

Article

Responsible Knowledge Management in Energy Data Ecosystems

Valentina Janev ^{1,*}, Maria-Esther Vidal ^{2,3}, Dea Pujic ¹, Dušan Popadić ¹, Enrique Iglesias ³, Ahmad Sakor ^{2,3} and Andrej Čampa ⁴

¹ Mihajlo Pupin Institute, University of Belgrade, 11060 Belgrade, Serbia; dea.pujic@pupin.rs (D.P.); dusan.popadic@pupin.rs (D.P.)

² TIB-Leibniz Information Centre for Science and Technology, 30167 Hannover, Germany; maria.vidal@tib.eu (M.-E.V.); ahmad.sakor@tib.eu (A.S.)

³ L3S Research Center, Leibniz University of Hannover, 30167 Hannover, Germany; enrique.iglesias@tib.eu

⁴ ComSensus, 1233 Dob, Slovenia; andrej.campa@comsensus.eu

* Correspondence: valentina.janev@instituteupin.com

Abstract: This paper analyzes the challenges and requirements of establishing energy data ecosystems (EDEs) as data-driven infrastructures that overcome the limitations of currently fragmented energy applications. It proposes a new data- and knowledge-driven approach for management and processing. This approach aims to extend the analytics services portfolio of various energy stakeholders and achieve two-way flows of electricity and information for optimized generation, distribution, and electricity consumption. The approach is based on semantic technologies to create knowledge-based systems that will aid machines in integrating and processing resources contextually and intelligently. Thus, a paradigm shift in the energy data value chain is proposed towards transparency and the responsible management of data and knowledge exchanged by the various stakeholders of an energy data space. The approach can contribute to innovative energy management and the adoption of new business models in future energy data spaces.

Keywords: data integration systems; energy big data; knowledge graphs; data exchange; semantic interoperability; big data analytic



Citation: Janev, V.; Vidal, M.-E.; Pujic, D.; Popadić, D.; Iglesias, E.; Sakor, A.; Čampa, A. Responsible Knowledge Management in Energy Data Ecosystems. *Energies* **2022**, *15*, 3973. <https://doi.org/10.3390/en15113973>

Academic Editors: Marek Matejun, Bożena Ewa Matusiak and Izabela Różańska-Bińczyk

Received: 18 April 2022

Accepted: 24 May 2022

Published: 27 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The digital transformation of the electricity sector from traditional electric grids to smart grids [1] is driven by multiple factors [2], while the emerging information technologies (e.g., big data, semantic technologies, machine learning algorithms) play a relevant role of automation and control of the energy value chain. Despite being recognized as crucial applications for efficiently generating and consuming energy, big data applications in the energy domain are still underdeveloped and fragmented. Challenges related to controlled data exchange and data integration are still not fully achieved. Hence, the fragmented applications are developed against energy data silos, and data exchange is limited. After the announcement of the Google Knowledge Graph [3] in 2012, semantic technologies and knowledge graphs (KGs) gained in popularity. They have been applied in various domains, especially to enhance the integration of distributed resources over the Internet, e.g., for facilitating product/service discovery [4], managing business registers and company data [5], managing drug data [6], or emergency management [7]. In this paper (it is an extension of a conference paper on knowledge-driven frameworks for managing energy data spaces, see Valentina Janev, Maria Esther Vidal, Kemele Endris, and Dea Pujic. 2021. Managing Knowledge in Energy Data Spaces. In Companion Proceedings of the Web Conference 2021 (WWW'21 Companion), 19–23 April 2021, Ljubljana, Slovenia. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3442442.3453541>, accessed on 23 May 2022), the authors

assess the applicability of the technologies for managing knowledge [8] in the energy sector and consequently for specifying new business models.

1.1. Data Ecosystems

Data ecosystems (DEs) are data-driven infrastructures that enable stakeholders to exchange and integrate data [9,10]. DEs comprise various computational methods to overcome interoperability issues while preserving data privacy, security, and sovereignty. They can be aligned to international data strategies, e.g., the European Data Strategy [11], representing, thus, crucial technological building blocks for digitalization and data markets, as well as for enhancing competitiveness and digital sovereignty. A DE can be centralized, and maintain shared data sources and host services on top of these sources. In this case, several DEs can be interconnected into a DE network [9,12]. As an individual DE, each node maintains and exchanges data; it can also perform data management and analytical tasks. DEs resort to semantic data models for providing a uniform view of heterogeneous data sources. Moreover, mapping rules state how data sources are defined in terms of semantic data models exported as unified schemas. Lastly, a DE can also be enhanced with a meta-layer that describes business models, data access regulations, and data exchange contracts.

1.2. The EU Energy Data Ecosystem

A priority on the European Union (EU) Political Agenda for the next period (2019–2024) is the European Green Deal strategy (2019) that aims to position Europe as the first climate-neutral continent by the year 2050. Integrated energy systems play a crucial role in implementing this vision. Hence, a new document—the EU Strategy for Energy System Integration (COM(2020) 299 final, 8 July 2020), was adopted that envisions coordinated planning and operation of the energy system as a whole, across multiple energy carriers, infrastructures, and consumer sectors. The increased volume of data generated from distributed renewable data sources creates data integration and processing challenges on different levels (processing in the cloud, processing on edge). Therefore, there is a need to develop computational methods for ingesting, managing, and analyzing big data. More importantly, considering the bidirectional flow of information and energy in smart grids, knowledge needs to be extracted from this data to uncover actionable insights. Hence, the future energy infrastructure will be based on intelligent power electronics, smart meters, context-aware devices, IoT, and AI-driven services. Interoperability problems caused by currently fragmented applications will be overcome in the new generation of grids, thus enabling data exchange between different players in the energy sector. For instance, the EU Data Strategy envisages the establishment of energy data spaces based on semantic web technologies and W3C standards. The information model (<https://github.com/International-Data-Spaces-Association/InformationModel>, (accessed on 23 May 2022)) proposed in the context of the International Data Space includes exemplary data models for describing datasets and services metadata needed to facilitate information search, service matching, and data exchange.

1.3. Overview of Main Contributions

The work presented in this paper is built on our previous work (conference paper) on knowledge-driven frameworks for managing energy data spaces [8] and in the knowledge-driven data ecosystems [12]. With the focus of achieving the targets envisioned in the latest EU energy strategy and the European Green Deal Action Plan [13], we present a knowledge-driven data ecosystem to encapsulate, communicate, and manage the distributed assets in the energy value chain. The main contributions of this paper are the following:

- A new approach can combine data and knowledge management and enhance the analytics services portfolio of various energy stakeholders. Thus, energy expert users can develop analytical methods for two-way electricity flows and information and optimize electricity generation, distribution, and consumption on top of heterogeneous data sources.

- An abstract architecture for semantic data integration and business analytics. On top of this architecture, various knowledge-driven services for processing data contextually and intelligently are devised.
- A unified knowledge graph that converges data and knowledge collected from the data ecosystem. The knowledge graph is connected to existing encyclopedic knowledge graphs (e.g., DBpedia [14] and Wikidata [15]). Additionally, a federated query engine allows for query processing on top of the connected knowledge graphs in a unified way. This engine provides the basis for the development of interactive and explainable AI-based services on top of the knowledge-driven data ecosystem.
- An analytical layer composed of advanced analytical services (statistical and ML models that work on edge and on top of integrated data). Depending on the stakeholders' needs and the available data, the services offered are related to renewable energy source (RES) production forecast, RES effects calculation, buildings operation optimization, and asset predictive maintenance.

The description of the detailed design and implementation of advanced analytical services (statistical and ML models that work on top of integrated data) is out of the scope of this paper. The large-scale validation is still underway.

The paper is organized as follows. Section 2 presents motivation scenarios from the energy sector. Our approach for big data management and analytics in the energy domain is introduced in Section 3. Sections 4 and 7 present proof-of-concept and discuss the results. Finally, related work is summarized in Sections 8, and 9 wraps our lesson learned up.

2. The Electricity Value Chain: Overview of Challenges

2.1. Example Case Study

The recently adopted EU energy-related strategies create opportunities to modernize the energy system, making it competitive and environmentally sustainable. Herein, we will use the example of the electricity system from Serbia (see Figure 1). The SCADA system of the Institute Mihajlo Pupin has been deployed at many parts of the national electricity grid. The system monitors and controls energy production, distribution, and usage with different objectives, including improving energy efficiency, increasing flexibility and renewable generation share, and reducing energy costs. Hence, in this paper, the authors describe a case study for an innovative energy management service layer on top of existing SCADA based on reusable semantic models or knowledge graphs. The proposed approach facilitates the integration of data silos and their fine-grain semantic description. Further, the semantic description using knowledge graphs provides a common understanding of the energy domain based on existing domain-based vocabularies. Additionally, this approach provides a ground for new business models and facilitates integration in the EU energy data space.

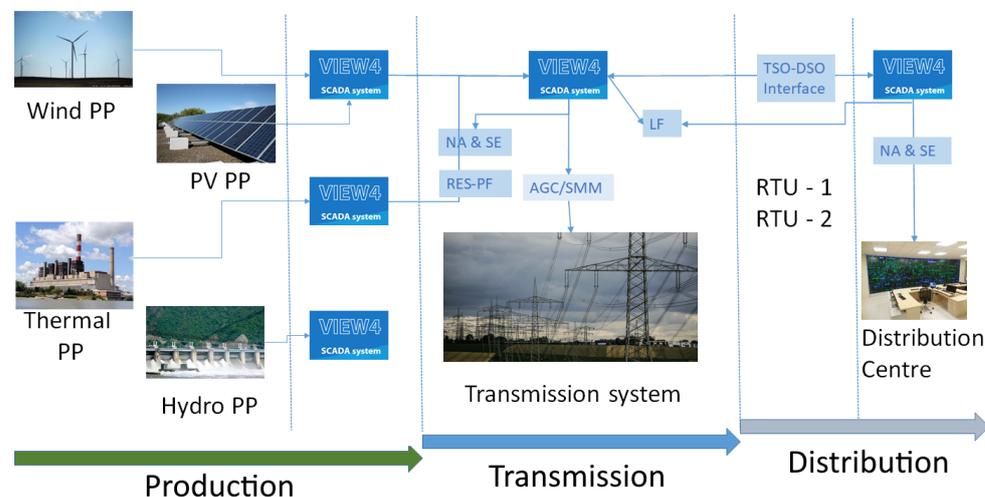


Figure 1. The electricity value chain.

Because the national electricity infrastructure is not isolated, interoperability should be ensured at different levels (i.e., legislation, functional, syntactic, and semantic) and in different parts of the energy value chain, i.e., electricity generation, transmission, and consumption. Figure 1, for instance, gives a simplified illustration of information and electricity flows between stakeholders, while in reality, the electricity infrastructure and data exchange processes are very complex, i.e., infrastructure consists of many energy systems/infrastructures (generation, transmission, and demand infrastructures).

2.2. EU Energy Data Spaces and New Business Models

Interoperability and the possibility of building cross-border and cross-sector services are the focus of many initiatives in Europe; see, for instance, the Interoperable Europe program (<https://joinup.ec.europa.eu/collection/interoperable-europe/interoperable-europe>, (accessed on 23 May 2022)). The high-level vision of the European Union for 2030 is to create a single internal market through a standardized laws' system transposed in the national legislation of all member states and a single European data space for data exchange. The overall idea behind data spaces for Europe lies in setting up the needed software infrastructure that fosters data reuse, data valorization, and the creation of new business models and actors along the value chain of each industry, based on agreement on common standards and design principles. Hence, such software infrastructure will differentiate between data platforms where data and services reside and market platforms that facilitate matchmaking and the exchange of data and services. Implementing analytics and big data processing pipelines for more efficient and targeted services is part of the data platforms. The following innovative scenarios provide the playground to position our research questions.

In order to drive data-driven innovations, standardization should be applied, for instance, using metadata schemata, data representation formats, license terms for data and services, data integration, and data exchange approaches. The International Data Spaces (IDS) [16], launched in Germany at the end of 2014, follow the DE concept introduced above and have been foreseen for establishing the EU energy data space. The IDS reference architecture aims at

- Data governance according to regulations imposed by data providers;
- Ensuring a trusted and secure data exchange;
- Semantically representing main data concepts and relationships;
- Exchanging formats and protocols;
- Providing software design principles for guiding the implementation of the reference architecture components.

IDS provides building blocks for the development of data-driven services, while data sovereignty for data providers is guaranteed. IDS propose a message-based infrastructure to enable the communication of the different nodes and components in a DE. Moreover, IDS resorts to the Semantic Web standards to express the content and meaning of the shared data source. The resource description framework (RDF) and ontologies defined using RDF are proposed to specify metadata, and data control and protection in a decentralized or federated DE. The IDS shared information model states standards for representing content, concept, community of trust, commodity, and communication. Proposed W3C standards including SHACL (<https://www.w3.org/TR/shacl/>, (accessed on 23 May 2022)) are proposed to express content and integrity constraints; SKOS (<https://www.w3.org/2004/02/skos/>, (accessed on 23 May 2022)) for modeling concepts and relationships; and PROV (<https://www.w3.org/TR/prov-overview/>, (accessed on 23 May 2022)) for representing data and service provenance.

2.3. Example Scenario: The RES Forecasting

Modernization of the grid implies fast integration of RESs, adapted power system planning, new forecasting methods, more flexible use of power plants, standardized data exchange, increased transfer capacity, and others. The volatile production of renewable energy sources creates particular challenges for the daily electricity balancing process, i.e., balancing the deviations between the planned or forecast production and demand on the one side and the actual performance in real-time on the other side [17]. Given that renewable energy sources are increasing their share in the electricity market, to maintain the stable grid; i.e., to match the production and the demand, it is crucial to have accurate predictions of the expected accessible energy. In this regard, the need for a precise RES production forecaster is obvious. Addressing Europe's current energy crisis due to under-performance by wind power [18] demands an accurate forecast of RES production capacities (wind and PV plants) and estimates the effects of the production on the grid. Moreover, interoperability between different analytical services and cross-service integration requires harmonizing domain-specific vocabularies applied in the information layer and reusing the models to expose analytical services on marketplace platforms. As standardization at different levels (such as metadata schemata, data representation formats, and licensing conditions of open data) are demanded, the authors formulated the following research questions as pillars of the work:

- **RQ1** How to establish a software platform taking into consideration open-standards and reference architecture (e.g., SGAM [19], BRIDGE Data Management Reference Architecture [20])?
- **RQ2** Which ontologies cover the needs for modeling the energy value chain and ensure uniform access [21] to data collected with the proprietary SCADA system?
- **RQ3** How to build a knowledge graph that will be ready for integration with services in future energy marketplaces?
- **RQ4** From a business perspective, what are the benefits of advanced analytics for different kinds of energy actors?

Example Data Sources

The RES data source provides relevant data and information regarding renewable energy source (RES) systems, and, therefore, could provide the following datasets:

- **Production** dataset contains historical wind power production measurements from the wind power plant.
- **Predictive maintenance** dataset contains high-resolution measurements collected by the phasor measurement unit (PMU) installed on the renewable power plant.
- **Meteorological** dataset contains both historical and forecasted meteorological data, which are crucial for providing precise RES production forecast.
- **RES effects** contains estimations regarding the effects of the renewable energy source on the power system based on the PMU measurement (predictive maintenance dataset).

3. Developing a Multi-Layer Software Architecture

The approach presented in this section is inspired by the International Data Space (IDS) initiative and the EU Data Strategy. It will showcase how a "network" of distributed data integration platforms can be instantiated in the energy value chain for establishing a »network of trusted data«. Herein, we propose an energy big data integration platform as an instantiation of a data ecosystem (DE) [12]; see Figure 2.

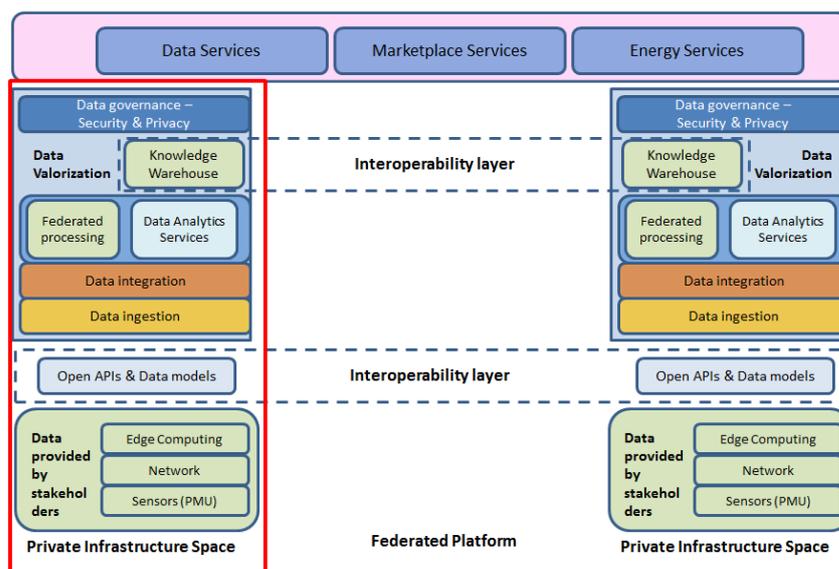


Figure 2. The energy big data integration platform as a knowledge-driven data ecosystem.

3.1. Energy Big Data Integration Platform

An energy big data integration platform is composed of several data integration platforms (one per Node i). Each node corresponds to a DE and can be integrated on the central level through mappings among nodes, data sharing, and service agreements. Each node (in Figure 2 denoted by Node, see red rectangle) applies a data integration process to a specific use case and can deploy its services for query processing, analytics as well as dashboards. Communication between nodes needs to be through an access agreement and can employ data connectors (IDS connectors) to secure data exchange according to data access contracts and regulations. Nodes have control over their data and may have data integrated in unified knowledge graphs. Moreover, each individual knowledge graph can be linked to knowledge graphs in other nodes, or to external knowledge graphs such as DBpedia [14], Wikidata [15], or others in the Linked Open Data cloud (<https://lod-cloud.net/>, (accessed on 23 May 2022)). Metadata are expressed using common semantic data models (e.g., CIM (https://ontology.tno.nl/IEC_CIM/, (accessed on 2 May 2022)), DCAT (<https://www.w3.org/TR/vocab-dcat-3/>, (accessed on 2 May 2022)), and SKOS (<https://www.w3.org/2004/02/skos/>, (accessed on 2 May 2022))), and the RDF mapping language (RML [22]) is utilized to define each pilot dataset in terms of the energy semantic data models. This framework enables pilots to preserve data sovereignty, privacy, and protection of data and analytical outcomes, as foreseen in IDS. More importantly, it represents a decentralized infrastructure empowered with the components that pave the way for interoperability across stakeholders.

3.2. Instantiating a DE

The main features of the energy data integration platform are illustrated in the instantiation of a DE; for instance, in the Serbian pilot depicted in Figure 3, DEs shall be instantiated at

- Producer site (e.g., at a wind power plant, a unified knowledge graph shall be integrated with the production forecast and the predictive maintenance services);
- Supplier site, an organization that integrates data from many producers and sells electricity to TSO (e.g., the power industry of Serbia might be interested to integrate the data sources from power plants it owns and manages);
- Transmission system operator site, an organization that operates and balances the grid (e.g., the joint stock company EMS might be interested in improving the data integration and the transparency of data exchanged with other actors).

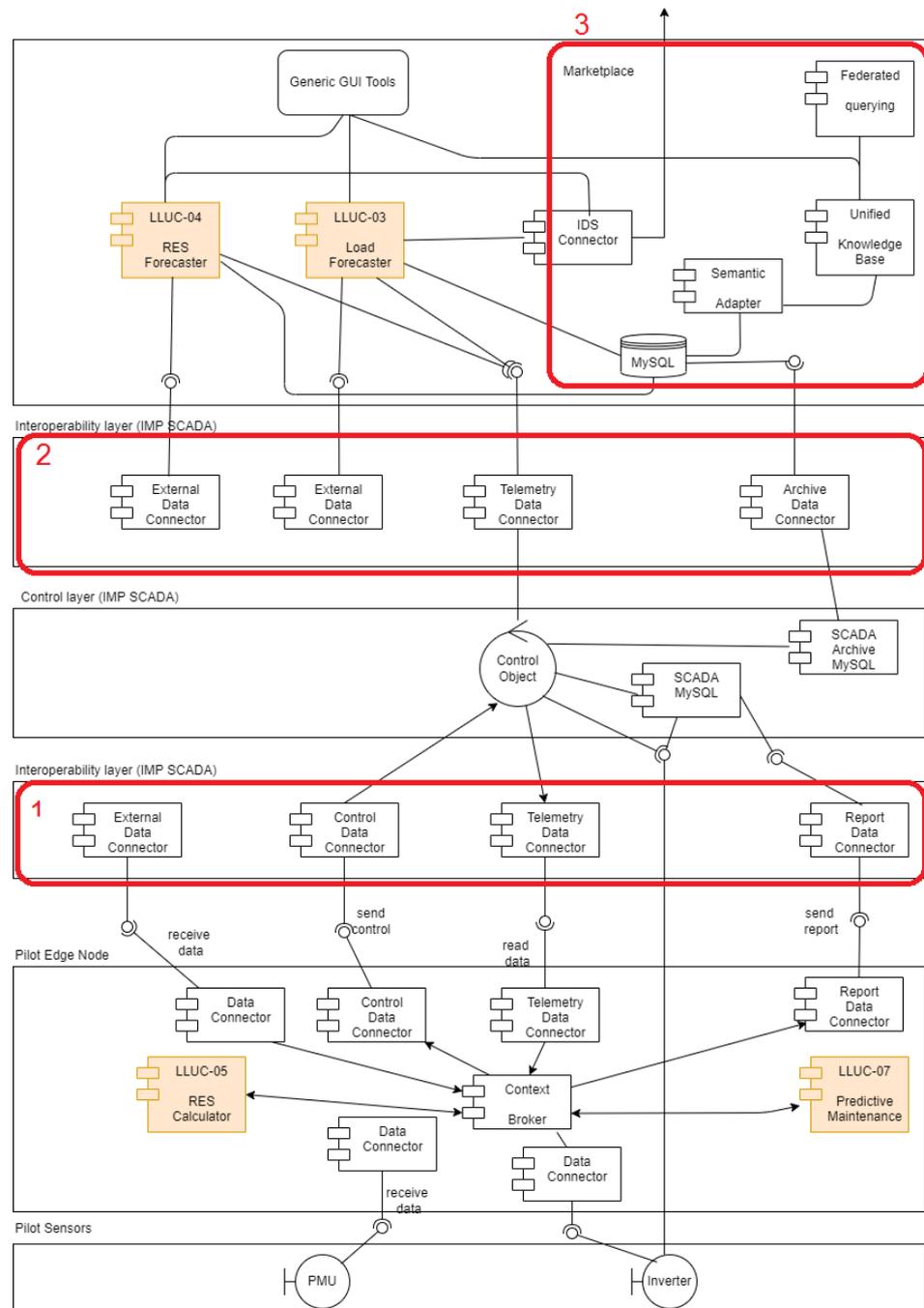


Figure 3. Software architecture for one node (Institute Mihajlo Pupin, 2021).

3.2.1. Description of the DE Architecture on the Node Level

Figure 3 illustrates the adopted reference architecture on a DE node level. As already discussed, the energy domain is characterized by the presence of many actors, often large organizations, and there are many technological solutions and proprietary systems. In order to connect the existing platforms and advanced business services, an interoperability layer is responsible for transforming data collected from data sources into structures that can be managed by analytical applications. Figure 3 depicts three interoperability layers:

- The first layer (denoted with number 1) ensures syntactic interoperability and communication with physical architecture, for example, phasor measurement units for collecting high-resolution data about the generating units (inverters of PV production

plant or turbines of wind plant); a building or a complex of buildings; or single devices, such as energy meters on the consumption side.

- The second layer (denoted with number 2) ensures syntactic interoperability and communication between the control SCADA system and the intelligent layer, analytical services that work on top of one kind of data, for instance, one MySQL base is used for retrieving the data.
- The third layer (denoted with number 3) ensures semantic interoperability for advanced business services where integration of different big data sources are needed because of different interoperability issues.

Interoperability issues explored in the process of building the unified knowledge graph are related to

- Representation of attributes' values: Timestamps standardization, measurement unit generalization, and measurement scale.
- Granularity: different aggregations (daily vs. hourly, weather at wind farm vs. at the city level); different measurement for same time intervals (example temperature from wind farm sensor and temp in WeatherBit of the city).
- Structuredness: SCADA—structured (MySQL); Weatherbit—semi-structured (JSON), ENTSO-E—semi-structured data (XSML).
- Schematic interoperability: various representations of attributes and concepts are used for modeling the same semantic concept (*outtemperature* in Wind RES database vs. *temp* at WeatherBit; *obtime* at WeatherBit vs. *timestamp* in Wind RES database).

3.2.2. Instantiating a Node at the Producers' Site

SCADA RES data are available in real time through a MySQL database. Data operators for preprocessing, mapping, linking, transformation, and validation are applied to the pilot data sources for creating a materialized version of the unified knowledge graph. The mapping rules among data sources and the unified schema are part of the DE as well. Furthermore, mappings between concepts from different ontologies are included in each DE. Data sources are also described in terms of provenance and main properties; these descriptions are utilized for the creation of a knowledge graph (e.g., by using *SDM-RDFizer* [23]) and during query processing (e.g., by using *Ontario* [24]). Links between entities in the knowledge graph and external data sources can be made by performing entity linking. Tools such as *Falcon2.0* [25] can be applied to linking the pilots' datasets with external knowledge graphs such as *DBpedia* and *Wikidata*, while *SHACL* validation engines (e.g., *Trav-SHACL* [26]) enable the validation of integrity constraints. Lastly, *RDF* knowledge graph will feed the semantic-based analytics engine *SANSA* [27] to perform tasks of knowledge discovery and prediction.

3.2.3. Instantiating an IDS Data Connector

The *IDS Connector* is one of the central technological building blocks of *IDS-based* digital ecosystems that allow the participant (node) to exchange, share, and process digital content while the data sovereignty of the data owner is guaranteed. The data connector should provide metadata to the data consumer connector. Hence, the data harmonization on an ecosystem level is a prerequisite for the smooth integration of different data connectors. The data connector architecture (technical interface description, authentication mechanism, exposed data sources, and associated data usage) is out of the scope of this paper; see more information in *PLATOON D3.4* [28].

4. Data Standardization and Harmonization

4.1. Developing a Global Schema for the Energy Domain

Different ontologies are proposed in the literature for development of a global schema including (i) upper ontologies (e.g., *SUMO*, *Dolce*, *BFO*), (ii) core ontologies (e.g., agent ontology, time ontology), (iii) domain ontologies for a specific domain, and (iv) domain-specific ontologies that can be reused and extended in order to meet a specific need of the

Example—RES forecasting: Another use case is related to a resource connected to the grid. Independent producers (IPP) and producers (cim:Producer) from distributed and renewable sources (DER) will be actors in the balance reserve market in the future. The goal of this scenario is to develop and test a service for more accurate prediction of renewable energy generation from RES plants (cim:Plant). Electricity production, however, from solar and wind plants (cim:Plant) is subject to considerable forecast errors that drive demand for balancing, i.e., for (cim:ReserveReq). The amount for each reservation is defined by the agreement (cim:Agreement) on the provision of system services signed between the transmission system operator (cim:SystemOperator) and the balancing service provider (cim:BalanceSupplier). Once the global schema has been developed, it can be used across the nodes established in the energy data ecosystem.

Example SPARQL query: Showing the total energy produced (active power) by WindFarms in Montenegro, on 31 December 2017. An example SPARQL query is presented in Figure 5.

```

PREFIX cim: <http://www.iec.ch/TC57/CIM#>
PREFIX platoon: <https://w3id.org/platoon/>
PREFIX seas: <https://w3id.org/seas/>
PREFIX wgs84_pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX qudt: <http://www.qudt.org/2.1/schema/qudt/>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX sch: <https://schema.org/>
PREFIX qudt: <http://www.qudt.org/2.1/schema/qudt/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?windfarm SUM(?value) as ?totalPower ?unit
WHERE {
  ?windfarm a platoon:WindFarm .
  ?windfarm platoon:country <http://platoon.eu/Country/ME>.
  ?turbine seas:isMemberOf ?windfarm .
  ?turbine a platoon:OffshoreWindTurbine .
  ?turbine seas:producedElectricPower ?powerprop .
  ?powerprop seas:evaluation ?prod_eval .
  ?prod_eval seas:evaluatedSimpleValue ?value.
  ?prod_eval qudt:unit ?unit.
  ?prod_eval seas:hasTemporalContext ?context .
  ?context time:hasBeginning ?start.
  ?context time:hasEnd ?end
  Filter (xsd:date(?start) = xsd:date('2017-12-31'))
}
GROUP BY ?windfarm ?unit

```

Figure 5. Example SPARQL query.

4.2. Unified Knowledge Graph Creation Process

In this section, two scenarios of the knowledge graph creation process and their pros and cons are discussed. Creating a knowledge graph from heterogeneous data sources at the supplier site requires the description of the entities in the data sources using RDF vocabularies. Additionally, it requires data curation and entity alignment to enhance data quality, e.g., missing values or duplicates. Two types of knowledge graph creation strategies are materialized (i.e., data warehousing) and virtual (i.e., via semantic data lakes). Both strategies are applicable for the above discussed use cases.

Materialized Knowledge Graph Creation Process: In a materialized knowledge graph creation process, data from individual data sources are loaded and materialized into an RDF format and stored in a physical database, the so-called triplestore. Figure 6 shows the data curation and integration subcomponents for creating a unified knowledge graph. The ingestion and preprocessing component is the gateway to the knowledge graph creation process. Input from producers' data sources is first stored in a raw data repository, i.e., staging repository. Any preprocessing steps, such as cleaning, normalization, and aggregation, that are predefined for input data are applied and provenance is recorded. The data integrator component then orchestrates the knowledge graph creation process according to the data source's configuration by invoking the linking and enrichment, SDM-RDFizer/semantifier, and data validation subcomponents, and finally integrating data to the supplier's unified knowledge graph. The linking and enrichment component performs

entity linking and enrichment using external as well as existing materialized knowledge graphs. The SDM-RDFizer/semantifier component transforms non-semantic, i.e., raw, data to a RDF graph based on mapping rules. Data validation component checks data constraint conformance.

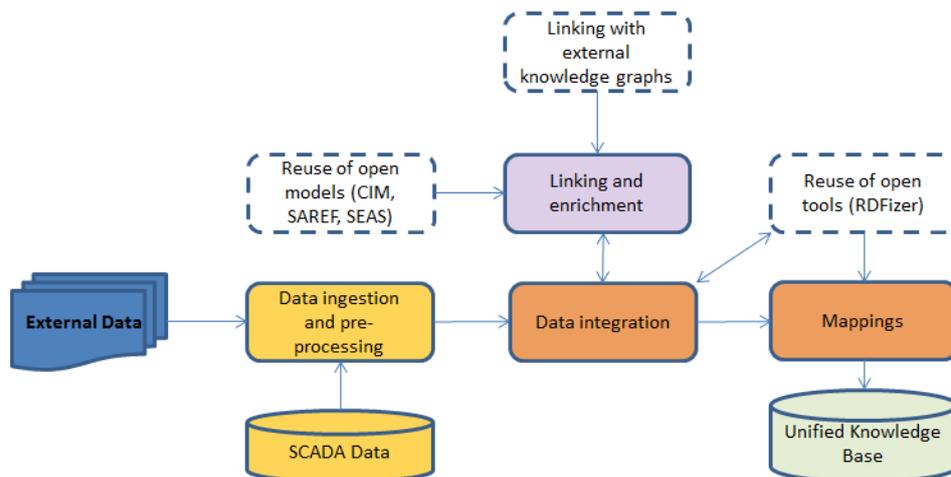


Figure 6. Unified knowledge graph creation process.

Virtual Knowledge Graph Creation Process: In a virtual knowledge graph creation process, data remain in the sources (in raw format) and are accessed as needed during query time. The federated query processing component can handle this process. The federated query processing component employs the data source descriptions stored in the metadata store to perform the integration during query time. Metadata about the number of data sources available, the provenance of the datasets, and mapping rules to transform data to RDF graph are stored in a separate data store available for both materialized and virtual data integration processes. If the datasets are already included in the materialized knowledge graph, then the federated query processing component can directly access them without performing data transformation at query time. However, if the data sources are stored in raw format, then the data transformation rules will be applied only for the part of the dataset required to answer the query; see also Janev et al. (2021) [8].

4.3. Data Harmonization

The data catalog (DCAT) vocabulary provides the basis for the harmonized data source and services description. In our approach, classes `dcatalog:Catalog` and `dcatalog:Dataset` have been used to describe the collection of datasets employed in a service in the way that it is understandable by humans and also by machines. The SDM-RDFizer/semantifier component is applied in the pipeline to make the use of the metadata, and guided by mapping rules, to generate a harmonized description that can be uploaded in a form of RDF knowledge base into a Virtuoso SPARQL endpoint for further exploration. Thus, datasets are annotated with concepts from the energy domain vocabularies whose meaning is commonly accepted by the energy sector community. Figure 7 presents the data harmonization pipeline, while Figure 8 visualizes the links between the target datasets, e.g., RES-PROduction. As observed, these annotations allow for establishing connections among the energy sector datasets. More importantly, they provide the basis for a semantic search based on classes of energy vocabularies, and enable a common understanding of the data collected from these datasets.

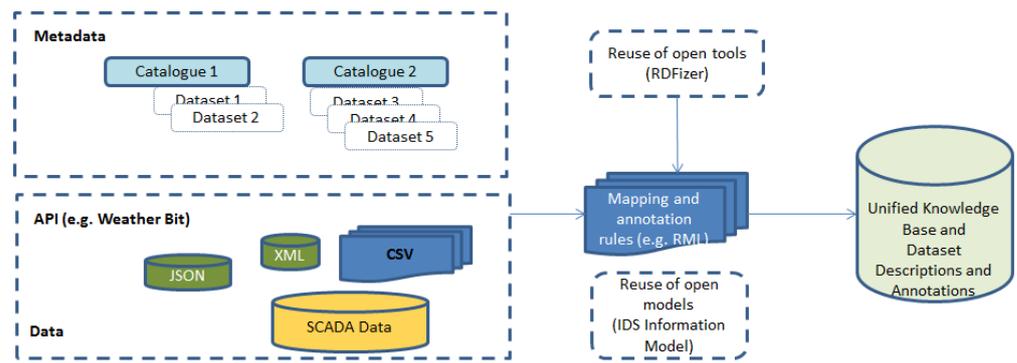


Figure 7. Data harmonization pipeline.

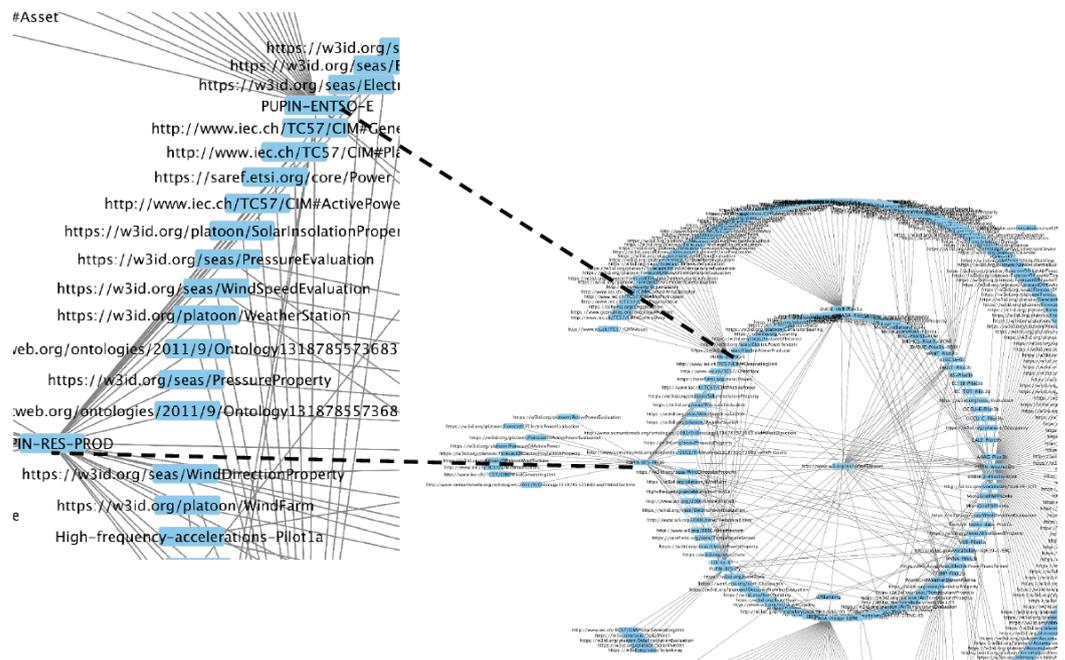


Figure 8. Annotation of data sources. Visualization powered by Cytoscape.

By using Cytoscape (<https://cytoscape.org/>, (accessed on 23 May 2022)), the main properties of annotation knowledge graph are analyzed in terms of graph measures (e.g., number of nodes, number of edges, average number of neighbors), see Table 1. The results place into perspective the relevant role that annotations from domain-specific vocabularies have in the discovery of connections among energy datasets.

Table 1. Datasets.

Dataset	Number of Annotations
PUPIN-RES-PROD	34
PUPIN-RES-PV	26
PUPIN-ENTSO-E	22
PUPIN-RES-Effects	4

5. Knowledge Exploitation

5.1. Traversing the Knowledge Graph

Once the knowledge graph creation process is established, exploring the knowledge base will be possible via a query engine. Additionally, data exploration and knowledge discovery services can be employed. Results of executing a federated query can be used as input of data analytics or knowledge discovery tasks. As the knowledge base is defined

through mapping to semantic data models for energy, the query processing engine is able to process queries posed using the SPARQL query language. If the materialization approach is applied and data are stored in a centralized triple store, e.g., Virtuoso, then the knowledge base can be accessed using SPARQL query over the query engine embedded in the triple store, see Figure 9. However, if the materialized knowledge base is large, then partitioning and distribution is necessary for timely response from the query engine and handling the resource requirements to store such large data in expensive servers. Such distribution of data needs to be accessed through a federated query engine that is able to distribute the posed query to each partition and merge data returned from them. Virtual integration approach can also be applied over heterogeneous data sources. In this case, the query processing engine not only queries each data source and merges results, but also should be able to transform raw data into the semantic models specified in the mappings during query time [8].

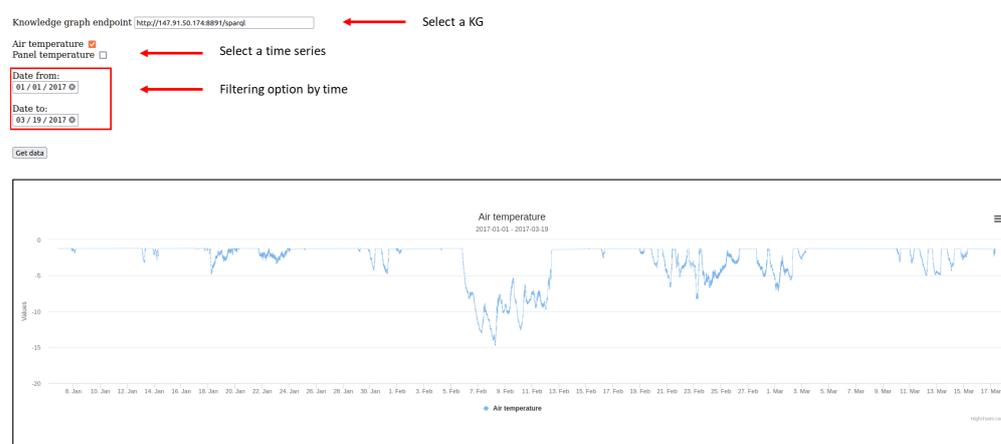


Figure 9. Energy analytics dashboard.

5.2. Federated Query Processing

A unified knowledge graph can be partitioned into various graphs accessible independently via SPARQL endpoints. A federation of knowledge graphs comprises all these graphs together with the external knowledge graphs linked from the unified knowledge graph (e.g., DBpedia [14] and Wikidata [15]). Although each graph in a federation can process SPARQL queries individually, queries that require data collected from one or more knowledge graphs need to be executed by a federated query engine [30]. Federated query processing demands the implementation of data management techniques for selecting the knowledge graphs that will answer a query and decomposing the queries into sub-queries on the chosen knowledge graphs. Moreover, a federated query engine must be equipped with physical operators capable of merging the answers produced by executing the sub-queries on top of the selected knowledge graphs. Furthermore, the federated query engine identifies query plans that minimize the execution time of the queries. The federated query engine interoperates across the unified knowledge graphs and DBpedia in this project.

6. Integration of Advanced Analytics

6.1. Res Forecasting

For the purpose of the development of the ML-based wind power production forecasting model for the Krnovo Wind Plant (Montenegro), the RES-PROD dataset was created. Meteorological data were collected from the local meteo-station and included the following characteristics: wind speed, wind direction, and temperature. The collected dataset covered a six-month-long period and had measurements with hourly time resolution. Production measurements were obtained from 26 wind turbines installed with 2.85 MW capacity. When wind power production forecast is considered, hybrid neural network approaches intend to provide improved estimation performances in comparison with other methods, which is

the main reason why authors decided to choose the hybrid approach within this research. Apart from the optimization regarding the number and type of the hidden layers, activation functions were chosen carefully as well. Namely, since the network output is limited with the turbine capacity, *tansig* activation function was selected. Training process for the final model was carried out using ADAM optimization method with the learning rate 0.001 and mean square error as the criterion function. The *dockerized* service was integrated with the knowledge graph discussed above. The federated query engine can select results based on the user query set via a graphical user interface; see Figure 9.

6.2. Data Analytics on the Edge

The integration of RES into the low-voltage (LV) grid together with the fusion of environmentally friendly technologies entering the low-voltage grid at the consumer's site presents a new challenge for the design of a reliable and manageable power grid [31]. Data services on edge, together with the sensors, represent the bottom-most layer in the architecture. There are many heterogeneous data sources in the field, from hardware sensors with analog output to more advanced intelligent electronic devices (IEDs) with standardized protocols and APIs. We developed a new method to analyze the impact of PV power plants before they are integrated into the grid, which can be easily extended to other types of RESs. In the case where the PV power plant is already installed, we first estimate the situation without the PV power plant to understand the impact of PV integration. By comparing the worst-case scenarios with and without the PV power plant (e.g., maximum or minimum daily voltages) over a longer period of time, we can estimate the impact and calculate the grid insertion capacity at that time. Once the data are available from a node, it can be semantically enriched to better understand where it came from and what exactly the processed values are. Once many nodes are integrated with the edge data, various big data analyses can be performed, for example, to reconstruct the topology of the critical grid infrastructure and to better understand and monitor the bidirectional energy flows in the power grid.

7. Discussion

In the last decade, the big data paradigm has gained momentum and is generally employed by businesses on a large scale to create value that surpasses the investment and maintenance costs of data. The energy sector is an example where tremendous amounts of data are collected from numerous sensors, which are generally attached to different plant subsystems. The new paradigm of DEs for smart grids that includes renewable energy sources challenges the existing network infrastructure and the energy management systems even more. In the EU project PLATOON, we have explored the possibilities for

- The use of new approaches capable of data managing and processing for extending the analytics services portfolio of various energy stakeholders. Examples include ESCOs, DSOs, and utilities to achieve two-way flows of electricity and information for optimized generation, distribution, and electricity consumption.
- Distributed/edge processing and data analytics technologies to optimize the operation of the real-time energy system management and automate the “monitor–forecast–optimize–control” loop.
- Effective integration of relevant digital technologies. It will transform energy systems from the top down and move from centralized production and rigid distribution framework into a collaborative ecosystem of self-managed prosumers able to act independently on the liberalized energy markets.

Next, we showcase how “networks” of distributed data integration platforms can be instantiated in the energy value chain for establishing a »network of trusted data«. Some benefits for main actors are the following:

- **Secure data exchange:** For instance, using the industrial data space concept that features various levels of protection, data are exchanged securely across the entire data supply chain (and not just in bilateral data exchange).

- **Data governance and sovereignty:** In a network of energy DEs, data owners determine the terms and conditions of use of the data provided, while data sovereignty always remains with the respective data provider. A provider makes data available to be requested by certain contractors in a data space by its own rules. Additionally, the provider can also offer data services (e.g., via an »AppStore«) to be found by all DE participants.
- **Innovative scalable and replicable energy management services:** a network of energy DEs opens opportunities for new data-driven and model-driven services that will complement and enhance the existing, e.g., balancing services, energy generation and consumption intelligent forecasts services, and energy performance assessment services.

8. Related Work

Ref. [32] highlights the value of data-driven solutions in the digitization era and outlines the challenges that need to be addressed in DEs in emerging areas such as maritime, manufacturing, and science. Controlled and secured data exchange in a traceable way are among the most relevant challenges. However, despite years of research in data governance and management, trustability is still affected by the absence of transparent and traceable data-driven pipelines. The need for responsible data management (see [33]) intensifies with the growing impact of data on society.

As shown in the described scenarios, DEs for energy big data are demanded to provide computational methods and semantic-based formalisms (e.g., ontologies) to represent the meaning of the data to be shared and processed. The metadata layer comprises unified schemas, mappings between datasets and concepts in the unified schema, and alignments across ontologies. Furthermore, following the IDS reference architecture, integrity constraints are represented using declarative formalisms (e.g., SHACL), while data provenance and quality are described based on standard vocabularies (e.g., PROV and DQV). These semantic descriptions provide building blocks for documenting data sharing, integration, and processing. As a result, services for tracking down DE components can be provided.

Several approaches have been defined to follow the DE architecture with the aim of solving interoperability across heterogeneous datasets during query processing time; they are usually named as federated query engines. Exemplary approaches include GEMMS [34], PolyWeb [35], BigDAWG [36], Constance [37], and Ontario [24]. These systems collect metadata about the main characteristics of their datasets, e.g., formats and query capabilities. Additionally, they resort to a global ontology to describe contextual information and the relationships among datasets, for purposes of optimized data integration, query processing, and automated schema discovery in quasi-central settings. This metadata has shown to be crucial for enabling these systems to perform query processing needed in advanced business services effectively. Knowledge-driven DEs are built on these results and make available the semantic description of the data collections made available by stakeholders. Furthermore, a DE empowers federated query processing engines with factual statements about the integrity constraints satisfied by the data retrieved and merged during query processing. As a consequence, a new paradigm shift in data management is devised towards tracing down data integration during query processing.

9. Concluding Remarks

Smart grids are cyber-physical energy systems, the next evolution step of the traditional power grid, and are characterized by a bidirectional flow of information and energy. One of the requirements related to data access procedures in future business solutions for electricity markets is related to interoperability of energy services. Therefore, the overall goal of the paper was to showcase and evaluate data ecosystems (DEs) and the International Data Space concept for advanced business services in the energy sector. The International Data Space initiative is based on the use of semantic technologies for creation of knowledge-based systems that will aid machines in integrating and processing resources contextually

and intelligently. In our work, we showed how DES provides the building blocks for enhancing the interoperability of energy management applications/services; they also enable the integration of energy data in the European Energy Data Space. The metadata layer in DEs, together with the internal SCADA information model, can be used as an information hub (“knowledge graphs”) for (1) building data connectors that will facilitate integration of services in future integrated energy systems and (2) improving the explainability of machine learning services/analytical applications. The selection of models was made based on a set of scenarios (electricity balancing services, predictive maintenance services, and services for residential, commercial, and industrial sector). The proposed approach is being used in the EU-funded H2020 project PLATOON. The validation of all the computational components and unified schemas to fulfill the analytic requirements on a large scale (country level) is part of our future agenda.

Author Contributions: Investigation, V.J.; Methodology, M.-E.V.; Project administration, V.J.; Software, D.P. (Dea Pujić), D.P. (Dušan Popadić), E.I., A.S. and A.Č.; Supervision, V.J. and M.-E.V.; Validation, D.P. (Dea Pujić), D.P. (Dušan Popadić), E.I., A.S. and A.Č.; Visualization, D.P. (Dea Pujić), D.P. (Dušan Popadić), E.I., A.S. and A.Č.; Writing—original draft, V.J.; Writing—review & editing, M.-E.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the EU H2020 funded projects PLATOON (GA No. 872592), the EU project LAMBDA (GA No. 809965), the EU project SINERGY (GA No. 952140) and partly by the Ministry of Science and Technological Development of the Republic of Serbia (No. 451-03-9/2021-14/200034) and the Science Fund of the Republic of Serbia (Artemis, No.6527051).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. El-hawary, M.E. The smart grid—State-of-the-art and future trends. *Electr. Power Compon. Syst.* **2014**, *42*, 239–250. [CrossRef]
2. Liggesmeyer, P.; Rombach, D.; Bomarius, F. Smart Energy. In *Digital Transformation*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 335–351. [CrossRef]
3. Noy, N.; Gao, Y.; Jain, A.; Narayanan, A.; Patterson, A.; Taylor, J. Industry-Scale Knowledge Graphs: Lessons and Challenges. *Commun. ACM* **2019**, *62*, 36–43. [CrossRef]
4. Jain, S. Exploiting Knowledge Graphs for Facilitating Product/Service Discovery. *arXiv* **2020**, arXiv:2010.05213.
5. Roman, D.; Alexiev, V.; Paniagua, J.; Elvæsæter, B.; Marius von Zernichow, B.; Soylyu, A.; Simeonov, B.; Taggart, C. The euBusinessGraph ontology: A lightweight ontology for harmonizing basic company information. *Semant. Web J.* **2022**, *13*, 41–68. [CrossRef]
6. Lackshen, G.; Janev, V.; Vraneš, S. Arabic Linked Drug Dataset Consolidating and Publishing. *Comput. Sci. Inf. Syst.* **2021**, *18*, 729–748. [CrossRef]
7. Mijović, V.; Tomašević, N.; Janev, V.; Stanojević, M.; Vraneš, S. Ontology Enabled Decision Support System for Emergency Management at Airports. In Proceedings of the I-SEMANTICS 2011, International Conference on Semantic Systems, Graz, Austria, 7–9 September 2011; ACM: New York, NY, USA, 2011; pp. 163–166. [CrossRef]
8. Janev, V.; Vidal, M.E.; Endris, K.; Pujić, D. *Managing Knowledge in Energy Data Spaces*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 7–15.
9. Capiello, C.; Gal, A.; Jarke, M.; Rehof, J. Data Ecosystems: Sovereign Data Exchange among Organizations (Dagstuhl Seminar 19391). In *Dagstuhl Reports*; Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik: Dagstuhl, Germany, 2020; Volume 9.
10. Oliveira, M.I.S.; Lóscio, B.F. What is a Data Ecosystem? In Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, Delft, The Netherlands, 30 May–June 1 2018; pp. 1–9.
11. European Commission. A European Strategy for Data (19 February 2020, COM(2020) 66 Final). 2020. Available online: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en (accessed on 23 May 2022).
12. Geisler, S.; Vidal, M.; Capiello, C.; Lóscio, B.F.; Gal, A.; Jarke, M.; Lenzerini, M.; Missier, P.; Otto, B.; Paja, E.; et al. Knowledge-Driven Data Ecosystems Toward Data Transparency. *ACM J. Data Inf. Qual.* **2022**, *14*, 3:1–3:12. [CrossRef]
13. European Commission. The European Green Deal (11 December 2019, COM(2019) 640 Final). 2019. Available online: https://ec.europa.eu/info/publications/communication-european-green-deal_en (accessed on 23 May 2022).
14. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 722–735.
15. Vrandečić, D.; Krötzsch, M. Wikidata: A free collaborative knowledgebase. *Commun. ACM* **2014**, *57*, 78–85. [CrossRef]

16. Otto, B.; Haas, C.; Pettenpohl, H.; Lohmann, S.; Huber, M.; Pullmann, J.; Auer, S. Reference Architecture Model for the Industrial Data Space. 2017. Available online: https://www.fit.fraunhofer.de/content/dam/fit/en/documents/Industrial-Data-Space_Reference-Architecture-Model-2017.pdf (accessed on 23 May 2022).
17. Janev, V.; Jakupović, G. Electricity Balancing: Challenges and Perspectives. In Proceedings of the 2020 28th Telecommunications Forum (TELFOR), Belgrade, Serbia, 24–25 November 2020; pp. 1–4. [CrossRef]
18. Blackmon, D. How Europes Energy Crisis Could Force The EU To Adopt More Sensible Policies. 2022. Available online: <https://www.forbes.com/sites/davidblackmon/2022/01/03/how-europes-energy-crisis-could-force-the-eu-to-adopt-more-sensible-policies/?sh=4e5e9a6e3ed3> (accessed on 23 May 2022).
19. Hooshyar, H.; Vanfretti, L. A SGAM-based architecture for synchrophasor applications facilitating TSO/DSO interactions. In Proceedings of the 2017 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, USA, 23–26 April 2017; pp. 1–5. [CrossRef]
20. Commission, E. BRIDGE European Energy Data Exchange Reference Architecture. 2021. Available online: https://ec.europa.eu/energy/sites/default/files/documents/bridge_wg_data_management_eu_reference_architecture_report_2020-2021.pdf (accessed on 23 May 2022).
21. Mami, M.; Graux, D.; Scerri, S.; Jabeen, H.; Auer, S.; Lehmann, S. Uniform Access to Multiform Data Lakes using Semantic Technologies. In Proceedings of the 21st International Conference on Information Integration and Web-based Applications and Services, Munich, Germany, 2–4 December 2019; pp. 313–322. [CrossRef]
22. Dimou, A.; Vander Sande, M.; Colpaert, P.; Verborgh, R.; Mannens, E.; Van de Walle, R. RML: A generic language for integrated RDF mappings of heterogeneous data. In Proceedings of the 7th Workshop on Linked Data on the Web, Seoul, Korea, 8 April 2014.
23. Iglesias, E.; Jozashoori, S.; Chaves-Fraga, D.; Collarana, D.; Vidal, M.E. SDM-RDFizer: An RML interpreter for the efficient creation of RDF knowledge graphs. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual Event, 19–23 October 2020; pp. 3039–3046.
24. Endris, K.M.; Rohde, P.D.; Vidal, M.; Auer, S. Ontario: Federated Query Processing Against a Semantic Data Lake. In *Lecture Notes in Computer Science, Proceedings of the Database and Expert Systems Applications—30th International Conference, DEXA 2019, Linz, Austria, 26–29 August 2019*; Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I., Eds.; Springer: Berlin/Heidelberg, Germany, 2019; Volume 11706, pp. 379–395. [CrossRef]
25. Sakor, A.; Singh, K.; Patel, A.; Vidal, M. Falcon 2.0: An Entity and Relation Linking Tool over Wikidata. In Proceedings of the CIKM'20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, 19–23 October 2020; d'Aquin, M., Dietze, S., Hauff, C., Curry, E., Cudré-Mauroux, P., Eds.; ACM: New York, NY, USA, 2020; pp. 3141–3148. [CrossRef]
26. Figuera, M.; Rohde, P.D.; Vidal, M. Trav-SHACL: Efficiently Validating Networks of SHACL Constraints. In Proceedings of the The Web Conference WWW, Ljubljana, Slovenia, 19–23 April 2021.
27. Lehmann, J.; Sejdiu, G.; Bühmann, L.; Westphal, P.; Stadler, C.; Ermilov, I.; Bin, S.; Chakraborty, N.; Saleem, M.; Ngomo, A.N.; et al. Distributed Semantic Analytics Using the SANSA Stack. In *Lecture Notes in Computer Science, Proceedings of the Semantic Web—ISWC 2017—16th International Semantic Web Conference, Vienna, Austria, 21–25 October 2017*; d'Amato, C., Fernández, M., Tamma, V.A.M., Lécué, F., Cudré-Mauroux, P., Sequeda, J.F., Lange, C., Heflin, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; Volume 10588, pp. 147–155. [CrossRef]
28. Maggio, M.; Savarino, V.; Rashid, T.T.; Harish, T.M.; Idoia Murua, U.I. Deliverable D3.4 Open Source Data Connector. 2021. Available online: <https://cordis.europa.eu/project/id/872592/results> (accessed on 23 May 2022).
29. Janev, V.; Popadić, D.; Pujčić, D.; Vidal, M.E.; Endris, K. Reuse of Semantic Models for Emerging Smart Grids Applications. *arXiv* **2021**, arXiv:2107.06999.
30. Endris, K.M.; Vidal, M.E.; Graux, D. Chapter 5 Federated Query Processing. In *Knowledge Graphs and Big Data Processing*; Janev, V., Graux, D., Jabeen, H., Sallinger, E., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 73–86. [CrossRef]
31. Fathabad, A.; Cheng, J.; Pan, K.; Qiu, F. Data-Driven Planning for Renewable Distributed Generation Integration. *IEEE Trans. Power Syst.* **2020**, *35*, 4357–4368. [CrossRef]
32. Gelhaar, J.; Otto, B. Challenges in the Emergence of Data Ecosystems. In Proceedings of the 24th Pacific Asia Conference on Information Systems, PACIS, Dubai, United Arab Emirates, 22–24 June 2020; Vogel, D., Shen, K.N., Ling, P.S., Hsu, C., Thong, J.Y.L., Marco, M.D., Limayem, M., Xu, S.X., Eds.; 2020; p. 175.
33. Stoyanovich, J.; Howe, B.; Jagadish, H.V. Responsible Data Management. *Proc. VLDB Endow.* **2020**, *13*, 3474–3488. [CrossRef]
34. Quix, C.; Hai, R.; Vatov, I. GEMMS: A Generic and Extensible Metadata Management System for Data Lakes. In Proceedings of the 28th International Conference on Advanced Information Systems Engineering (CAiSE 2016), CEUR-WS, Ljubljana, Slovenia, 13–17 June 2016; pp. 129–136.
35. Khan, Y.; Zimmermann, A.; Jha, A.; Gadepally, V.; D'Aquin, M.; Sahay, R. One Size Does Not Fit All: Querying Web Polystores. *IEEE Access* **2019**, *7*, 9598–9617. [CrossRef]
36. Duggan, J.; Elmore, A.J.; Stonebraker, M.; Balazinska, M.; Howe, B.; Kepner, J.; Madden, S.; Maier, D.; Mattson, T.; Zdonik, S. The BigDAWG Polystore System. *SIGMOD Rec.* **2015**, *44*, 11–16. [CrossRef]
37. Hai, R.; Geisler, S.; Quix, C. Constance: An Intelligent Data Lake System. In Proceedings of the 2016 International Conference on Management of Data, SIGMOD, San Francisco, CA, USA, 26 June–1 July 2016; ACM: New York, NY, USA, 2016; pp. 2097–2100. [CrossRef]