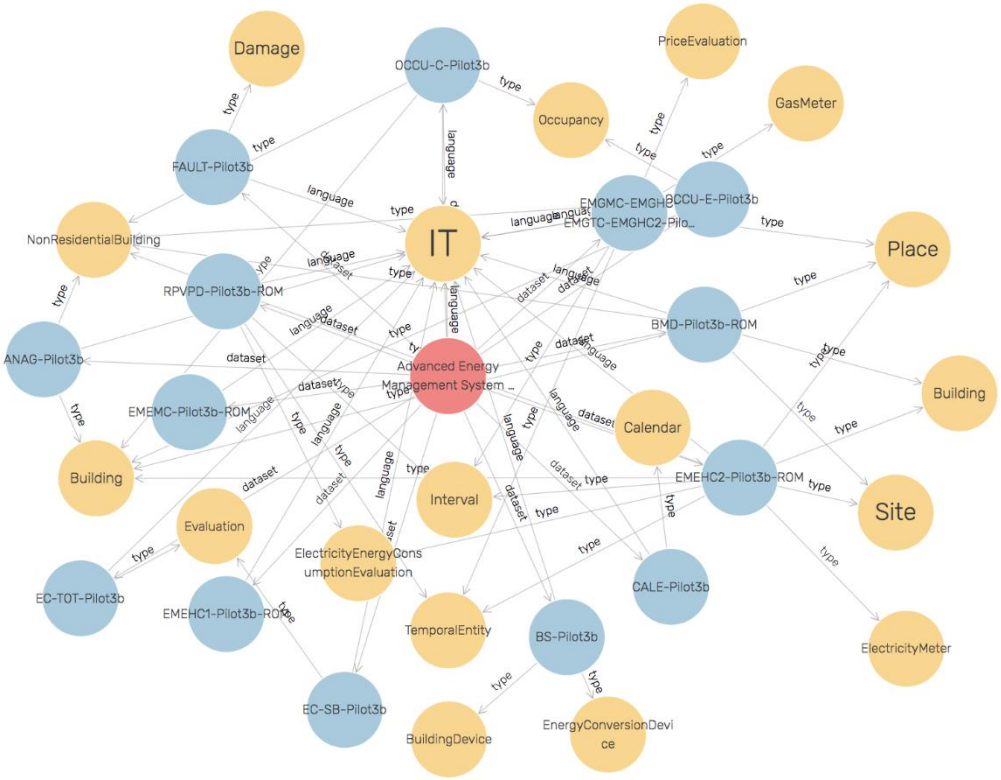

Figure 27: RDF Knowledge Base with the Description of Pilot3b Catalog and the Data Source Descriptions.

5.6 Pilot 3c, Advanced Energy Management System and Energy Efficiency and Predictive Maintenance in the Smart Tertiary Building Hubgrade

5.6.1 Data Sources

- **Energy Huybgrade dataset (SCADA-Pilot3c)**: SCADA data about buildings (up to 300) with observations collected from thermal, electric, and gas meters. Collected values correspond to temperatures, water, electricity and thermal consumption, position of the valves, dumpers. Weather data and forecasts are also part of this dataset. The observations are registered every 10 minutes and the database grows 1.5MB per day per building.

5.6.2 Mapping Rules and Annotations Using the PLATOON Semantic Data Models

A relational database stores metadata characterizing the Pilot 3c datasets. Figure 28 depicts the mapping rules that are used to generate dcat:Dataset and annotations. The mapping rules in Figure 28 produce the entries for the remaining DCAT predicates.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix plt: <https://w3id.org/platoon/>.
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dctype: <http://purl.org/dc/dcmitype/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix ql: <http://semweb.mmlab.be/ns/ql#> .
@base <https://w3id.org/platoon/>.

<Pilot3c>
    rml:logicalSource [
      rml:source <RDB_source>;
      rr:sqlVersion rr:SQL2008;
      rml:query """SELECT distinct dataSource, Class FROM datasources_classespilot3c""";
    ];
    rr:subjectMap [
        rr:template "https://w3id.org/platoon/entity/{dataSource}";
        rr:class dcat:Dataset
    ];
    rr:predicateObjectMap [
        rr:predicate dct:type
        rr:objectMap [
            rr:template "{Class}"
        ]
    ].
```

Figure 28: Portion of the RML Mapping Rule to Define the Annotations of the Pilot3c Data Sources.

5.6.3 Pilot 3c Data Harmonization in Numbers

As we can see in Table 9, Pilot3c has a single dataset annotated with 39 annotations from the PLATOON semantic data models.

Table 9: Pilot 3c datasets and their annotations

Pilot3c Dataset	Number of Annotations from the PLATOON Semantic Data Models
SCADA-Pilot3c	39

Figure 29 presents the RDF graph describing pilot 3c datasets. The blue node represents the pilot 3c dataset. All the other yellow nodes represent the annotations and types assigned to the SCADA-Pilot3c dataset. The dataset is described in the Spanish language (ES) and is related to the catalog Energy Efficiency and Predictive maintenance.

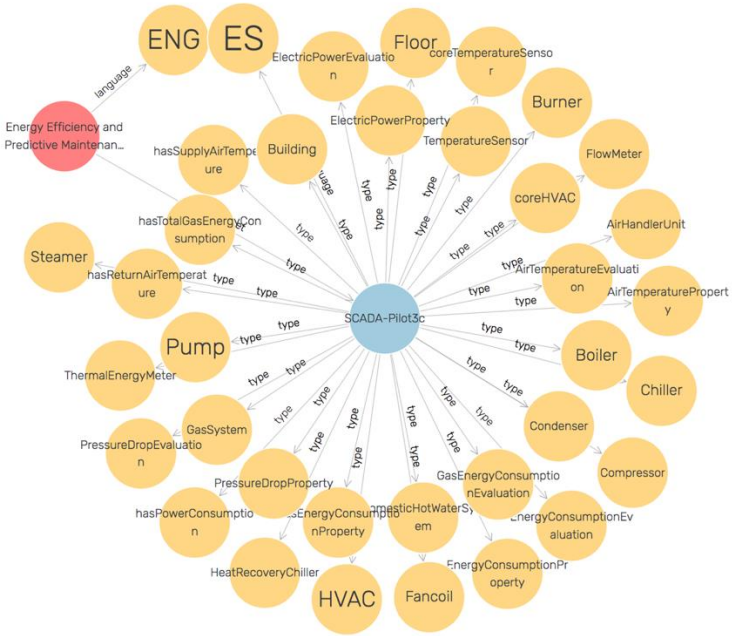

Figure 30: RDF Knowledge Base with the Description of Pilot3c Catalog and the Data Source Descriptions.

5.7 Pilot 4a, Energy Management in microgrids

5.7.1 Datasets

Pilot 4a consists of four datasets from the area of Milan, Italy:

● **Microgrid PV power production and forecast (MicroGridPVPilot4a)**: 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 (MicroGridBatteryPilot4a)**: 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 (MPWPPilot4a)**: 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 (MicroGridWeatherStationPilot4a):** consist of observations sensed by a weather station. It reports ambient temperature, wind speed, wind direction, relative humidity, rain, and irradiance. The growth trend is 65K records per day, and observations are registered every 10 seconds.

● **Microgrid full skype imaging (MicroGridFSIPilot4a):** comprises full-sky images in JPEG format. It grows in more than 250 records per day every 5 minutes.

5.7.2 Mapping Rules and Annotations Using the PLATOON Semantic Data Models

The metadata describing the Pilot 4a datasets, similarly to the other pilots, is maintained in a relational database. Figure 30 depicts the mapping rules used to generate dcat:Dataset and annotations; the shown mapping rules generate entries for the remaining DCAT predicates.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix plt: <https://w3id.org/platoon/>.
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dctype: <http://purl.org/dc/dcmitype/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix ql: <http://semweb.mmlab.be/ns/ql#> .
@base <https://w3id.org/platoon/>.

<Pilot4a>
    rml:logicalSource [
      rml:source <RDB_source>;
      rr:sqlVersion rr:SQL2008;
      rml:query """SELECT distinct dataSource, Class FROM datasources_classespilot4a""";
    ];
    rr:subjectMap [
        rr:template "https://w3id.org/platoon/entity/{dataSource}";
        rr:class dcat:Dataset
    ];
    rr:predicateObjectMap [
        rr:predicate dct:type
        rr:objectMap [
            rr:template "{Class}"
        ]
    ].

```

Figure 31: Snapshot of the RML Mapping Rule defining the annotations of the Pilot 4a Data Sources.

5.7.3 Pilot 4a Data Harmonization in Numbers

Table 10 depicts the annotations of the Pilot 4a datasets, an average of 6.5 (median 3) classes is used to annotate the datasets.

Table 10: Pilot 4a datasets and their annotations

Pilot4a Dataset	Number of Annotations from the PLATOON Semantic Data Models
MicroGridWeatherStationPilot4a	18
MicroGridPVPilot4a	4
MicroGridBatteryPilot4a	2
MicroGridFSIPilot4a	2

Figure 32 presents the RDF graph with the descriptions of the pilot4a datasets. The blue nodes represent the dataset of the pilot. While the yellow nodes are the annotations of the mentioned datasets. All the datasets in pilot4a are described in the Italian language (IT). The datasets are related to the Energy Management in Microgrids catalog. As we can see in Figure 32, the datasets are annotated with different annotations, which is consistent with the nature of each dataset from the pilot. The dataset MicroGridWeatherStation has most of the annotations in this pilot (18) compared to the other datasets (2-4).

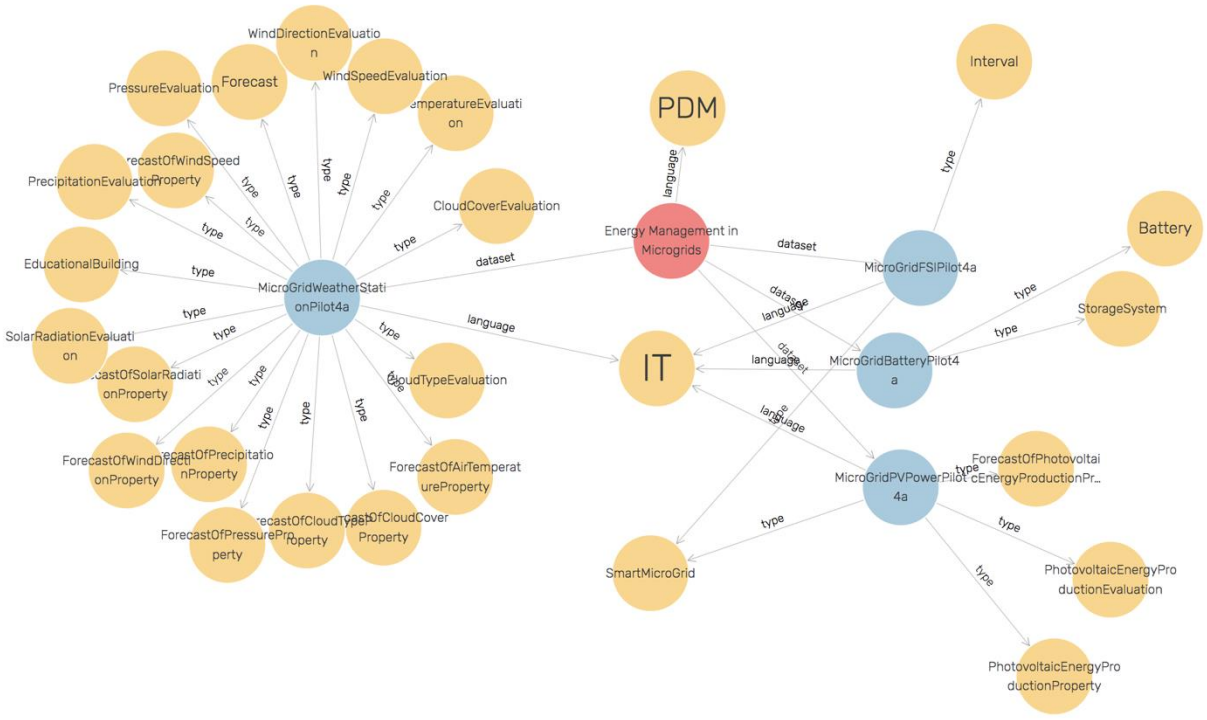

Figure 32: RDF Knowledge Base with the Description of Pilot4a Catalog and the Data Source Descriptions.

6. Conclusions and Future Work

This document reports on the outcomes of performing task T5.3 – Data Collection and Harmonization, that started in month nineteen (M19) of the PLATOON project. The characterization of the PLATOON data sources conducted in T2.4 represented the starting point for the harmonization process. Although datasets are presented in diverse formats (e.g., CSV, JSON, RDB, JPEG), they comprise data collected characterized by concepts in the energy domain represented in the PLATOON semantic data models.

The reported results provide evidence of the relevance role of the PLATOON semantic data models defined in T2.3 and how they allow for the development a common understanding of the meaning of concepts that characterize the energy sector. Furthermore, T5.3 outcomes provide illustrate the potential that data harmonization using W3C recommended vocabularies and energy data models has into interoperability resolution.

Expressive vocabularies from the International Data Space information model (e.g., DCAT) facilitate the generation of machine-readable description of the PLATOON data sources. During the next months of 2022, the outcomes of this task will be utilized to populate the data market catalog implemented in WP3 and the pipeline to describe the PLATOON analytical tools developed in WP4.

7. References

[1] S. Geisler, M.-E. Vidal, C. Capiello, B. Farias, A. Gal, M. Jarke, M. Lenzerini, P. Missier, B. Otto, E. Paja, B. Pernici and J. Rehof, “Knowledge-driven Data Ecosystems Towards Data Transparency,” *ACM J. Data Inf. Quality*, vol. 14, no. 1, pp. 1-12, 2022.

[2] M. Lenzerini, “Data Integration: A theorical Perspective,” in *ACM SIGMO-SIGACT-SIGART*, 2002.

[3] F. Maali, J. Erickson and P. Archer, “Data Catalog Vocabulary (DCAT),” W3C, 2017.

[4] S. Bader, J. Pullmann, C. Mader, S. Tramp, A. Quix, W. Muller, H. Akyurek, M. Bockmann, B. Imbusch, J. Lipp, S. Geisler and C. Lange, “The International Data Spaces Information Model- An Ontology for Sovereign Exchaned of Digital Content,” in *International Semantic Web Conference*, 2020.

[5] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborg and E. Mannens, "RML: A Genric Language for Integrated RDM Mappings of Heterogeneous Data," in *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International WWW*, 2014.

[6] D2.3, “PLATOON D2.3: PLATOON Common Data Models for Energy,” 2020.

[7] E. Iglesias, S. Jozashoori, D. Chaves-Fraga, D. Collarana and M.-E. Vidal, “SDM-RDFizer; An RML Interpreter for the Efiicient Creation of RDF Knowledge Graphs,” in *Procedding of the ACM International Conference on Information & Knowledge Management*, 2020.

[8] K. Alexander, R. Cyganiak, M. Hausenblasm and J. Zhao, “Describing Linked Datasets with the VoID Vocabulary.,” W3C, 2011.

[9] R. Cyganiak and D. Reynolds, “The RDF Data Cube Vocabulary,” W3C, 2014.

[10] H. Knublauch and D. Kontokostas, “Shapes Constraint Language (SHACL),” W3C, 2017.

[11] A. Miles and S. Bechhofer, “SKOS Simple Knowledge Organization System Reference,” W3C, 2009.

[12] D. Reynolds, “The Organization Ontology,” W3C, 2014.

[13] T. Lebo, S. Sahoo and D. McGuinness, “PROV-O: The PROV Ontology,” 2013.

[14] R. Iannella, “Open Digital Rights Language (ODRL) Version 1.1,” W3C, 2002.

[15] R. Albertoni and A. Isaac, “Data on the Web Best Practices: Data Quality Vocabulary,” W3C, 2016.

[16] D. Brickley, “Basic Geo (WGS84 lat/long) Vocabulary,” W3C, 2006.

[17] S. Cox and C. Little, “Time Ontology in OWL,” W3C, 2017.