On the Use of Virtual Knowledge Graphs to Improve Environmental Sensor Data Accessibility

Jiantao Wu, Fabrizio Orlandi, Declan O'Sullivan, and Soumyabrata Dev, *Member, IEEE*

***Abstract*—The rapid proliferation of environmental sensor networks (ESNs) used for monitoring environmental systems, such as meteorology and air quality, and advances in database technologies [e.g., structured query language (SQL)] has made significant progress in sensor data management. Notwithstanding the strength of these databases, they can inevitably lead to a data heterogeneity problem as a result of isolated databases with distinct data schema, which are expensive to be accessed and preprocessed when the data are consumed spanning multiple databases. Recently, knowledge graphs have been used as one of the most popular integration frameworks to address this data heterogeneity problem from the perspective of establishing an interoperable semantic schema (also known as ontology). However, the majority of the proposed knowledge graphs in this domain are a product of an extraction–transform–load approach with all the data physically stored in a triplestore. In contrast, this article examines an approach of virtualizing knowledge graphs on top of the SQL databases as the means to provide a federated data integration approach for enhanced heterogeneous ESNs' data access, bringing with it the promise of more cost efficiency in terms of input/output, storage, etc. In addition, this work also considers some motivating application scenarios regarding the efficiency of time-series data access. Based on a performance comparison between the proposed integration approach and some popular triplestores, the proposed approach has a significant edge over triplestores in multiple time-series structuring and acquisition.**

***Index Terms*—Environmental sensor networks (ESNs), ontology-based data access, structured query language (SQL) databases, virtual knowledge graphs (VKGs).**

I. Introduction

With the growing usage of information and communication technologies in the Internet of Things sector [1], [2], environmental measurements are able to be recorded in a timely manner through wireless environmental sensor networks (ESNs) and station-based ESNs, providing substantial data sources for environmental data consumers. In general, most of these ESN data sources, such as National Oceanic and Atmospheric Administration (NOAA)[1] and PurpleAir,[2] are made accessible on the Internet at scale through the use of structured query language (SQL) databases as the back-end data server. However, due to the vast content coverage of the earth's environment ecosystem, each environmental sensor SQL-based data service is typically the outcome of a limited monitoring scope in terms of time, location, or in variety of scales. This issue is also impacting downward data analysis model development in this field [3], [4], since the approach to design robust modeling frameworks, such as popular deep learning methods, is intrinsically linked to the amount and diversity of upward environmental data input [5], [6]. Consequently, how to preprocess heterogeneous environmental sensor data from diverse data suppliers in order to provide easy access to a broader range of consistent ESN data to facilitate environmental data (e.g., climate) analysis [7], [8], [9] has become one of the most critical questions confronting academia and industry in this area.

Acquisition and integration of heterogeneous environmental sensor data from various sources continue to require considerable work even today [10]. Technically, this difficulty often occurs as a result of the interoperability gap [11] among ESNs, including different hypertext transfer protocols (HTTPs) schemas, and naming conventions [12], [13]. In a more general context, "interoperability" refers to the degree to which information may be exchanged across systems. It varies among domains and may be considered at various degrees, ranging from no interoperability to complete interoperability [14]. In our research setting, the interoperability gap mainly occurs in the data access to various data sources. For example, for data consumers who intend to acquire and integrate data from multiple data sources, they must handle the schema mismatch before the data are ready for their data processing tasks.

A. Semantic Approaches for Data Integration

Recently, semantic research has made significant progress in establishing semantic interoperability across data from diverse areas through ontology-based data access/integration (OBDA/I) [15]. The knowledge graph (KG) is a critical conceptualizing framework in the semantic web research community, which serves as a universal model to model everything shared by heterogeneous entities [16] by utilizing the data-graph-based World Wide Web Consortium (W3C) standard representation resource description framework (RDF) [17]. The augmentation of access to KGs on the World Wide Web can be significantly heightened through the publication and annotation of entities and relationships within KGs in adherence to linked

Manuscript received 11 July 2023; revised 9 January 2024; accepted 9 February 2024. Date of publication 27 February 2024; date of current version 20 March 2024. This work was supported by the Science Foundation Ireland (SFI) through the SFI Research Centres Program under Grant 13/RC/2106_P2. *(Corresponding author: Soumyabrata Dev.)*

Jiantao Wu and Soumyabrata Dev are with the ADAPT SFI Research Centre, School of Computer Science, University College Dublin, D04 V1W8 Dublin, Ireland (e-mail: soumyabrata.dev@ucd.ie).

Fabrizio Orlandi and Declan O'Sullivan are with the ADAPT SFI Research Centre, School of Computer Science and Statistics, Trinity College Dublin, D02 PN40 Dublin, Ireland.

Digital Object Identifier 10.1109/JSTARS.2024.3370389

[1][Online]. Available: https://www.noaa.gov/

[2][Online]. Available: https://www2.purpleair.com/

TABLE I
LIST OF DOMAIN ONTOLOGIES FOR ESN MODELING

Ontology	W3C standards[a]	Subject area	Shared by relevant ESN studies
SOSA/SSN [41], [42] (https://www.w3.org/TR/vocab-ssn/)	✓	Sensor networks, observations, sampling methods	Weather data publication [43], [44], marine data management [45]
SWEET [46] (http://sweetontology.net/sweetAll)	ESIPFed (https://www.esipfed.org/)	Thesaurus of earth sciences	Generic earth sciences [47], hydrogeology [48], pedology [49]
SEAS (https://ci.mines-stetienne.fr/seas/index.html)	W3ID (https://w3id.org/)	Smart energy, smart meter, energy statistics	Building Management System [50], power systems [51], global energy [52]
QUDT (http://qudt.org/2.1/schema/qudt)	QUDT.org (http://www.qudt.org/)	Units of measure, quantity kinds, dimensions and data Types.	Weather data publication [43], urban sustainability [53], marine geoinformatics [54], health sensing [55]
OWL-Time (https://www.eea.europa.eu/themes/air)	✓	Generic time annotation, inference...	Weather data publication [43]
GeoSPARQL (https://www.ogc.org/standards/geosparql)	OGC (https://www.ogc.org/)	Geographical representation and topology inference	Weather data publication [43], [56], Copernicus big data [57], green energy [58]
RDF Data Cube Vocabulary (https://www.w3.org/TR/vocab-data-cube/)	✓	Statistical data analysis	Weather data publication [59]–[61]

[a] Ontologies compliant with W3C standards are standardized for the usage in Semantic Web (https://www.w3.org/standards/semanticweb/).

data principles [18], as outlined by established best practices for data dissemination. These practices encompass the following recommendations: 1) use uniform resource identifiers (URIs) as names for things; 2) use HTTP URIs so that people can look up those names; 3) when someone looks up a URI, provide useful information; and 4) include links to other URIs so that they can discover more things. In comparison to the widespread adoption of semantic data structuring in social networks and encyclopedias, such as Facebook Graph Search[3] and Google Knowledge Graph,[4] the higher level of interoperability with respect to a broader range of subjects, such as the domain, the features, and the events that may be made accessible and shared via conceptual meanings using KGs, is underexplored in the area of ESNs. However, the most prevalent approaches in this area (as shown in Table I) seek to construct KGs in triplestores via an extract–transform–load (ETL) approach, which directly ingests data from data services and preprocesses them with semantic annotations. The most significant disadvantage of using triplestores in this manner is the extra storage capacity needed for the materialization of the data, which is often much bigger than the original datasets. In addition, an effective synchronization mechanism for on-the-fly data materialization must be implemented for triplestore-based KGs derived from original data providers in order to maintain their currency. This mechanism even demands more efficacy when data providers update large amounts of data frequently (e.g., streaming data) [19].

To circumvent the disadvantages of a triplestore-based approach, an approach has been introduced that uses an SQL database to house the data and enables the virtualization of the KG [also known as virtual knowledge graph (VKG)] by translating SPARQL protocol and RDF query language (SPARQL) [20] queries into SQL queries on the fly at runtime and the return of the data to the consumer as graph data. The semantic annotation is then undertaken during the query translation in accordance with the W3C standard RDB to RDF mapping language (R2RML) [21] uplift mapping language. The VKG approach is gaining popularity in a variety of contexts [22] for delivering OBDA/I; nevertheless, to the best of our knowledge, there are still knowledge gaps about the extent to which a VKG can be comparable to a triplestore-based KG in terms of ESN data integration. Specifically, this article will address the following gaps in the state of the art by: 1) proposing a VKG-based federation approach for the efficient integration of ESN data with less physical data storage cost involved and with the faster processing of tabular data when compared to conventional triplestores and 2) providing an evaluated comparison of the VKG approach with conventional triplestores with respect to the efficiency of accessing semantically integrated ESN data. This will be evaluated according to their applications in some typical ESN time-series data access scenarios.

B. Contributions of This Work

We list the major contributions of this work as follows.

1) We propose a novel VKG-based federation approach for the semantic integration of ESN data using *Ontop* [23] and PostgreSQL [24]. This approach utilizes the expanded climate analysis (CA) ontology and VKG to integrate various ESN data using shared semantics seamlessly. In addition, it leverages the efficient tabular data processing capabilities offered by SQL-type databases.
2) We use *Ontop* to incorporate the linked data principles for ESN data publication, which makes the ESN data more accessible to other SPARQL endpoints on the web by leveraging the RDF data model.

[3][Online]. Available: https://en.wikipedia.org/wiki/Facebook_Graph_Search

[4][Online]. Available: https://en.wikipedia.org/wiki/Google_Knowledge_Graph

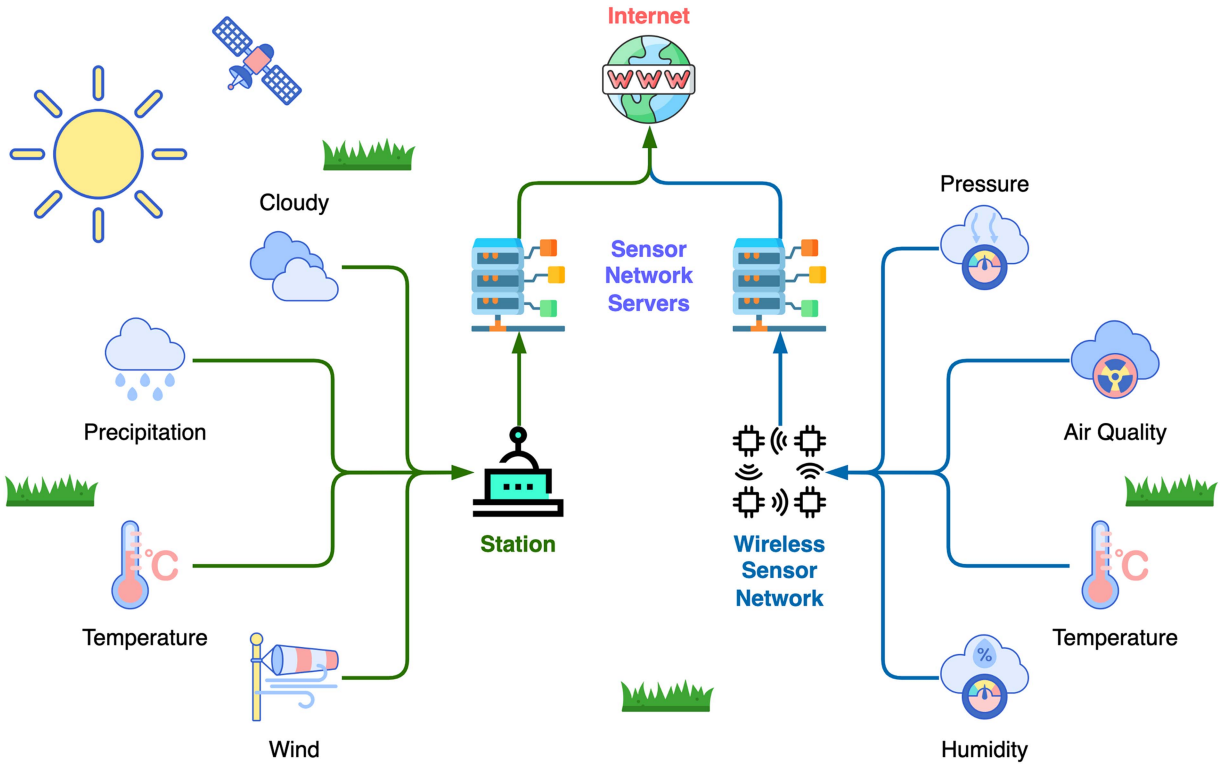


Fig. 1. Schematic diagram of ESNs, with the left half representing station-based ESNs and the right half representing wireless sensor networks.

3) We have evaluated the query processing efficiency of the proposed VKG-based integration approach, specifically for accessing multiple time series. We conclude that it significantly outperforms traditional triplestores in terms of time-stamped alignments, which has not been addressed by previous studies in the area.

The rest of this article[5] is organized as follows. Section II presents the related literature that informs this research. Section III highlights the technical aspects in relation to the *Ontop*-PostgreSQL federated integration approach, including the architectural design, and performance benchmark comparing conventional triplestores for time-series alignment using a set of SPARQL queries. Finally, Section V concludes this article.

II. Related Work

A. Heterogeneous Data From ESNs

We begin by defining an ESN based on a literature discussion. An ESN consists of a collection of sensor nodes and a communication system that transmits their data to a server [25]. The left half of Fig. 1 illustrates the data flow: the sensor nodes (e.g., temperature and precipitation sensors) collect data autonomously and transmit the data to one or more base stations, which then send them to a sensor network server. ESNs can also be wireless-sensor-based networks (see the right half of Fig. 1) and are regarded as a subtype of ESN, in which the base stations are removed from the network to increase the mobility of the sensor nodes [26]. Notably, the station-based ESNs have a significant advantage over today's wireless network in terms of their capacity for long-term monitoring stability. Some of NOAA's National Centers for Environmental Information (NCEI)[6] stations in Ireland have climate data that span over 150 years. This provides ideal long-term data evidence for classic climate classification [27] and trend analysis [28].

An ESN can be categorized as homogeneous or heterogeneous based on whether or not its sensor nodes are identical [29]. However, the ESNs in our research are all heterogeneous (most modern sensor networks are heterogeneous networks [30]), and we have focused primarily on the heterogeneity of the top hierarchy of ESNs—the open heterogeneous environmental sensor data. Fig. 1 is essential to readers because it illustrates how the top-level heterogeneity among different ESNs accumulates as data travel from sensors to servers and are ultimately distributed independently over the web by the top servers (or the data suppliers). The heterogeneity in a single ESN is often managed separately by each ESN prior to data publishing; nevertheless, this may ultimately result in inter-ESN data heterogeneity since there is presently no consensus on protocols or standards for ESN data publication.

B. KG as Semantic Integration Framework

The KG [31] consists of a graph data model or ontology of the domain to conceptualize the things, events, and ideas and

[5]In the spirit of reproducible research, the source code is available at https://zenodo.org/record/8082674

[6][Online]. Available: https://www.ncei.noaa.gov/

how they are connected, composed, and generated in web ontology language (OWL) and/or resource description framework schema (RDFS), as well as the content and other data that are informed by and can be integrated with the aid of this model. All types of data may be merged using a semantic-based KG, and all of that data can be kept alongside the data model(s), with everything, including data models or schemas and instance data, being encoded using the W3C-standard-based RDF representation. Being standard based ensures that the KG can be easily harnessed by applications, making it possible to simplify, avoid duplication, and scale application development beyond organizational boundaries.

Nowadays, KGs have been used in many areas as a powerful way to represent and integrate data. They can be constructed from either unstructured data (e.g., news and reports) using popular natural language processing technologies or structured data (e.g., SQL, JavaScript object notation (JSON), and representational state transfer (REST)ful) using declarative mapping rules such as the W3C-standard-based R2RML language [21]. Recently, many studies have focused on exploring KGs from the perspective of domain knowledge inference [32], [33], [34], [35]. For example, Shiri et al. [35] proposed the "Maritime DeepDive"—a probabilistic KG to describe a set of random variables and how they are correlated for enhancing inference about maritime events. Nevertheless, these studies tend to define their own data graph schemas and build offline KGs. A major limitation of these offline KGs is that data cannot be easily shared across them. In other words, the extensibility of a KG to a broader scope of data for enhanced KG functionality is absent. In contrast, our research examines the application of KGs to ESNs in terms of the improvement of data accessibility. We explore the environmental monitoring area and use KGs to enhance data accessibility. As outlined in Section I, a KG may be materialized in a triplestore or virtualized as a VKG, depending on whether the real data are kept as graph data or in other forms (e.g., relational model and plain tables). Currently, the majority of VKG implementations only natively support SQL database virtualization. For nonrelational data types, such as JSON and eXtensible markup language (XML), an SQL wrapper for data type transition is often required in the technology stack, which can significantly reduce the semantic processing efficiency of VKGs. Prospective VKG implementations are currently undergoing active development to achieve the same level of semantic processing (such as question answering) as triplestores [22]. Even though a VKG has a fraction of a triplestore's semantic processing capabilities, this article demonstrates that it has a considerable advantage over triplestores for ESN time-series data processing.

C. Integrating ESNs for Improved Data Accessibility

The existing ESNs (including wireless sensor networks) built for environmental monitoring are unable to provide dense coverage in terms of geographical locations or environmental parameters, particularly for those that monitor the environment globally with station-based sensor networks, such as NOAA's NCEI. Several NCEI stations are situated in Ireland, and yet the available measurements primarily include only temperature and precipitation. Thanks to the publication of integrated ESN data sources by a number of projects, data consumers can obtain a denser coverage of environmental sensor data all at once. The Royal Netherlands Meteorological Institute (KNMI) Climate Explorer[7], for instance, is a large climate data provider that incorporates several high-quality and reliable ESN data producers, such as NCEI, the European centre for medium-range weather forecasts (ECMWF)[8], the European climate assessment & dataset (ECA&D)[9], and others. To aid users in navigating data sources for analytical purposes, KNMI [36] is developed with a graphical user interface (GUI) that allows users to fill out a form with climate indices of interest and conduct direct analyses on selected data. OpenSensorWeb [37] is another sizable project that aims to provide an all-inclusive ESN data explorer. So far, OpenSensorWeb has recorded data from 44 ESNs with a total of 2 800 000 devices and 72 700 sensors. For severe weather event analysis, it has also been strengthened by the EXTRUSO project [38], which combines remote sensing, geoinformatics, and other techniques. OpenSensorWeb, in contrast to KNMI, offers a more contemporary graphical dashboard with a map view to identify all the gathered sensors and data of environmental elements, such as atmosphere, soil, water, etc.

1) Limitations of Triplestore-Based ESN Integration Frameworks: In order to deliver consistent multisource ESN data access, the aforementioned ESN integration approaches tackle the data heterogeneity issue by designing a new schema to rearrange different data forms from the original data sources. However, this design is often employed only for a system of integrated ESN data sources and typically results in minimal cross-system interoperability. In addition, data consumers must undergo a learning curve (e.g., learning from documentation) whenever they encounter a new system of integrated ESNs due to the lack of information exchange across schemas of various systems. As indicated in Section II-B, due to the shareable schema (also known as an ontology) of a KG, the KG is now widely used as a framework to integrate ESN data while resolving heterogeneity issues at a higher level via the semantic exchange of information across systems. Specifically, we characterize some important ontologies in this field in Table I, which includes information on the subject areas and some example ESNs they link. In state-of-the-art semantic integration frameworks, exemplified by advanced systems such as KARMA [39] and evolving semantic knowledge and aggregation processing engine (ESKAPE) [40], a notable challenge arises wherein the requisite storage of data in a KG incurs substantial costs associated with storage for semantic annotations. In addition, the processing of tabular data, including time-series data, demands considerable computing power.

2) Advantages of Adopting VKGs for ESNs Integration: The majority of relevant studies in Table I employ triplestore-based KGs, which can potentially pose a synchronization problem for KG materialization from dynamic data sources impacting critical requirements on materializing bandwidth and overhead. The emergence of a VKG brings the possibility of eliminating

[7][Online]. Available: https://climexp.knmi.nl/
[8][Online]. Available: https://www.ecmwf.int/
[9][Online]. Available: https://www.ecad.eu/

materialization while sharing the semantics among data of different domains. According to Calvanese's study [23] and the Norwegian Petroleum Directorate[10] benchmark [62], the well-known efficiency of relational databases (i.e., better transaction handling, table locking, input/output (I/O), caching, etc.) can be viewed as the decisive factor to empower a VKG to outperform a triplestore in some use cases other than where advanced SPARQL features are required (e.g., navigation of arbitrary paths through the graph). Nevertheless, to the best of our knowledge, the available literature on VKG applications [63], [64], [65] does not adequately represent the reality of ESN-specific analytical tasks, such as semantic queries to acquire multiple time series at once [66]. By contrast, examining the usage of VKGs to integrate ESN data for increased data accessibility will be the primary emphasis of this work, which will principally address this research gap. Due to the fact that VKG applied to data formats other than the relational models will compromise the performance by adding SQL wrappers on the stack, this work will only focus on the VKG applied to the derived relational databases from a variety of ESN data sources and so retain the benefits of harnessing the schema information during the query translation [23]. For example, integrity constraints, including primary and foreign keys, may be employed to mitigate the size and intricacy of the query within the SPARQL-to-SQL conversion process. This can be achieved, for instance, by eliminating redundant self-joins and identifying conditions that are either unsatisfiable or trivially satisfiable. Notably, this also reveals a fundamental limitation of this strategy, since the transformation of ESN data sources into SQL databases would incur an inevitable data materialization cost. However, compared to the transformation of ESN data sources into triplestore-based KGs as accomplished by frameworks like KARMA [39], employing VKGs can still minimize the semantic annotation cost, which often occupies a significant fraction of the storage space in a triplestore-based KG, and exploit the power of underlying SQL databases. This is also known as data integration through database federation in studies such as those by Haas et al. [67] and Gu et al. [68], but this work relies on the ETL process to effectuate the transformation of data sources into relational databases, thereby facilitating expedited processing of tabular data. We have chosen the *Ontop*-PostgreSQL combination as the VKG unit, as its performance has been lauded in Lanti's work and is comparable to *Ontop* with other relational back ends (e.g., MySQL) [62]. Moreover, we advocate the adoption of PostgreSQL also due to its recognized scalability and the presence of a dynamic and engaged open-source community. We compare the performance of the proposed VKG-based federated integration approach with two popular open-source triplestores: Apache Jena Fuseki[11] and GraphDB's free version[12].

[10][Online]. Available: https://www.npd.no/en/
[11][Online]. Available: https://jena.apache.org/documentation/fuseki2/
[12][Online]. Available: https://graphdb.ontotext.com/

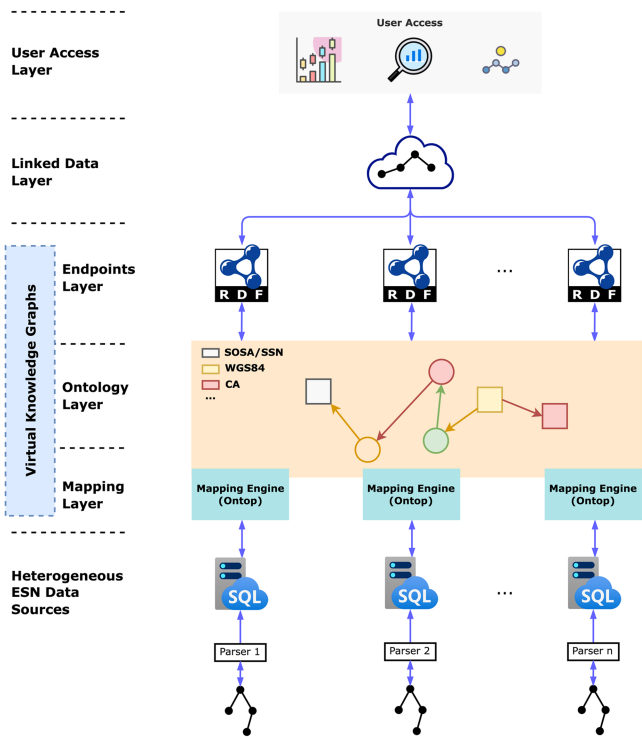


Fig. 2. VKG architecture designed for integrating heterogeneous ESN data.

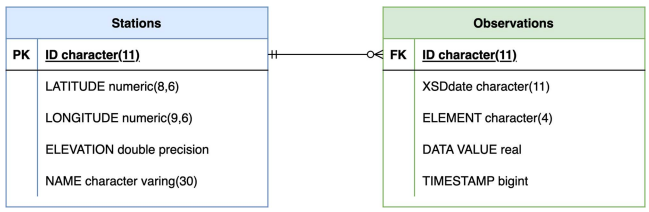


Fig. 3. Minimal relational model for an NOAA's GHCND station (blue) and an observation (green); the numbers in brackets represent the length of the data in bytes.

III. Methodology

A. Proposed VKG Architecture

We introduce the proposed VKG-based federated integration architecture (see Fig. 2) as easy access to various sources of ESN data. One thing to note here is that there is yet no usability development for the approach such as a GUI for practical applications. The proposed approach is evaluated in Section III-B exploring how much power can be exploited from the proposed integration approach in terms of the speed of SPARQL query responses in completing typical time-series data access scenarios in the ESN domain.

1) Heterogeneous ESN Data Sources: Each source of heterogeneous ESN data will be retrieved online and parsed for storage in an SQL database. For example, we modeled NOAA's Global Historical Climatology Network daily (GHCND) data, as seen in Fig. 3. Here, we chose PostgreSQL as the SQL connection to "mapping layer." In this architecture, each ESN data source has a parser and *Ontop* mapping engine since different data sources can hold different protocols for data access.

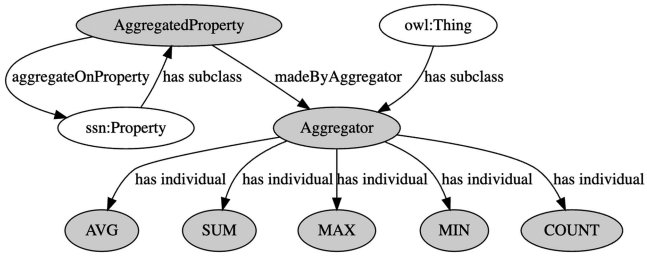


Fig. 4. Representation of aggregative properties in CA ontology; the concepts and the relationships between concepts are denoted by nodes and directed edges, respectively. Terminologies of CA ontologies are in filled nodes. ssn and owl denote terminologies adopted from SSN and OWL ontologies (see Table I), respectively.

2) Mapping Layer: The *Ontop* is used as the mapping engine. The role of the *Ontop* mapping engine is to execute the SPARQL-to-SQL rewriting according to the mapping rules, which can define a VKG by declaring the way of populating the ontological models (see "ontology layer" in Fig. 2) with corresponding SQL tables. Listing 1 provides an example set of *Ontop* mapping declarations for NOAA's data[13]. Importantly, the VKG will not be materialized through the mapping. Instead, the actual data access is finalized by querying the underlying SQL databases (PostgreSQL). It is acknowledged that users must establish distinct schema mapping rules through *Ontop*'s integrated application programming interfaces (APIs) for data sources outside the purview of this study. This process entails creating mapping declarations akin to those illustrated in Listing 1, customized to diverse database management system schema structures. Presently, the mapping definition is dependent on *Ontop* configuration files. However, there is a future aspiration to introduce a GUI-based mapping definition approach, catering to nonexpert users and mitigating the complexity associated with *Ontop* configuration files.

3) Ontology Layer: The schema unifying heterogeneous ESN data sources happens on the "ontology layer," where each *Ontop* VKG applies the same sets of ontological models to various SQL databases. Typically, we reuse the sensor, observation, sample, and actuator/semantic sensor network (SOSA/SSN) ontology (see Table I) and our CA ontology [69] as the major domain ontologies for ESN data modeling. The advantage of using the CA ontology over other relevant domain ontologies is its ontological expressivity for aggregation functions (see Fig. 4), which are often used in ESN data (NOAA's GHCND data use "TMAX" for maximum temperature, for example). Fig. 5 gives an example of ontological modeling of tables in Fig. 3.

4) Endpoint Layer: In the endpoint layer, *Ontop* sets an endpoint for each VKG virtualization from an SQL database. One thing to note is that *Ontop* endpoint does not support SPARQL federated queries, which means that *Ontop* VKGs can provide data access independently but data in each VKG are not accessible between each other. However, the *Ontop* VKGs are mutually accessible through an intermediate linked data triplestore (see Section III-A5) such as Apache Jena Fuseki.

5) Linked Data Layer: The integration of *Ontop* endpoint services is implemented by using a triplestore to assemble all the *Ontop* VKGs as one linked data endpoint in this layer.

6) User Access Layer: On the top layer, there is user access; this system can be designed to supply dialog-based query formulations to allow users to perform online data preprocessing based on semantic queries (SPARQL).

B. Multiple Time-Series Accessibility Benchmark

Multiple ESN time-series data structuring is the major use case examined (with SPARQL queries) in this article. Multiple time-series structuring is an important time-series preprocessing step for enabling end users to swiftly get an initial understanding of historical data. Finely aligned multiple time series in consistent sequence of time are easily usable as multivariate time-series input in further data analyses, such as the correlation analysis and the preparation of the multistep forecasting datasets [66]. A comparable multiple time-series preprocessing platform is provided by the KNMI Climate Explorer.[14] Recapping the semantic interoperability issue discussed in Section I, the key strength of our work over KNMI is that the VKGs are readily available to all the linked data endpoints, allowing users to easily query a larger dataset (sometimes spanning domains) despite schematic heterogeneities. For example, in our earlier work, we complemented the administrative geography of ESN sensors by using data from Wikidata SPARQL endpoints [73].

Due to the rising effort required by end users for multiple time-series acquisition from diverse ESN data sources, semantically integrated ESN data providers need to prioritize SPARQL query processing efficiency to improve multiple time-series data accessibility. The following are two common multiple ESN time-series data access scenarios that are generally regarded by end users in this field for downstream multistep time-series forecasting tasks [66], [74]. The performance comparison of our approach against conventional triplestores w.r.t. for each of the two scenarios was conducted in the following configurations.

1) *Hardware:* Intel Xeon W-1290P (ten cores at 3.7 GHz), 64 GB of RAM, and 1-TB M.2 PCIe SSD.
2) *Software:*

a) *OS:* Ubuntu 18.04 LTS;
b) *Java:* 11.0.13;
c) *SQL database:* PostgreSQL 14.4;
d) *VKG implementation: Ontop* 4.2.1;
e) *Triplestores:* i) Apache Jena Fuseki 4.4.0 (with 4-GB memory of JVM) and ii) Ontotext GraphDB 10 Free Edition (with 4-GB memory of JVM).

3) *Launch type:* Hot start, i.e., benchmarks are conducted when the processes of VKGs and triplestores are already running in the background.
4) *Cost measurement:* Arithmetic mean of the execution time of each query in different platforms; each query is run 20 times to get the mean value.
5) *Queries used per scenario:*

a) property navigation as per class (see Section III-B1 for more details): Listing 2;

[13]Refer to our repository at https://zenodo.org/record/8082674 for the full set of mapping rules.

[14][Online]. Available: http://climexp.knmi.nl/start.cgi

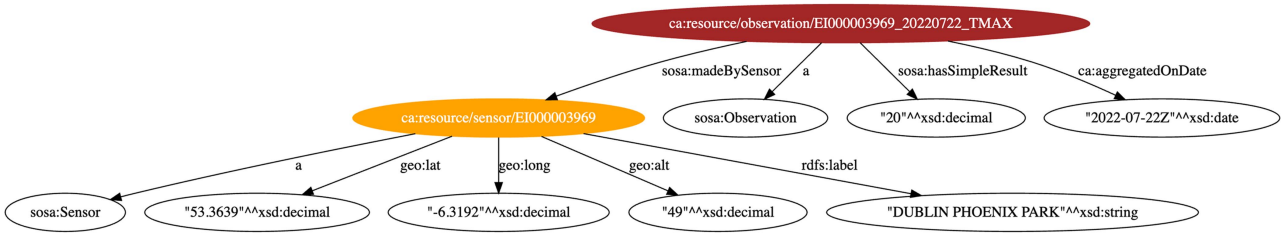


Fig. 5. Minimal ontological models representation of an NOAA's GHCND station (in orange) and an observation (in brown); the concepts and the relationships between concepts are denoted by nodes and directed edges, respectively. sosa, ca, rdfs, and geo denote terminologies adopted from SOSA, CA, RDFS [70], and WGS84 [71] ontologies, respectively. Different literal types (e.g., decimal and string) are defined by XML schema definition (XSD) ontology [72].

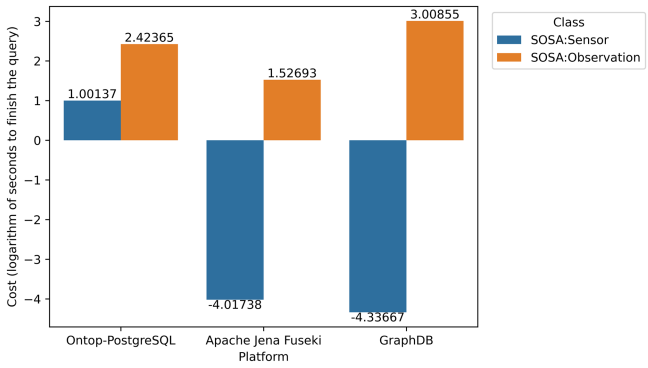


Fig. 6. Performance benchmark for arbitrary path navigation.

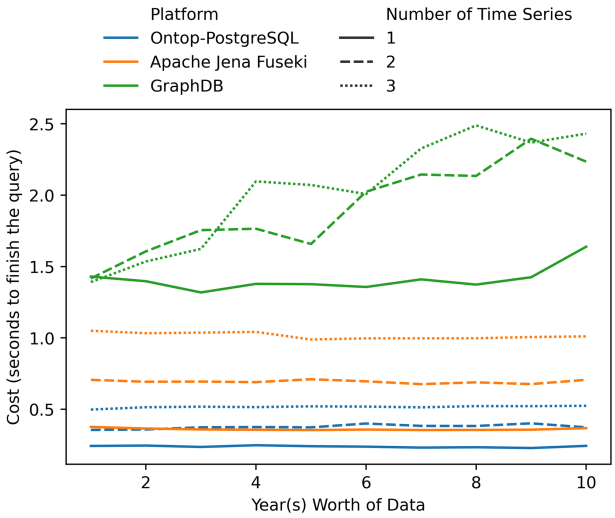


Fig. 7. Performance benchmark for 1, 2, 3 time-series alignment based on known stations.

b) multivariate time-series structuring (see Section III-B2 for more details): Listings 3 and 4.

6) *Dataset:* NOAA's GHCND within the spatial extent of Ireland.[15]

1) Property Navigation as per Class: For domain ESN data consumers, understanding how the ESN data are modeled is a major problem for first-time users ingesting the time-series data. Particularly, in ontological modeling, ontologies may be partially reused and coupled to others in a flexible manner. This may result in a learning curve before being acquainted with the terminology used in ontologies. In our work, `SOSA:Sensor` and `SOSA:Observation` are two primary classes taken from SOSA ontology. However, the properties of an `SOSA:Sensor (Observation)` can be represented by other ontologies. For instance, we use `WGS84`'s `lat/long` as an `SOSA:Sensor`'s latitude/longitude coordinate properties. To ensure how an `SOSA:Sensor (Observation)` is modeled using various ontologies, the query Listing 2 might be made. The logarithmic (base e) query response time comparison between the proposed VKG and native triplestores is shown in Fig. 6. As seen in the figure, the VKG is significantly less efficient than the Apache Jena Fuseki triplestore during the sensors' property search but considerably better than GraphDB during the observations' property search. For any separate database, i.e., a triplestore or an SQL database, the cost of this query is mostly determined by the number of traversed objects. In this experiment, we have collected from NOAA's GHCND all the sensors (in the quantity of 25) and observations (in the quantity of over 1 000 000) in Ireland. This result indicates that GraphDB performs better in the navigation of arbitrary paths while traversing a small number of entities, but its triplestore-based advantage diminishes substantially as the number of traversed entities increases.

2) Multivariate Time-Series Structuring: ESN data analysis makes growing use of multivariate time series that may be obtained by aligning multiple time series that are assumed to be interrelated. During the analysis of environmental data, several environmental factors might occasionally interact to decide the ultimate result. For instance, the air pressure and humidity time series may influence the precipitation time series [66]. To prepare consistent multiple time series simultaneously, one of the difficulties when creating SPARQL queries is unstacking and aligning the different types of observational data to the same sequence of time steps. In this scenario, multiple daily ESN time series are aligned using simple and only SPARQL queries, which can essentially minimize the bandwidth and compute resource demands of procedural programming techniques.

Listing 3 is an example SPARQL query of aligning two daily time series (station `EI000003969` is given explicitly) on the same date sequence of year 2022. The comprehensive performance comparison regarding length of time series and number of time series being aligned is given in Fig. 7. As shown in the figure, generally, the proposed VKG-based federated integration outperforms the other two triplestores by a

[15][Online]. Available: Archived at https://zenodo.org/record/8082674

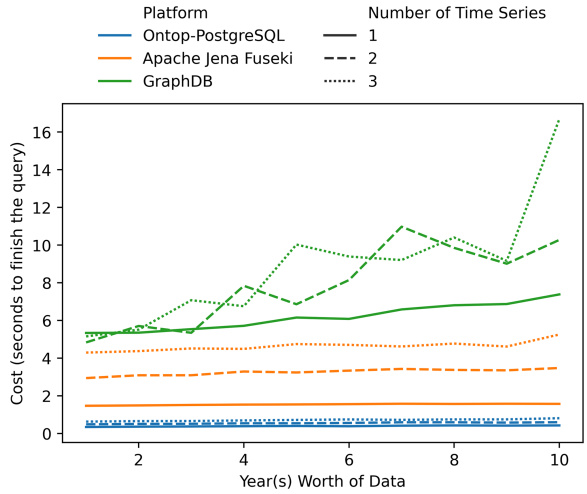


Fig. 8. Performance benchmark for 1, 2, 3 time-series alignment based on unknown stations.

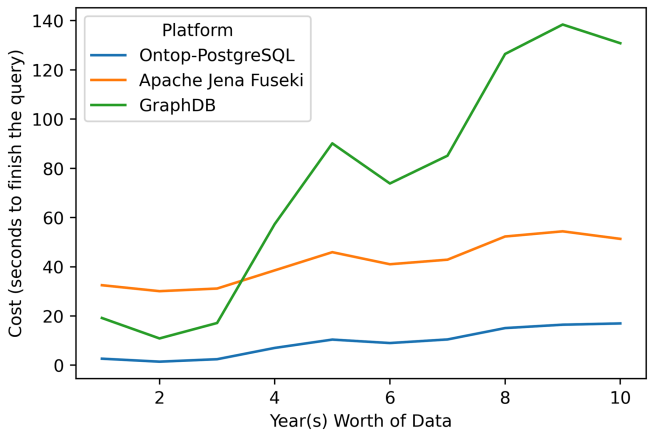


Fig. 9. Performance benchmark for 3 time-series alignment based on 1) unknown stations and 2) optional results.

considerable margin, and query costs rise as the number of time series for alignment increases on all the platforms. However, as time-series length increases, *Ontop* VKG and Fuseki triplestore have constant query costs, but GraphDB has a growing query cost (almost linear for the number of time series over 2). A similar performance discrepancy can be observed in Fig. 8 of which the queries for performance comparison were made by parameterizing `EI000003969` in Listing 3 as a variable being projected onto the `SELECT` clause for stations retrieving.

Prior SPARQL queries have been capable of aligning various time series in the same time order. However, only dates with values for all the aligned time series can be tied to query solutions. In fact, ESN data records are often absent at certain time intervals. Sometimes, ESN data consumers are more interested in time-series alignments with missing values, which are subsequently filled using interpolation techniques. Listing 4 is a more complex variant of Listing 3, which is written by an expert to compare the platforms' performance of time-series alignment with missing values, and the results are shown in Fig. 9. According to the cost curves, the proposed *Ontop* VKG-based federated integration approach requires the least amount of time (but growing with the quantity of data) for time-series alignment in response to the costly SPARQL `OPTIONAL` clause (see Listing 4), whereas the GraphDB triplestore has a much more rapid cost increase as the amount of data increase.

IV. Discussion

In this section, we summarize the findings of this study and discuss some significant points that may inform future prospective studies addressing the data heterogeneity problem of various ESN data sources.

A. Semantic Interoperability

One of the most significant rationales for proposing the VKG-based federated integration approach is the efforts made toward semantic interoperability among diverse ESN data sources. Semantic interoperability seeks to facilitate data sharing across organizations or systems by ensuring a shared understanding and interpretation of data, employing domain concepts, context knowledge, and formal data representation. In a broader context, semantics, which explores the meaning of language, is crucial for fostering mutual understanding among individuals with diverse experiences or perspectives [75]. Interoperability, meanwhile, pertains to the capacity of multiple systems to collaborate seamlessly, irrespective of variations in interfaces, platforms, and adopted technologies [76].

Currently, the most sophisticated and widely used ESN data integration platforms, such as KNMI, are likely to be based on relational databases that expose RESTful APIs for data-consuming purposes. The major limit of such platforms is that they cannot be easily extended to additional data sources unless various heterogeneity problems (e.g., schemas, data models, and platforms) are addressed [77]. By contrast, the proposed VKG-based federated integration adheres to the linked data principles and RDF data model framework for data modeling. From the perspective of schematic heterogeneity, different ESN data sources can share the same "schema (i.e., ontologies)" when being transformed into KGs since semantic models are reusable due to the usage of the RDF data model as the representation backbone. This minimizes the schema unification cost during the data integration, as well as the learning curve of different schemas for data consumers. From the perspective of platform heterogeneity, incorporating linked data principles into the proposed approach provides interoperable access to various ESN data sources situated in different linked data endpoints. In other words, we expose SPARQL endpoints on the web so that the data in the VKG can also be queryable in other KGs using SPARQL queries (e.g., federated queries [78]), as well as the potential to reference in the VKG data in other linked data KGs.

B. Enhanced Time-Series Data Analytics

Semantic interoperability can bring many advantages for time-series data analytics. One of the most notable applications is its use in feature selection for machine learning tasks. Our previous studies [66], [79], [80] have examined triplestore-based KGs in terms of their ability to enable data consumers

to acquire time-stamped features from multiple data sources using SPARQL queries. This considerably reduces the I/O cost and preprocessing efforts during the training dataset preparation. Compared to triplestore-based KGs, the proposed VKG-based federated integration approach can further facilitate time-stamped feature alignment in two additional ways. First, the VKG performs better at aligning time-series data with respect to time stamps. This is due to the effectiveness of VKGs' underlying SQL-based databases (e.g., PostgreSQL) in optimizing tabular data processing. Second, the I/O and storage costs are decreased further due to the absence of semantic annotations of the underlying tabular data. For complex connections in a VKG (e.g., where the properties query in Section III-B1) that cannot be efficiently condensed into tables, the usage of VKGs can be inferior to triplestores. In practice, however, the ESN data may contain additional complex data other than the simple time-stamped structured data. In such a case, data partitioning techniques that divide data into triplestore-based KGs and VKGs in order to enhance the data processing efficiency might be explored. We view this as a prospective research direction for future studies in this area.

C. Limitations in Streaming Sensor Data Handling

The proposed VKG-based federated integration focuses on dealing with a derived SQL database (i.e., PostgreSQL) of various data sources. Though this approach can enable data consumers to consume data in an interoperable manner by using only SPARQL queries and requires less storage than directly using triplestore-based KGs, it still incurs additional I/O and storage costs during the data materialization from original data sources (e.g., NOAA) into PostgreSQL. This limit can be notable in streaming sensor data, posing a challenge of timely data processing. In this case, direct access to various ESN data sources for VKG-based virtualization and the incorporation of RDF stream processing [81] languages (e.g., C-SPARQL [82], and CQELS [83]) into the VKG-based federation integration approach would be an appropriate research direction for future studies in this area.

V. Conclusion and Future Work

The fast expansion of ESNs has caused a severe data heterogeneity problem, resulting in an expense growth of ESN data acquisition from diverse data sources. To enable efficient access to various data sources, we propose a novel VKG-based federation integration approach with *Ontop*-PostgreSQL for ESN data integration based on semantic models. In addition, we incorporate the linked data principles into the proposed approach, allowing other triplestores to easily use the VKGs for ESN data integration and vice versa. Compared to the existing sophisticated relational model-based integration platforms, such as the KNMI Climate Explorer, the proposed VKG-based federation integration approach is schema-less, i.e., independent of the schema isolation across different data sources, and is readily extensible to other KGs that adhere to linked data principles. To evaluate the performance of the data acquisition through semantic queries to the proposed approach, we compared our approach to that of two commonly used triplestores (Apache Jena Fuseki and GraphDB). In particular, the benchmark focused primarily on SPARQL queries in the processing of multiple time-series data access purposes, a subject not addressed by earlier research. Though the VKG architecture performs poorly in property navigation, it has a substantial benefit in multiple time-series alignments, which is expensive in conventional triplestores. In the future, based on the findings of this study, we will conduct data partition research to explore a more efficient semantic ESN data integration architecture that combines the benefits of both conventional triplestores and VKGs.

Appendix

```
[PrefixDeclaration]
:  http://example.org/ontology/
geo:  http://www.w3.org/2003/01/geo/wgs84_pos
    #
rdf:  http://www.w3.org/1999/02/22-rdf-syntax
    -ns#
xsd:  http://www.w3.org/2001/XMLSchema#
rdfs:  http://www.w3.org/2000/01/rdf-schema#
sosa:  http://www.w3.org/ns/sosa/

[MappingDeclaration] @collection [[
mappingId Station
target  :resource/sensor/{ID} a sosa:Sensor ;
     geo:lat {LATITUDE} ; geo:long {LONGITUDE
    } ; geo:alt {ELEVATION} ; rdfs:label {
    NAME}^^xsd:string .
source  SELECT * FROM public."Stations" ;

mappingId Observations
target  :resource/observation/{ID}_{DATE}_{
    ELEMENT} a sosa:Observation ; sosa:
    madeBySensor :resource/sensor/{ID} ; sosa
    :hasSimpleResult {DATA VALUE}^^xsd:
    decimal ; :aggregatedOnDate {XSDdate}^^
    xsd:date ; sosa:observedProperty :
    resource/datatype/{ELEMENT} .
source  SELECT * FROM public."Observations"
]]
```

Listing 1: Part of *Ontop* VKG mapping declarations for an NOAA's GHCND station and an observation.

```
#reused onotologies
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-
    schema#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>

#find properties of a sosa:Sensor (using sosa
    :Observation for the other case)
SELECT DISTINCT ?p {?x a sosa:Sensor; ?p ?o.}
```

Listing 2: Query to investigate how an SOSA:Sensor (SOSA:Observation) is modeled in our KGs.

```
PREFIX ca: <http://example.org/ontology/>
BASE <http://example.org/ontology/>
PREFIX ca_type: <http://example.org/ontology/
    resource/datatype/>
PREFIX sosa: <http://www.w3.org/ns/sosa/>

SELECT ?prcp ?tmax ?date WHERE { #2
  ?prcp_obs sosa:madeBySensor <resource/
      sensor/EI000003969>; sosa:
      hasSimpleResult ?prcp; sosa:
      observedProperty ca_type:PRCP; ca:
      aggregatedOnDate ?date .
    {
     SELECT ?tmax ?date WHERE{ #1
     ?tmax_obs sosa:madeBySensor <resource/
         sensor/EI000003969>; sosa:
         hasSimpleResult ?tmax; sosa:
         observedProperty ca_type:TMAX; ca:
         aggregatedOnDate ?date .
     FILTER(YEAR(?date)=2022)
     } ORDER BY ?date
    }
}
```

Listing 3: Two daily time series alignment.

```
#reused onotologies
PREFIX ca: <http://example.org/ontology/>
PREFIX ca_type: <http://example.org/ontology/
    resource/datatype/>
PREFIX sosa: <http://www.w3.org/ns/sosa/>

SELECT ?sta (MAX(?prcp) as ?prcp) (MAX(?tmin)
     as ?tmin) (MAX(?tmax) as ?tmax) ?date
    WHERE { #3
  ?prcp_obs sosa:madeBySensor ?sta; ca:
      aggregatedOnDate ?date .
  OPTIONAL {?prcp_obs sosa:hasSimpleResult ?
      prcp ; sosa:observedProperty ca_type:
      PRCP .}
  {
    SELECT ?sta ?tmin ?tmax ?date WHERE{ #2
       ?tmin_obs sosa:madeBySensor ?sta; ca:
           aggregatedOnDate ?date .
       OPTIONAL {?tmin_obs sosa:
           hasSimpleResult ?tmin ; sosa:
           observedProperty ca_type:TMIN .}
    {
     SELECT ?sta ?tmax ?date WHERE{ #1
     ?tmax_obs sosa:madeBySensor ?sta; ca:
         aggregatedOnDate ?date.
       OPTIONAL {?tmax_obs sosa:
           hasSimpleResult ?tmax ; sosa:
           observedProperty ca_type:TMAX .}
       FILTER(YEAR(?date)=2018)
       }ORDER BY ?date
     }}ORDER BY ?date
  }}
GROUP BY ?sta ?date
ORDER BY ?sta ?date
```

Listing 4: Time series alignment with missing values.

REFERENCES

[1] G. Fortino, W. Russo, C. Savaglio, W. Shen, and M. Zhou, "Agent-oriented cooperative smart objects: From IoT system design to implementation," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 48, no. 11, pp. 1939–1956, Nov. 2018.

[2] G. Fortino, C. Savaglio, G. Spezzano, and M. Zhou, "Internet of Things as system of systems: A review of methodologies, frameworks, platforms, and tools," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 51, no. 1, pp. 223–236, Jan. 2021.

[3] M. Jain, T. Alskaif, and S. Dev, "Validating clustering frameworks for electric load demand profiles," *IEEE Trans. Ind. Inform.*, vol. 17, no. 12, pp. 8057–8065, Dec. 2021.

[4] M. Jain, T. AlSkaif, and S. Dev, "A clustering framework for residential electric demand profiles," in *Proc. Int. Conf. Smart Energy Syst. Technol.*, 2020, pp. 1–6.

[5] T. AlSkaif, S. Dev, L. Visser, M. Hossari, and W. van Sark, "A systematic analysis of meteorological variables for PV output power estimation," *Renewable Energy*, vol. 153, pp. 12–22, 2020.

[6] S. Manandhar, S. Dev, Y. H. Lee, Y. S. Meng, and S. Winkler, "A data-driven approach for accurate rainfall prediction," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9323–9331, Nov. 2019.

[7] J. Wu, F. Orlandi, D. O'Sullivan, and S. Dev, "A workflow to convert live atmospheric sensor data into linked data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 4086–4089.

[8] J. Wu, F. Orlandi, D. O'Sullivan, and S. Dev, "Publishing climate data as linked data via virtual knowledge graphs," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 4090–4093.

[9] A. Poppick, J. Nardi, N. Feldman, A. H. Baker, A. Pinard, and D. M. Hammerling, "A statistical analysis of lossily compressed climate model data," *Comput. Geosci.*, vol. 145, Dec. 2020, Art. no. 104599.

[10] V. Beretta, J.-C. Desconnets, I. Mougenot, M. Arslan, J. Barde, and V. Chaffard, "A user-centric metadata model to foster sharing and reuse of multidisciplinary datasets in environmental and life sciences," *Comput. Geosci.*, vol. 154, Sep. 2021, Art. no. 104807.

[11] W. Litwin and A. Abdellatif, "Multidatabase interoperability," *Computer*, vol. 19, pp. 10–18, Dec. 1986.

[12] J. Wu, H. Chen, F. Orlandi, Y. H. Lee, D. O'Sullivan, and S. Dev, "Automated climate analyses using knowledge graph," in *Proc. IEEE USNC-URSI Radio Sci. Meeting*, 2021, pp. 106–107.

[13] J. Wu, F. Orlandi, M. S. Pathan, D. O'Sullivan, and S. Dev, "Augmenting weather sensor data with remote knowledge graphs," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 1264–1267.

[14] A. Tolk and J. A. Muguira, "The levels of conceptual interoperability model," in *Proc. Fall Simul. Interoperability Workshop*, 2003, pp. 1–11.

[15] G. Xiao et al., "Ontology-based data access: A survey," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 5511–5519.

[16] A. A. Salatino, T. Thanapalasingam, A. Mannocci, F. Osborne, and E. Motta, "The computer science ontology: A large-scale taxonomy of research areas," in *Proc. 17th Int. Semantic Web Conf.*, 2018, pp. 187–205.

[17] F. Manola et al., "RDF primer," *W3C Recommendation*, vol. 10, no. 1–107, p. 6, 2004.

[18] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data: The story so far," in *Semantic Services, Interoperability and Web Applications: Emerging Concepts*. Hershey, PA, USA: IGI Global, 2011, pp. 205–227.

[19] O. Corcho, F. Priyatna, and D. Chaves-Fraga, "Towards a new generation of ontology based data access," *Semantic Web*, vol. 11, no. 1, pp. 153–160, Jan. 2020.

[20] "SPARQL 1.1 query language." Accessed: Oct. 5, 2022. [Online]. Available: https://www.w3.org/TR/sparql11-query/

[21] "R2RML: RDB to RDF mapping language." Accessed: Oct. 5, 2022. [Online]. Available: https://www.w3.org/TR/r2rml/

[22] G. Xiao, L. Ding, B. Cogrel, and D. Calvanese, "Virtual knowledge graphs: An overview of systems and use cases," *Data Intell.*, vol. 1, pp. 201–223, 2019.

[23] D. Calvanese et al., "Ontop: Answering SPARQL queries over relational databases," *Semantic Web*, vol. 8, no. 3, pp. 471–487, 2017.

[24] B. Momjian, *PostgreSQL: Introduction and Concepts*, vol. 192. New York, NY, USA: Addison-Wesley, 2001.

[25] J. K. Hart and K. Martinez, "Environmental sensor networks: A revolution in the earth system science?," *Earth-Sci. Rev.*, vol. 78, no. 3, pp. 177–191, Oct. 2006.

[26] D. Kandris, C. Nakas, D. Vomvas, and G. Koulouras, "Applications of wireless sensor networks: An up-to-date survey," *Appl. Syst. Innov.*, vol. 3, no. 1, Feb. 2020, Art. no. 14.

[27] S. Khan and M.-U. Hasan, "Climate classification of Pakistan," *Int. J. Econ. Environ. Geol.*, vol. 10, no. 2, pp. 60–71, Sep. 2019.

[28] M. Mudelsee, "Trend analysis of climate time series: A review of methods," *Earth-Sci. Rev.*, vol. 190, pp. 310–322, Mar. 2019.

[29] V. Mhatre and C. Rosenberg, "Homogeneous vs heterogeneous clustered sensor networks: A comparative study," in *Proc. IEEE Int. Conf. Commun.*, 2004, pp. 3646–3651.

[30] S. Kolavennu, "Process control and diagnostics over wireless sensor networks," in *Industrial Wireless Sensor Networks*, R. Budampati and S. Kolavennu, Eds., Sawston, U.K.: Woodhead Publishing, Jan. 2016, pp. 39–55.

[31] A. Hogan et al., "Knowledge graphs," 2020, *arXiv:2003.02320*.

[32] B. Liu, R. Song, Y. Xiang, J. Du, W. Ruan, and J. Hu, "Self-supervised entity alignment based on multi-modal contrastive learning," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 11, pp. 2031–2033, Nov. 2022.

[33] L. Yang et al., "Collective entity alignment for knowledge fusion of power grid dispatching knowledge graphs," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 11, pp. 1990–2004, Nov. 2022.

[34] J. Lü, G. Wen, R. Lu, Y. Wang, and S. Zhang, "Networked knowledge and complex networks: An engineering view," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 8, pp. 1366–1383, Aug. 2022.

[35] F. Shiri et al., "Toward the automated construction of probabilistic knowledge graphs for the maritime domain," in *Proc. IEEE 24th Int. Conf. Inf. Fusion*, 2021, pp. 1–8.

[36] V. Trouet and G. J. Van Oldenborgh, "KNMI climate explorer: A web-based research tool for high-resolution paleoclimatology," *Tree-Ring Res.*, vol. 69, no. 1, pp. 3–13, Jan. 2013.

[37] "OPEN ENSOR WEB—Environmental data at your fingertips." Accessed: Aug. 30, 2022. [Online]. Available: https://www.opensensorweb.de/

[38] S. Wiemann et al., "On the monitoring and prediction of flash floods in small and medium-sized catchments—The EXTRUSO project," in *Proc. EGU Gen. Assem. Conf. Abstr.*, 2017, Art. no. 4862.

[39] S. Gupta, P. Szekely, C. A. Knoblock, A. Goel, M. Taheriyan, and M. Muslea, "Karma: A system for mapping structured sources into the semantic web," in *Proc. Extended Semantic Web Conf.*, 2012, pp. 430–434.

[40] A. Pomp, A. Paulus, S. Jeschke, and T. Meisen, "ESKAPE: Information platform for enabling semantic data processing," in *Proc. 19th Int. Conf. Enterprise Inf. Syst.*, 2017, pp. 644–655.

[41] K. Janowicz, A. Haller, S. J. D. Cox, D. Le Phuoc, and M. Lefrançois, "SOSA: A lightweight ontology for sensors, observations, samples, and actuators," *J. Web Semantics*, vol. 56, pp. 1–10, May 2019.

[42] M. Compton et al., "The SSN ontology of the W3C semantic sensor network incubator group," *J. Web Semantics*, vol. 17, pp. 25–32, 2012.

[43] R. Catherine, B. Stephan, A. Geraldine, and B. Daniel, "Weather data publication on the LOD using SOSA/SSN ontology," *Semantic Web*, vol. 11, pp. 581–591, 2019.

[44] H. N. M. Quoca, H. N. Quoca, M. Hauswirtha, and D. L. Phuoca, "Global weather sensor dataset." Accessed: Mar. 7, 2024. [Online]. Available: http://www.semantic-web-journal.net/content/global-weather-sensor-dataset

[45] M. Zárate, G. Braun, M. Lewis, and P. Fillottrani, "Observational/hydrographic data of the south atlantic ocean published as LOD," *Semantic Web*, pp. 1–12, Mar. 2021.

[46] R. Raskin, "Guide to sweet ontologies," NASA/Jet Propulsion Lab, Pasadena, CA, USA, 2006. Accessed: May 2011. [Online]. Available: http://sweet.jpl.nasa.gov/guide.doc

[47] N. DiGiuseppe, L. C. Pouchard, and N. F. Noy, "Sweet ontology coverage for earth system sciences," *Earth Sci. Informat.*, vol. 7, pp. 249–264, 2014.

[48] A. Tripathi and H. A. Babaie, "Developing a modular hydrogeology ontology by extending the sweet upper-level ontologies," *Comput. Geosci.*, vol. 34, no. 9, pp. 1022–1033, 2008.

[49] H. Du et al., "An ontology of soil properties and processes," in *Proc. 15th Int. Semantic Web Conf.*, 2016, pp. 30–37.

[50] I. Esnaola-Gonzalez, J. Bermúdez, I. Fernandez, and A. Arnaiz, "Ontologies for observations and actuations in buildings: A survey," *Semantic Web*, vol. 11, no. 4, pp. 593–621, 2020.

[51] G. Santos, T. Pinto, Z. Vale, and J. M. Corchado, "Semantic interoperability for multiagent simulation and decision support in power systems," in *Highlights in Practical Applications of Agents, Multi-Agent Systems, and Social Good. The PAAMS Collection*. New York, NY, USA: Springer, 2021, pp. 215–226.

[52] J. Cuenca, F. Larrinaga, and E. Curry, "DABGEO: A reusable and usable global energy ontology for the energy domain," *J. Web Semantics*, vol. 61/62, Mar. 2020, Art. no. 100550.

[53] C. Kuster, J.-L. Hippolyte, and Y. Rezgui, "The UDSA ontology: An ontology to support real time urban sustainability assessment," *Adv. Eng. Softw.*, vol. 140, Feb. 2020, Art. no. 102731.

[54] Y. Lassoued and A. Leadbetter, "Ontologies and their contribution to marine and coastal geoinformatics interoperability," in *Geoinformatics for Marine and Coastal Management*. Boca Raton, FL, USA: CRC Press, 2016, pp. 159–178.

[55] M. Hennessy, C. Oentojo, and S. Ray, "A framework and ontology for mobile sensor platforms in home health management," in *Proc. 1st Int. Workshop Eng. Mobile-Enabled Syst.*, 2013, pp. 31–35.

[56] N. Y. Ayadi, C. Faron, F. Michel, F. Gandon, and O. Corby, "WeaKG-MF: A knowledge graph of observational weather data," in *Proc. 19th Extended Semantic Web Conf.*, 2022, pp. 101–106.

[57] K. Bereta et al., "From Copernicus Big Data to big information and big knowledge: A demo from the Copernicus App Lab Project," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 1911–1914.

[58] A. Ishmanov and I. Alikhojayev, "Spatial data consolidation for decision support in the field of green energy," in *Proc. Int. Conf. Inf. Technol. Manage.*, 2015, pp. 90–91.

[59] L. Lefort et al., "The ACORN-SAT linked climate dataset," *Semantic Web*, vol. 8, pp. 959–967, 2017.

[60] *Use Cases and Lessons for the Data Cube Vocabulary*, Cambridge, MA, USA: World Wide Web Consortium, , 2013.

[61] L. M. Vilches-Blázquez, B. Villazón-Terrazas, O. Corcho, and A. Gómez-Pérez, "Integrating geographical information in the linked digital earth," *Int. J. Digit. Earth*, vol. 7, no. 7, pp. 554–575, Aug. 2014.

[62] D. Lanti, M. I. Rezk, G. Xiao, and D. Calvanese, "The NPD benchmark: Reality check for OBDA systems," in *Proc. 18th Int. Conf. Extending Database Technol.*, 2015, pp. 617–628.

[63] D. Chaves-Fraga, F. Priyatna, A. Cimmino, J. Toledo, E. Ruckhaus, and O. Corcho, "GTFS-Madrid-bench: A benchmark for virtual knowledge graph access in the transport domain," *J. Web Semantics*, vol. 65, 2020, Art. no. 100596.

[64] Y. Seonwoo, S. Yoon, F. Dernoncourt, T. Bui, and A. Oh, "Virtual knowledge graph construction for zero-shot domain-specific document retrieval," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 1169–1178.

[65] M. Arslan and C. Cruz, "Modeling virtual knowledge graphs using relevant news data by NLP methods for business analysis," in *Proc. 17th Int. Conf. Emerg. Technol.*, 2022, pp. 172–177.

[66] J. Wu, F. Orlandi, D. O'Sullivan, E. Pisoni, and S. Dev, "Boosting climate analysis with semantically uplifted knowledge graphs," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4708–4718, 2022.

[67] L. M. Haas, E. T. Lin, and M. A. Roth, "Data integration through database federation," *IBM Syst. J.*, vol. 41, no. 4, pp. 578–596, 2002.

[68] Z. Gu et al., "A systematic overview of data federation systems," *Semantic Web*, vol. 15, no. 1, pp. 107–165, 2024.

[69] J. Wu, F. Orlandi, D. O'Sullivan, and S. Dev, "An ontology model for climatic data analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 5739–5742.

[70] B. McBride, "The resource description framework (RDF) and its vocabulary description language RDFs," in *Handbook on Ontologies*. New York, NY, USA: Springer, 2004, pp. 51–65.

[71] D. Brickley, "W3C semantic web interest group: Basic geo (WGS84 lat/long) vocabulary," W3C Working Group Note, 2004.

[72] "XML schema datatypes in RDF and OWL." Accessed: Jan. 29, 2023. [Online]. Available: https://www.w3.org/TR/swbp-xsch-datatypes/

[73] J. Wu, O. Fabrizio, D. O'Sullivan, and S. Dev, *Link Climate: An Interoperable Knowledge Graph Platform for Climate Data*. New York, NY, USA: Elsevier, 2022.

[74] D. Sahoo, N. Sood, U. Rani, G. Abraham, V. Dutt, and A. Dileep, "Comparative analysis of multi-step time-series forecasting for network load dataset," in *Proc. 11th Int. Conf. Comput., Commun. Netw. Technol.*, 2020, pp. 1–7.

[75] P. Alexopoulos, *Semantic Modeling for Data*. Sebastopol, CA, USA: O'Reilly Media, Inc., Aug. 2020.

[76] P. Wegner, "Interoperability," *ACM Comput. Surv.*, vol. 28, no. 1, pp. 285–287, Mar. 1996.

[77] A. M. Ouksel and A. Sheth, "Semantic interoperability in global information systems," *SIGMOD Rec.*, vol. 28, no. 1, pp. 5–12, Mar. 1999.

[78] C. Buil-Aranda, M. Arenas, O. Corcho, and A. Polleres, "Federating queries in SPARQL 1.1: Syntax, semantics and evaluation," *J. Web Semantics*, vol. 18, no. 1, pp. 1–17, Jan. 2013.

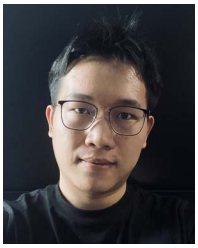
[79] J. Wu, J. Pierse, F. Orlandi, D. O'Sullivan, and S. Dev, "Improving tourism analytics from climate data using knowledge graphs," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2402–2412, 2023.

[80] J. Wu, F. Orlandi, D. O'Sullivan, and S. Dev, "Detecting rainfall events leveraging climate knowledge graphs," in *Proc. Photon. Electromagn. Res. Symp.*, 2021, pp. 2336–2341.

[81] D. Dell'Aglio, E. D. Valle, J.-P. Calbimonte, and O. Corcho, "RSP-QL semantics: A unifying query model to explain heterogeneity of RDF stream processing systems," *Int. J. Semantic Web Inf. Syst.*, vol. 10, no. 4, pp. 17–44, 2014.

[82] D. F. Barbieri, D. Braga, S. Ceri, E. Della Valle, and M. Grossniklaus, "C-SPARQL: SPARQL for continuous querying," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 1061–1062.

[83] D. Le Phuoc, M. Dao-Tran, A. Le Tuan, M. N. Duc, and M. Hauswirth, "RDF stream processing with CQELS framework for real-time analysis," in *Proc. 9th ACM Int. Conf. Distrib. Event-Based Syst.*, 2015, pp. 285–292.

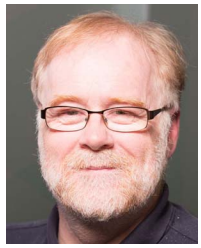


Jiantao Wu received the M.Sc. degree in materials for energy and environment from University College London, London, U.K., in 2017. He is currently working toward the Ph.D. degree in computer science under the joint supervision of Prof. Soumyabrata Dev with University College Dublin, Dublin, Ireland, and Dr. Fabrizio Orlandi and Prof. Declan O'Sullivan with Trinity College Dublin, Dublin.

He was a Software Engineer with China Electronics Technology Group Corporation, Beijing, China, from 2017 to 2019. His research interests include knowledge graphs, machine learning, and sensor data processing.

Fabrizio Orlandi received the B.Eng. and M.Eng. degrees in computer engineering from the University of Modena and Reggio Emilia, Modena, Italy, in 2005 and 2008, respectively and the Ph.D. degree in computer science from the University of Galway, Galway, Ireland, in 2013.

He is a Senior Knowledge and Data Engineer with Inter IKEA Systems, Delft, The Netherlands . He is also Visiting Research Fellow with the ADAPT Research Centre, Trinity College Dublin, Dublin, Ireland. Prior to this, he was a Marie Skłodowska-Curie EDGE Fellow with Trinity College Dublin. Prior to joining ADAPT—at Fraunhofer IAIS in Germany—he was a coordinator of the EU H2020 OpenBudgets.eu project and contributed to several large European and industry research projects, such as BigDataOcean.eu and SLIPO.eu. In his research areas, he has experience on foundational and applied research on both EU-funded and industry projects. His research interests include knowledge graphs, linked open data, knowledge representation, and the application of semantic technologies to different domains, such as social media, cultural heritage, law, and open data.



Declan O'Sullivan received the B.A., M.Sc., and Ph.D. degrees in computer science from Trinity College Dublin, Dublin, Ireland, in 1985, 1988, and 2006, respectively.

He is a Professor of Computer Science with the School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland, where he is also a co-applicant Principal Investigator with the ADAPT SFI Research Centre. He has authored more than 260 scientific peer-reviewed papers and international journals. He is a Member of three journal editorial boards and has undertaken more than 12 chair roles in IEEE and IFIP conferences over the years. He has won competitive research funding as principal investigator and co-principal investigator of approximately 7.8 million euros. Funding has been won across a range of funding programs: European Commission (H2020 and Marie Curie), Science Foundation Ireland (FAME, CNGL, and ADAPT), HEA PRTLI (NEMBES and TGI), and from industry: Huawei, Accenture, Ericsson, Nokia Bell Labs, Ordnance Survey Ireland, and Central Statistics Office. He was elected as a Fellow in Trinity College Dublin in 2019 in recognition for the quality of his contributions.

Soumyabrata Dev (Member, IEEE) received the B.Tech. degree (*summa cum laude*) in electronics and communication from the National Institute of Technology Silchar, Silchar, India, in 2010, and the Ph.D. degree in electrical and electronics engineering from Nanyang Technological University, Singapore, in 2017.

He is currently an Assistant Professor with the School of Computer Science, University College Dublin, Dublin, Ireland, and an SFI-Funded Investigator with the ADAPT SFI Research Centre, Dublin. In 2015, he was a Visiting Student with the Audiovisual Communication Laboratory, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. He has authored more than 120 publications in leading journals and conferences. His research interests include remote sensing, statistical image processing, machine learning, and deep learning.