LinkedGeoClimate: An Interoperable Platform for Climate Data Access Within Geographical Context

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

***Abstract*—Climate data (e.g., air temperature and precipitation) are used in a variety of environmental sectors, such as determining which crops to cultivate for farmlands, and optimizing the placement of products in retail stores. Currently, access to climate data is popularly managed by sophisticated database management systems, which are efficient for data processing but have limited data exchangeability across systems. By contrast, many semantic studies focus on the use of RDF knowledge graphs (KGs) for climate data access, which is semantically interoperable such that data can be easily shared between different RDF KGs based on predefined domain ontologies. However, climate data is often consumed in a certain geographical context to understand its relationships with other environmental sectors. For example, the geographical context of farmyards is needed to determine which climate stations nearby are used. The interoperability proposed for the geographical context of climate data access is under-explored by relevant semantic studies, resulting in additional resource waste in integrating heterogeneous geospatial information for climate data access. In this article, we propose LinkedGeoClimate, which is an interoperable RDF KGs platform for climate data access within an enriched geographical context. LinkedGeoClimate provides the necessary geographical and geospatial information for climate data access and further advances interoperable climate data access when mutual spatial relationships with other environmental sectors are concerned.**

***Index Terms*—Climate data, GeoSPARQL, geographical data, knowledge graphs (KGs), RDF.**

I. Introduction

TODAY'S climate data accessibility has been substantially facilitated by successful application of database management system (DBMS) techniques during the retrieval, processing, and distribution of climate data. Especially, relational DBMS (RDBMS) such as MySQL and PostgreSQL, is now widely used by various climate data distributors (e.g., GHCND [1], UKMO [2]). However, RDBMS systems present fragmentary climate data on the Web since the relational models are not designed for reuse purposes. A typical downside can be seen from today's climate data integration: data exposed from RDBMS (e.g., using RESTful APIs) needs significant efforts (e.g., defining a new relational model) to collate the data heterogeneity in terms of naming conventions, data formatting, and other protocols, resulting in increased data complexity and non-interoperable databases, where data cannot be shared with each other [3]. For instance, consider the KNMI Climate Explorer,[1] a valuable online tool designed for the analysis and visualization of climate data. While it offers access to an extensive array of climate datasets, each of these datasets is independently published and exhibits distinct data structures and formats. To tackle the issue of data heterogeneity, KNMI has undertaken a proactive initiative, maintaining a project repository [4]. This project involves the utilization of customized scripts to procure various datasets from the Web. Subsequently, these datasets are subjected to a unification process within the platform, involving the imposition of a new data schema for structured organization. This approach enables the consolidation of data from diverse sources, rendering it accessible via a single platform. However, from the perspective of data consumers, query protocols directed toward various platforms can exhibit substantial disparities. This divergence underscores the overarching challenge of data heterogeneity, significantly complicating the process of accessing and utilizing climate data for scientific analysis and informed decision-making. Furthermore, climate data can often be analyzed together with data sources from other domains [5], for example, to understand the implications of climate change, especially for a reciprocal relationship with systems concerned with geographical features, such as land, soil, and agriculture. Using DBMSs for climate data distribution cannot provide consistent accessibility where climate data is needed for external data and vice versa.

The use of RDF knowledge graphs (KGs) to uplift heterogeneous climate data is a current trend in research [6], [7] to achieve interoperability. In contrast to RDBMS, RDF [8], [9] is used as the data model standardized with the World Wide Web Consortium (W3C) for data interchange. A key advantage of using RDF to build KGs is that it allows multiple schemas to be applied, interconnected, queried as one and modified without altering the data instances in a KG [10]. The schema applied in a KG is also known as the "Ontology" and can be defined independently of the KG databases; in other words, the ontologies are reusable rather than specific to the databases. Taking into account the modeling of observational climate data (e.g., temperature) in

Manuscript received 15 May 2023; revised 26 August 2023 and 15 November 2023; accepted 3 April 2024. Date of publication 22 April 2024; date of current version 30 May 2024. This work was supported by Science Foundation Ireland under Grant 13/RC/2106_P2 at the ADAPT SFI Research Centre of University College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, was supported by Science Foundation Ireland through the SFI Research Centres Programme. *(Corresponding author: Soumyabrata Dev.)*

Jiantao Wu and Soumyabrata Dev are with the ADAPT SFI Research Centre, School of Computer Science, University College Dublin, D04 C1P1 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.2024.3391922

[1] [Online]. Available: http://climexp.knmi.nl

KGs, different data sources can adopt the same "SOSA" ontology [11] (`https://www.w3.org/TR/vocab-ssn/`) to build the taxonomy of observational data. By reusing the ontologies, the climate data in disparate KGs are exchangeable according to the ontological expressivity without the database isolation that occurs in the RDBMS. However, climate data in relevant RDF KGs are more likely to be uplifted primarily by W3C recommended ontologies (e.g., SOSA/SSN ontology) that are generic and have a limited variety to cover the semantics for more analytical data usage [12], [13], [14] from aspects of environmental sectors.

In reality, environmental measurements can be associated with a geospatial extent, including air temperature observations made by stations, farming productions of farmyards, and soil types of land regions, where the stations, farmyards, and land regions imply the geographical contexts for these measurements. When using data-driven analytical methods to understand the mutual relationships between climate and other environmental sectors, climatic measurements (e.g., air temperature) are often linked to these systems according to some spatial constraints applied to geographical contexts. For example, the nearest weather station of a farmyard can be used to approximate the local weather of the farmyard. Land cover may be attached to weather stations to improve weather forecast quality [15], [16] or air temperature estimation based on remote sensing images [17]. This research underscores the necessity of incorporating a geographical context into climate data for effective climate–environment analytics. Notably, our investigation reveals that the incorporation of such geographical context remains relatively under-explored within existing RDF KGs. In response, our study introduces LinkedGeoClimate, an RDF KG platform meticulously crafted to address this gap. The primary aim of LinkedGeoClimate is to establish a seamlessly interoperable geographical context for accessing climate data. By achieving this objective, our platform facilitates the cohesive integration and consumption of climate data in conjunction with geographical data, all within the ambit of spatial conditions.

A. Contributions

The contributions of our work are listed as follows.

1) We extend the climate analysis (CA) ontology [18] for semantic annotations on the geographical metadata of climate data sources to improve geo-contextual access to climate data.
2) We create RDF KGs of geographical vector data from OpenStreetMap (OSM) [19], CORINE [20] land cover map, and EPA soil map [21]. These endpoints are open for enriching the geographical information of climate datasets.
3) We propose using a triple store to manage the data integration and access based on geospatial relationships. In particular, we use correlated queries to provide efficient processing of complex geospatial constraints for climate data access when other environmental sectors are involved. For example, finding the closest climate/weather stations of OSM farmyards.
4) We present an initial scalability assessment (Section III-D) conducted on LinkedGeoClimate, focusing on the evaluation of its performance in handling correlated queries for the consumption of geo-contextual climate data. The findings substantiate the expedited execution of spatial queries through the utilization of PostGIS databases empowered with spatial indexing. It is pertinent to regard these outcomes as a foundational reference point to inform forthcoming refinements aimed at enhancing the efficiency of LinkedGeoClimate.
5) We released LinkedGeoClimate on Github[2] for future iterations of development.

B. Structure of the Article

The rest of this article is organized as follows: Section II discusses related studies on addressing climate data accessibility. We will emphasize the lessons learned and the aspects in which we advance the state-of-the-art. In Section III, we illustrate the details of the proposed ontological modeling approach. In Section IV, we discuss the value of this approach, as well as the limitations that challenge the studies in this area. Finally, Section V concludes this article.

II. Related Work

We discuss two main kinds of approaches during the storyline of addressing the multisource climate data accessibility: 1) RDBMS-based approaches and 2) KG-based approaches. Although RDBMS has an obvious database isolation problem, it is still currently the most popular technique used in practice to build climate data integration platforms. There are many lessons that can be learned from existing RDBMS-based platforms. For example, the efforts made to justify and collate the necessary metadata of different climate data sources. For KG-based approaches, there have been many previous ontological modeling studies, but they have been limited to the general representation of observational data. These approaches have limited concern for the geographical context while providing climate data access.

A. RDBMS-Based Approaches

In these approaches, the relational models are not reusable as mentioned in the Section I. To integrate heterogeneous climate data from RDBMS systems, data warehousing frameworks are now popularly studied to achieve uniform access to multisource climate data. Due to the fact that the data warehousing framework is technique heavy, including approaches relating to more than data accessibility, such as parallel processing, here we only characterize how these frameworks address the data exchange of multiple data sources.

One kind of data warehousing framework is implemented through a central data service (e.g. cloud platform) in which data are migrated from various data sources through the extract-transform-load [22] approach. A local schema is often defined as per the application to ensure that datasets are connected

[2]In the spirit of reproducible research, all the source code is available at https://github.com/futaoo/LinkedGeoClimate.

as a whole for querying. The European Climate Assessment & Dataset [23] (ECA&D) is a well-known Web database of multisource climate data that has been successfully made available for use. ECA&D has negotiated with 82 participants (e.g., GHCNd [1], UKMO [2]) to build their own MySQL RDBMS of daily climate station data from 65 countries. ECA&D has a structured metadata template formulated based on the 82 climate data sources gathered, including geographical location, land use, surface coverage, etc. However, these metadata have not been sufficiently exploited for more technical applications. Especially, for queries that involve geospatial calculations, such as "I want maximum temperature series available in the weather station nearest to Dublin Phoenix Park.", and "I want all precipitation data from stations within 10 kilometers of Dublin Airport.", ECA&D is incapable of giving direct access to data under certain geospatial constraints.

Data warehousing can also be performed virtually through a middleware capable of distributing requests of a client query to the actual data sources. Grid is an important proposal by Foster et al.[24] for collaborative data integration. DBMSs (including relational ones) of different organizations can be bundled as a virtual organization in the Grid. The data flow between DBMSs and clients is managed by a middleware, the Open Grid Services Architecture (OGSA [25]), which defines the Grid standards that should be observed for creating virtual organizations. OGSA heavily focuses on the metadata of each data source to provide so-called "metadata-driven" [26] data access. In particular, OGSA-DAI [27] and OGSA-DQP [28] (more advanced in distributed query planning) are essential components in the Grid to exploit the metadata of each DBMS to provide consistent data access. The proposal of OGSA successfully enables consistent data transport between clients and multiple groups in the Grid, but it still needs further interoperability for handling the data sharing between different groups that use diverse policies, protocols, resource descriptions, etc., in their DBMSs.

B. KG-Based Approaches

Today, the RDF KG is an increasingly popular integration framework for improving the accessibility of climate data [29]. Compared to conventional RDBMSs, data (including metadata) in a KG can be exchanged with other KGs due to the exclusive usage of the RDF data model [30]. The schema of a KG is often called "Ontology," which represents data in terms of their relationships corresponding to human knowledge. In the climate domain, ontological modeling is one of the key research directions to address the data accessibility problem.

In previous work, the authors of this article [18] published the ontology "CA," providing a semantic backbone for the construction of general climate KGs. Quoc et al. used the W3C-recommended ontologies (e.g., SSN ontology [31]) to model and publish the global GHCND dataset as linked data. The key idea of these studies is to model domain data using the interoperable RDF model to share the underlying information of the data. This can significantly contribute to easy access to heterogeneous data sources even across domains. For example, data schemas (i.e., ontologies) can be reused by various data sources, allowing the same semantic query to get data from multiple sources. However, the design of ontological access to climate data should consider not only the quantity of data sources, but also what domain knowledge is necessary for applications, especially in today's cross-domain climate data analytics environment. We find that most of the KGs [32], [33], [34], [35] developed for climate data access are heavily based on the standardized (per W3C recommendations) SOSA/SSN [11] ontology, which can help improve general climate data accessibility, but few provide the geographical context of multisource climate data for cross-domain climatic studies such as learning the impacts of climate change on other environmental sectors. The geographical context of climate data access is the main focus of this study, which differs from other studies that examine more general aspects.

C. Relevant Platforms

In this section, we present a list of contemporary platforms proposed to facilitate access to climate data, highlighting their distinguishing features. ECA&D [23] and GHCNd [1] emerge as notable RDBMS-based solutions for the management of climate data. The efficacy of data utilization on such platforms is intricately linked to the functionality of their exposed RESTful APIs. However, it is important to acknowledge that the scope of spatial operations is confined to administrative regions. In addition, the extendability of these platforms to encompass diverse databases, including geographic cartography (e.g., OSM), to facilitate advanced queries spanning multifarious environmental domains is rather limited.

In the domain of KG-based platforms, the graph of things (GoT) [36] has been conceptualized as an integrative mechanism for an array of data sources encompassing NOAA, Camera, Flight, Ship, Twitter, among others. An innovative approach is pursued by converting all data into tailored triplestores, effectively indexed through Elasticsearch (https://www.elastic.co/) and OpenTSDB (http://opentsdb.net/) to accommodate spatio-temporal inquiries. It is pertinent to acknowledge that the geospatial queries within GoT are realized through bespoke functions that deviate from the contemporary GeoSPARQL standards, potentially constraining interoperability. Moreover, it is noteworthy that GoT's visualization lacks integration with a cartographic service, thereby precluding the interactive alignment of data with maps to offer users more intuitive spatial manipulations.

Recent research endeavors have prominently embraced the SOSA/SSN ontology for representing observational data within KGs, coupled with the utilization of GeoSPARQL-enabled triplestores for KG storage. For instance, OceanGraph [37] aptly employs the SOSA/SSN ontology to delineate the process of aggregating information from stationary oceanographic stations, further leveraging GraphDB^{TM}(https://graphdb.ontotext.com/) for triple storage, endowed with inherent GeoSPARQL functionality. WeKG-MF [35] materializes as a triplestore-centric KG tailored to meteorological observations furnished by Météo-France (https://meteofrance.com/). It exhibits an extension of the SOSA/SSN ontology to encapsulate specific meteorological attributes, embracing GeoSPARQL vocabularies to encapsulate

TABLE I
COMPARISON BETWEEN LINKEDGEOCLIMATE AND RELEVANT PLATFORMS

Platforms	Interoperability			Geographical context				Visualization		
	Data can be equally identified across databases/KGs	Data join is allowed across databases/KGs	Data is queryable across databases/KGs	Climate data access based on geospatial constraints	Land features, *e.g.*, land cover, soil type	Places of interest, *e.g.*, OpenStreetMap	Administrative regions	Interactive geospatial operations	Geographical data visualization	GeoSPARQL compatibility
ECA&D [23]†	-	-	-	●	-	◐	◐	●	-	-
GHCNd [1]†	-	-	-	●	-	-	◐	●	-	-
GoT [36]	●	●	●	●	-	-	●	●	●	-
ACORN-SAT [40]	●	●	●	●	-	-	-	●	-	-
WeKG-MF [35]	●	●	●	●	-	-	-	-	-	-
OceanGraph [41]	●	●	●	●	-	-	●	-	-	-
KnowWhereGraph [38]	●	●	●	●	-	-	●	●	●	-
LinkClimate [32]	●	●	●	●	◐	-	-	●	-	-
LinkedGeoClimate*	●	●	●	●	●	●	●	●	●	●

● = satisfied; ◐ = partially satisfied; - = not satisfied;
†RDBMS-based infrastructure; * infrastructure proposed in this work.

the geospatial attributes of stations. KnowWhereGraph [38] constitutes a pivotal environmental KG platform that probes the tenability of environmental events (e.g., extreme weather), delving into facets such as spatial characteristics, historical context, and comparative analyses of specific regions. To this effect, KnowWhereGraph introduces the utilization of a discrete global grid [39], colloquially termed the "S2 Grid System," to index geospatial entities through a collection of S2 cells. This indexing technique employs a tradeoff strategy, balancing data precision against access speed at scale. Notably, KnowWhereGraph integrates a user-friendly interface to facilitate data exploration and furnishes APIs primed for assimilation into geographic information system (GIS) analytical platforms, such as ArcGIS, thereby enhancing the accessibility of GIS queries for nonexperts.

By contrast, the essence of LinkedGeoClimate underscores the enrichment of geospatial constraints pivotal for the comprehensive assimilation of geo-contextual climate data. Furthermore, LinkedGeoClimate embarks on an exploration of SPARQL federation mechanisms to establish connections between climate data and geospatial data sources, encompassing elements like points of interest, land cover, and soil type. This diverges from conventional platforms where geospatial relationships are primarily ascertainable within the integrated datasets confined within a KG. Consequently, the distinctiveness of LinkedGeoClimate lies in its extended reach, fostering climate data interoperability with diverse geospatial data sources while concurrently investigating the efficacy of federated querying mechanisms across multifarious KGs from heterogeneous origins.

For comparison, we define a matrix consisting of criteria that are particularly focused on in this work, namely interoperability, geographical context, and visualization. The interoperability here is defined as the ability to share data across databases/KGs. The geographical context represents how many sources of geographical data can be used for contextual access to climate. For visualization, we focus on the ability to display geographical data, the allowance of interactive geospatial operations, and the compatibility with the GeoSPARQL standards. Details of the results are given in Table I.

III. METHODOLOGY

In this section, we illustrate the construction of the LinkedGeoClimate platform to improve the accessibility of climate data in a geographical context. A bottom-up overview of the platform architecture is given in Fig. 1. The construction process follows a general order of ontological modeling, and then the creation of RDF KGs. In particular, the geographical context (dashed area in Fig. 1) can be used to enrich climate metadata or impose geospatial constraints to climate data access. The enriched geographical metadata can be used to define the local environment of the climate data. Geospatial constraints are typically useful when climate data is consumed for researching the reciprocal relationship between climate and other environmental sectors. In terms of the technology stack, the proposed platform uses virtual knowledge graphs for data storage and a triple store for data enrichment and access management. The detailed proposal of this work is given in the following sections.

A. Ontological Modeling of Climate Metadata

We first examine the metadata structure template used in the ECA&D project (see Section II-A) to design the ontological model for the representation of climate metadata. We use ECA&D because the metadata template has been validated throughout 82 different climate sources, making it a suitable

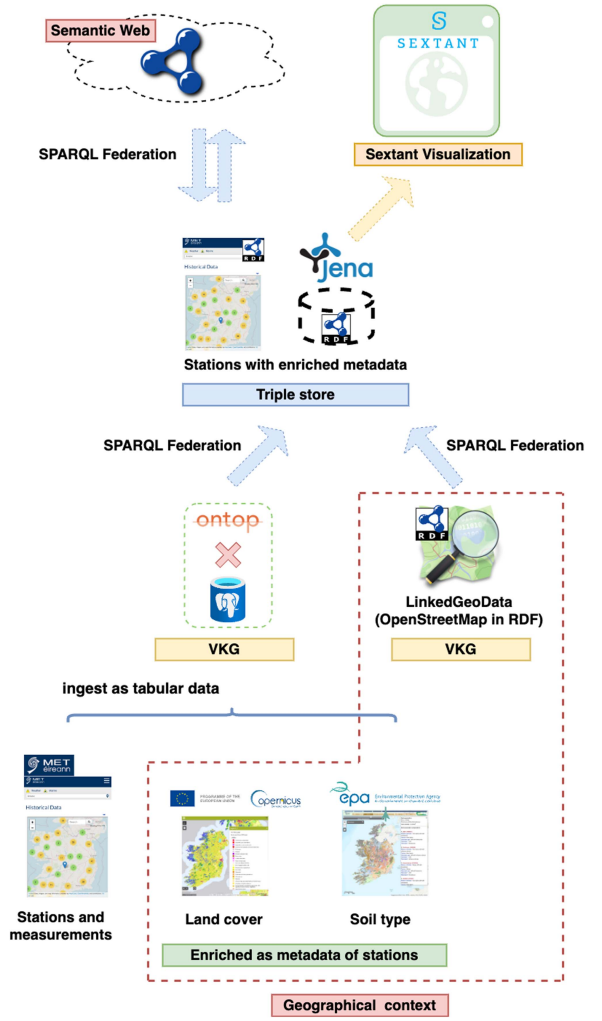


Fig. 1. Overview of the RDF KG platform.

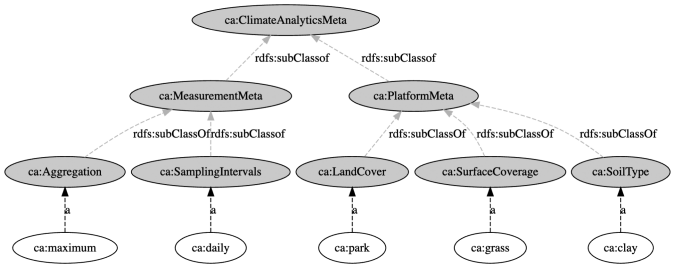


Fig. 2. Ontological representation of observation metadata.

reference for the creation of our ontological model. An example of the metadata encoded for "Dublin Phoenix Park" in the ECA&D project is given in Listing 1. The parameters collected for the station include name, country, geographical coordinates, land use, soil type, surface coverage, and history, which are particularly related to geographical environments. In Fig. 2, our ontological model categorizes these parameters as `ca:PlatformMeta` to describe the metadata of the platform that generates the climate measurements. For brevity, Fig. 2 lists

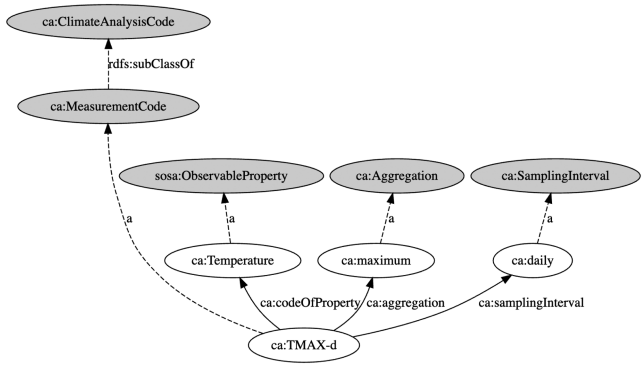


Fig. 3. Daily maximum temperature compiled in a code.

only a subset of ontological representations of these parameters, namely, `ca:LandUse`, `ca:SurfaceCoverage`, and `ca:SoilType`.

```
-----------------------------------------------------------------
121,DUBLIN PHOENIX PARK,IRELAND,+53:21:50,-006:19:10,49 m
-----------------------------------------------------------------
* Land use:
Private area within large public urban park. Surrounds: trees from SE - S - W - N.
      Low buildings in other directions.
* Soil type:
Clay / Grey-brown podzolics predominant in the region
* Surface coverage:
Grass
* History:
There were station moves in 1842, 1855, 1936 and 1976. However, the post 1976
      location is central in that the previous locations were all within about 100
      metres of this location. The land in the area is quite flat. <br>During the
      second half of 1975 a building was constructed about 15m from the instrument
      screen. A new instrument screen about 32m from nearest corner of building
      became operational from 1 Jan 1976.
```

Listing 1. Metadata of weather station – Dublin Phoenix Park.

Measurements also have metadata including the observable property (e.g., temperature), aggregation function (e.g., maximum), and sampling intervals (e.g., daily). We categorize these parameters as `ca:MeasurementMeta` in Fig. 2. These metadata are summarized from the files available on the ECA&D website but will be presented in our ontological model with a clearer hierarchical structure. For example, in Fig. 3, we compiled `ca:Temperature`, `ca:maximum`, and `ca:daily` into a `ca:MeasurementCode` code `ca:TMAX-d` to compactly denote these parameters. A `ca:MeasurementCode` can be attached to a climatic measurement in a way similar to using the W3C standardized `sosa:ObservableProperty` but contains more information, i.e., aggregation function, and sampling interval, for general analytical purposes. Depending on the scope of the CA ontology application, `ca:MeasurementCode` can be expanded with subclasses to differentiate between instantaneous measurements and aggregated measurements. This study focuses primarily on aggregated measurements that are more specific to climate analysis, such as climate change. Currently, the minimum sampling interval is currently up to daily, which is ideal for climate change analysis based on the 27 core indices [42] proposed by the CCl/CLIVAR/JCOMM Expert Panel on Climate Change Detection and Indices [43].

The proposed ontological modeling of metadata for climate datasets demonstrates to data consumers the configuration of

domain parameters. Data consumers can obtain more domain-specific knowledge (e.g., aggregation methods and platform environment) from the implemented RDF KGs using the standard query language—SPARQL [44]. However, the availability of domain parameters can pose a big problem, as not all climate sources have all of the parameters that our ontological model specifies. Thanks to the powerful underlying RDF triple stores of KGs, SPARQL naturally supports federated queries across multiple RDF KGs. This affords us the ideal opportunity to incorporate external datasets to enrich the parameter sets for climate analytics. We investigate the data sources presented in the ECA&D platform and find that most data sources have clearly indicated the measurement metadata on the aggregation methods, observed properties, and sampling intervals, but the platform metadata is often incomplete. Because most of the platform data focus on geographical information, which is a key to be used as geographical constraints for climate data acquisition and climate analytics, we further propose using the OGC GeoSPARQL standards [45] to model the geographical parameters of platforms and taking advantage of the Semantic Web to enrich the geographical metadata modeling for platforms. More details on the modeling and enrichment of geographical metadata are presented in Sections III-B and III-C, respectively.

B. Geospatial Representation Based on GeoSPARQL

The OGC GeoSPARQL standard extends the SPARQL query language to process geospatial data. It specifies a set of vocabulary for the representation of geospatial data in RDF. The semantic meanings of the ontological terms in the GeoSPARQL vocabulary can be implemented in a triple store so that qualitative geospatial reasoning and quantitative geospatial computations can be done using SPARQL queries. Moreover, recapping the OGSA-DAI services in Section II-A, one of the essential principles in the design of the OGSA Grid is *Avoid unnecessary data movement:* wherever possible move the computation to the data. Using GeoSPARQL queries is naturally in favor of this principle in that the computing resources for geospatial reasoning and computations in a KG come from the host triple store. This can reduce the effort of data consumers to do the calculations themselves, such as their redirecting of geographical coordinate data into programmable pipelines.

The geographical coordinates of a platform determine the geometry of the platform, such as a point, or polygon. Fig. 4 presents the geographical part of the ontological modeling of the "Dublin Phoenix Park" weather station. To link a platform with its geometry, we add a term `ca:hasPointGeometry` in the CA ontology. `ca:hasPointGeometry` is often used when the geometry of platforms does not have to be distinguished for climate analytics. This has been adopted by most climate data sources. The terms belonging to OGC GeoSPARQL start with the prefix `geo:` and are in red in the figure. We adopt the Well Known Text (WKT) as one of the GeoSPARQL-validated serializations for geometric values. The point geometry of the station in Fig. 4 is then encoded as `"Point(-6.34972 52.36361)"'eos:wktLiteral`. After being encoded as WKT strings, the topological

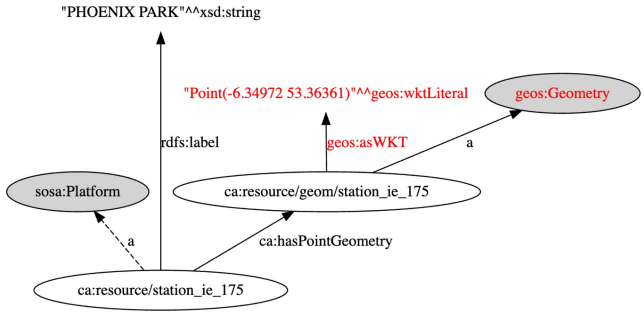


Fig. 4. Ontological representation of geographical features for platforms based on GeoSPARQL standards.

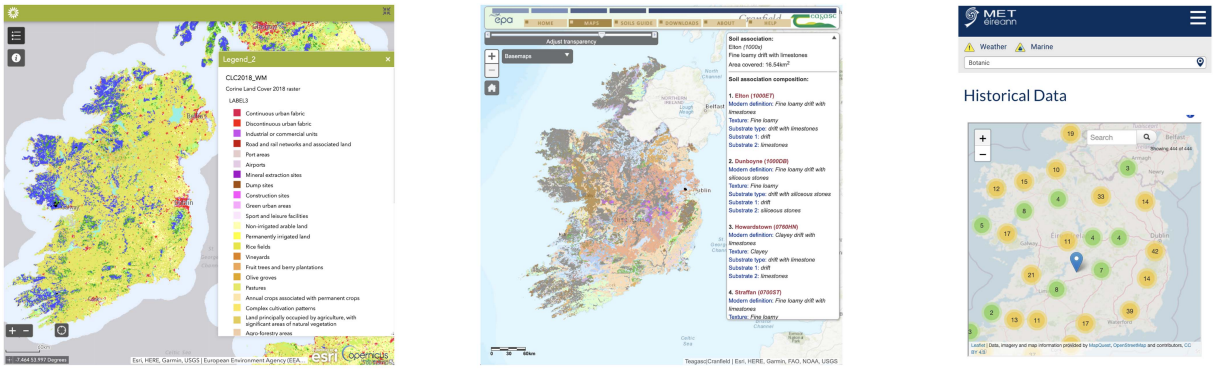
relationships between spatial objects can be queried in a KG using SPARQL queries. Furthermore, notice that we are also enabled to query other datasets over the powerful interoperable Semantic Web consisting of diverse RDF KGs. By running GeoSPARQL queries across multiple KGs, we will demonstrate how RDF KGs can be used to increase the completeness of the geographical metadata of the climate data in Section III-C.

C. Geographical Metadata Enrichment for Stations in RDF

In this section, we demonstrate the detailed proposal for enriching geographical metadata using RDF KGs. We create RDF KGs based on the following datasets (previews are given in Fig. 5): a) 2018 CORINE land cover map, b) 2014 EPA soil map, and c) Met Éireann [46] climate data. In particular, these datasets are up to date and capable of reflecting the full spatial extent of Ireland.

1) Choice Between RDF Virtualization and Triple Stores: We discuss two approaches that can be adopted to implement RDF KGs for climate and geographical datasets: RDF virtualization (*a.k.a.* virtual RDF KGs), and triple stores. The proposal of RDF virtualization over RDBMS takes advantage of RDBMS in the fast tabular data processing and meanwhile provides interoperable ontological access to data. Due to the nature of RDBMS, virtual RDF KGs inherently enable direct ingestion of homogeneous tables, making them suitable for RDF virtualization of frequently updated data streams or massive data dumps. The major limit of RDF virtualization is the inability to ingest schema-less RDF data, since the underlying data keep the relational schemas, i.e., data join from other RDF KGs via SPARQL is inefficient. For triple stores, semantic RDF annotations of large data dumps request a significant amount of storage and I/O bandwidth. The power of triple stores lies in their ability to perform schema-less RDF data join from external RDF KGs via SPARQL federated queries.

In the context of climate data integration, we propose the use of RDF virtualization for climate data, including stations, measurements, and third-party geographical datasets. Distinguished from other studies, we still maintain a triple store for a copy of materialized RDF data of stations from the virtual RDF KGs, as well as any possible external RDF data joined with stations. This architectural framework offers distinct advantages beyond



(a) (b) (c)

Fig. 5. Previews of datasets; (a) and (b) are available as vector data in `.shp` format that contain the multipolygons of the land features while (c) contains stations and their observations in `.csv` format. (a) 2018 CORINE land cover map [47]. (b) 2014 EPA soil map [48]. (c) Met Éireann climatic stations [49].

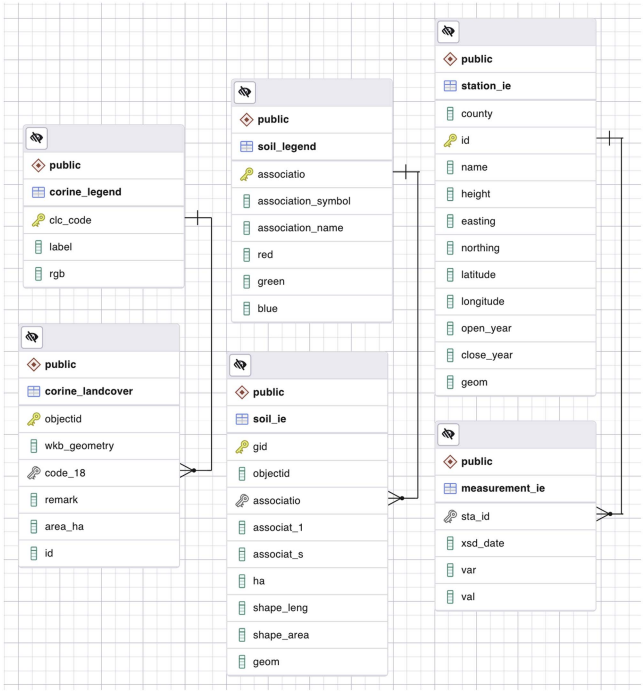


Fig. 6. ERD of the proposed PostgreSQL database.

the exclusive utilization of a triplestore, RDF virtualization, or RDBMS. Its primary advantage, in contrast to the sole reliance on RDBMS or RDF virtualization, lies in its capacity to facilitate novel knowledge discovery within the context of the Semantic Web [49] (e.g., SPARQL federation for geo data illustrated in Section III-C4), encompassing diverse Linked Data endpoints. Furthermore, compared to the exclusive deployment of a triplestore, it demonstrates heightened efficiency in terms of tabular data storage and processing, owing to its underlying RDBMS infrastructure.

2) Virtual RDF KGs Construction: We upload a data dump of the CORINE land cover map, the EPA soil map, and Met Éireann climate data into a PostgreSQL database. The Entity Relationship Diagram of the database is given in Fig. 6. Different data sources are in different independent table groups. We then use the popular Protégé [50] and Ontop [51] bundle for ontological modeling and RDF virtualization, respectively. We connect Ontop with the PostgresSQL database and expose the relational tables as RDF KGs. In Ontop, SPARQL queries are transformed into SQL queries, to be executed in the PostgreSQL database, according to a set of predefined declarative mapping rules. An example of SPARQL-to-SQL mapping for the stations' geometries is given in Listing 2, where `target` is the RDF statements template, and `source` is the source relations to be retrieved.

```
[PrefixDeclaration]
ca:  http://example.org/CA/
geos:  http://www.opengis.net/ont/geosparql#
...

[MappingDeclaration] @collection [[
mappingId geom-station
target  ca:resource/geom/station_ie_{id} a
    geos:Geometry ; geos:asWKT {wkt}^^geos:
    wktLiteral .
source  SELECT id, ST_AsText(geom) AS wkt
    from public.station_ie

...
]]
```

Listing 2. SPARQL-to-SQL mapping of station geometry.

3) Triple Store Construction: We store an RDF data copy of stations in a triple store to make the climate datasets extendable to any external RDF KGs. The initialization of the RDF statements of the stations in the triple store can be consistently generated using Ontop materialization according to the predefined mapping rules (Listing 2) or simply by using SPARQL federated queries executed upon virtual RDF KGs. We choose the Fuseki triple store due to it being based on the open-source Apache Jena (https://jena.apache.org) stack and being extensible by adding additional features. We configure the vanilla Fuseki to support geospatial queries based on OGC GeoSPARQL standards, and correlated subquery [52], i.e., the outer query is evaluated before the inner queries for nested queries. Importantly, we see that the correlated subquery is a key enhancement to SPARQL federated query for geographical data enrichment in our work.

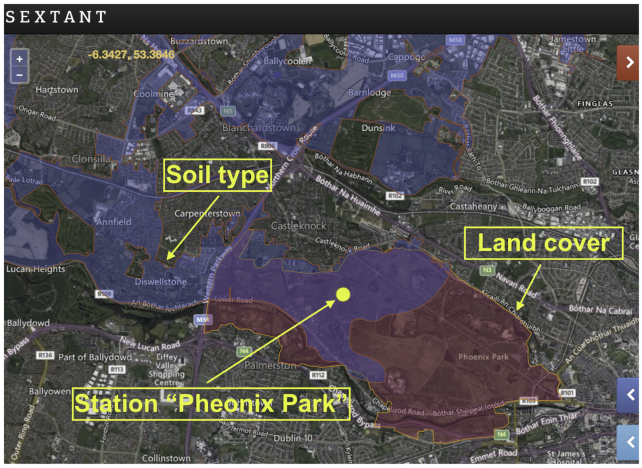


Fig. 7. Visualization of land cover and soil type of station "Phoenix Park" in Sextant; multipolygons of land cover, soil type, and point of the station are presented in red, purple, and yellow, respectively.

We will demonstrate this in more detail in Sections III-C4 and III-C6. To the best of our knowledge, this research is the first to explore the correlated subquery in SPARQL in enriching climate data integration. For graph management, we use named graphs to store data enriched from different sources for different purposes. For example, we store the copy of stations in a named graph <http://example.org/station>, enriched metadata from remote RDF KGs in a named graph <http://example.org/station/more-meta>. The named graphs can be further classified to keep the provenance of the enriched data. For query convenience, the union of all named graphs is used as the default graph for the SPARQL query.

4) Enriching Land Cover and Soil Type as Metadata for Stations: Land cover, soil type, and station data are managed in a SQL database (see Section III-C2), land cover and soil type enrichment for stations can be performed with RDBMS operations by creating related tables/views to be virtualized as part of the geographical RDF KG. However, the use of a DBMS will result in a portion of the enriched data (i.e., land cover and soil type) being maintained separately by the DBMS. In contrast, the triple store has access to both local and external RDF KGs. Hence, we always choose SPARQL federated queries from our triple store for data enrichment, such that the supplemented data is managed consistently by the triple store.

In this work, the land cover and soil type of a station is asserted by the CORINE land cover and the EPA soil data of which the multipolygons spatially contain the point location of the station. A graphical example created by the Sextant visualization technique [53] is shown in Fig. 7. We will explain more about the role of Sextant in this work in Section III-E. Listing 3 is a SPARQL federated query that links to our virtual RDF KGs to implement an assertion for the station "Phoenix Park (encoded in `car:station_ie_175`)". Because geographical coordinates are modeled in GeoSPARQL standards, spatial relationships between station points and land cover and soil type multipolygons can be examined with the GeoSPARQL

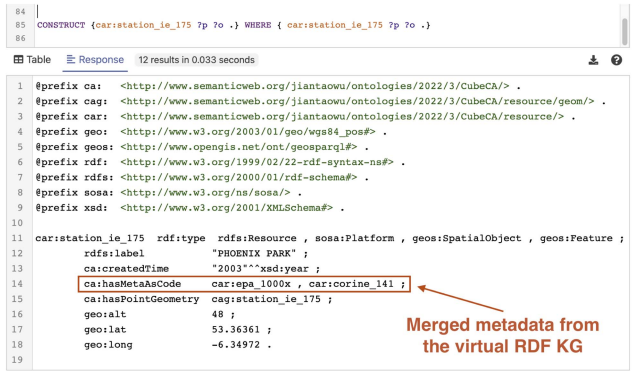


Fig. 8. Enriched metadata of land cover and soil type for station "Phoenix Park" after performing Listing 3.

function `geof:sfWithin` in Fuseki to determine if a station is spatially contained in an area labeled with CORINE land cover or EPA soil type. In addition, the asserted land cover and soil type of the stations are managed in the named graph `http://example.org/station/more-meta` to be distinguished from the default station metadata. The enriched metadata of land cover and soil type for station "Phoenix Park" is shown in Fig. 8.

```
PREFIX lgdo: <http://linkedgeodata.org/
    ontology/>
PREFIX spatialF: <http://jena.apache.org/
    function/spatial#>
PREFIX unit: <http://www.opengis.net/def/uom/
    OGC/1.0/>
...

SELECT ?farm ?sta ?dst {
  SERVICE <http://linkedgeodata.org/sparql> {
      SELECT * { ?farm a lgdo:Farmyard; geos:
      hasGeometry/geos:asWKT ?fWkt; rdfs:
      label ?wktLabel .}}

  SERVICE <loop:> {SELECT ?sta ?dst {?sta a
      sosa:Platform; ca:hasPointGeometry/geos
      :asWKT ?sWkt .
  BIND (spatialF:distance(?sWkt, ?fWkt, unit:
      kilometre) AS ?dst)} ORDER BY ASC(?dst)
       LIMIT 1}
}
```

Listing 3. SPARQL federated query to assert land cover and soil type for station "Pheonix Park".

Notably, Listing 3 breaks the SPARQL 1.1 standards in that the variable `?pWKT` is not in the graph pattern within the inner query to the remote SPARQL endpoint `http://example.vkg.org/sparql`. In other words, running Listing 3 in a regular triple store setting complying with SPARQL 1.1 standards, can fail. Thanks to Fuseki's additional extension for the correlated subquery, the URI of `SERVICE` can be prefixed with a keyword `loop:` to enforce the subquery to be evaluated after the main query. Therefore, the bindings of `?pWKT` in the triple store are resolved first in the outer query and then used in the inner query to find solutions. The decision

to use a correlated subquery should necessarily be based on the expected solutions to graph patterns of local and remote RDF KG. Under the condition that both local and remote RDF KGs (our virtual RDF KGs layering on the PostgreSQL) have spatial indexing enabled. In Listing 3, a standard SPARQL 1.1 federating query first resolves all solutions from the remote RDF KG before passing them into the main query. Nevertheless, this is much less efficient since the remote RDF KG contains more than 2 million multipolygon geometries for land cover and each of them will be examined with each station. The correlated subquery here makes judicious use of much fewer stations for geospatial function examination in the subquery.

5) Enhanced Geo-Contextual Climate Data Access With LinkedGeoData: The availability of places of interest (e.g., OpenStreetMap data) can be useful to improve the accessibility of climate data according to geospatial constraints, especially when the local climate is concerned about its impacts on other environmental sectors. For example, in some agricultural studies, researchers may demand local weather data to understand the impacts of climate on agricultural productivity [54], [55]. In this scenario, we assume that the nearest weather stations of all farmyards on the island of Ireland are needed. To further improve the local accessibility of climate data, we develop a LinkedGeoData [56] SPARQL endpoint of OpenStreetMap within the spatial extent of the island of Ireland. LinkedGeoData uses Ontop to virtualize OSM data in the RDF model, which is consistent with the method described in Section III-C2. LinkedGeoData in our work (LinkeGeoData@Ireland thereafter) is initialized and synchronized with the OSM data of the island of Ireland from GeoFabrick (https://www.geofabrik.de/). LinkedGeoData@Ireland makes places of interest (e.g., a restaurant, a park, or a university) available for SPARQL queries. Therefore, data consumers can use the SPARQL federated query to get the OSM data and link them with the climatic stations in our triple store (see Section III-C3). To find the nearest weather stations of farmyards in the island of Ireland, the following example of SPARQL query (Listing 4) can be made in the Fuseki triple store. In particular, the use of `loop:` here plays a key role in achieving the "loop" mechanism (the implementation of the correlated query in Fuseki) to find the "nearest" for each binding of the graph pattern (`SERVICE <http://linkedgeodata.org/sparql` > ...) in the outer query. This type of query cannot be achieved simply with other triple stores due to the "Top-K" problem in the SPARQL [57] query language. By posting this SPARQL query onto the Sextant, the results can be visualized as shown in Fig. 9.

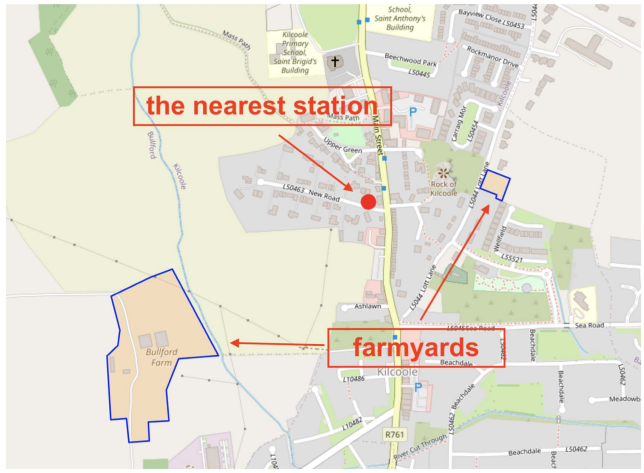


Fig. 9. Nearest stations found for farmyards in LinkedGeoData@Ireland (after zooming in the map).

```
PREFIX lgdo: <http://linkedgeodata.org/
    ontology/>
PREFIX spatialF: <http://jena.apache.org/
    function/spatial#>
PREFIX unit: <http://www.opengis.net/def/uom/
    OGC/1.0/>
...

SELECT ?farm ?sta ?dst {
  SERVICE <http://linkedgeodata.org/sparql> {
      SELECT * { ?farm a lgdo:Farmyard; geos:
      hasGeometry/geos:asWKT ?fWkt; rdfs:
      label ?wktLabel .}}

  SERVICE <loop:> {SELECT ?sta ?dst {?sta a
      sosa:Platform; ca:hasPointGeometry/geos
      :asWKT ?sWkt .
  BIND (spatialF:distance(?sWkt, ?fWkt, unit:
      kilometre) AS ?dst)} ORDER BY ASC(?dst)
       LIMIT 1}
}
```

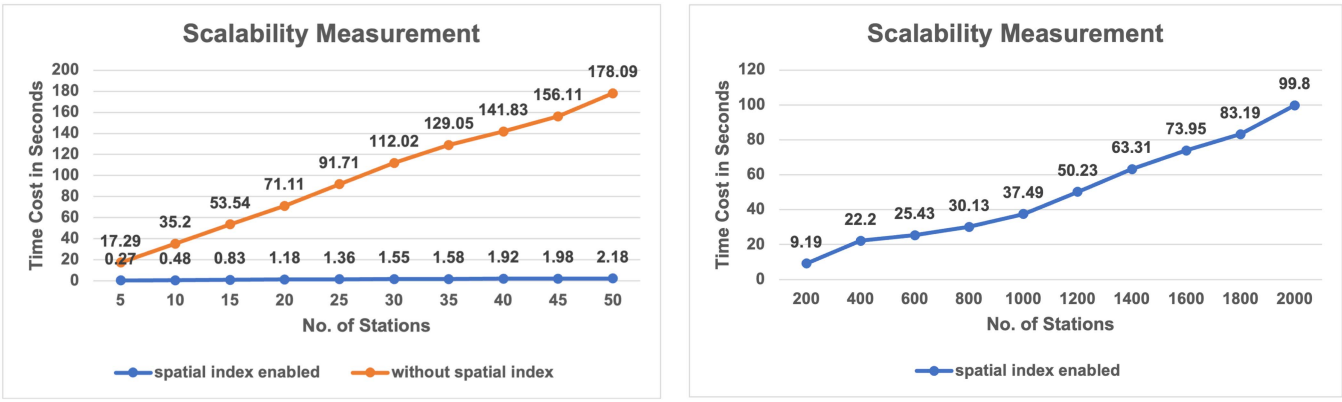
Listing 4. SPARQL federated query to find the nearest stations of OSM farmyards in Ireland where the bindings for `?farm`, `?sta`, and `?dst` denote OSM farmyards, the nearest corresponding stations and calculated distances between them, respectively; The explanations of other SPARQL keywords can be found in SPARQL 1.1 (https://www.w3.org/TR/sparql11-query/).

D. Scalability Analysis on Spatial Queries

The spatial queries in LinkedGeoClimate platform primarily benefit from Apache Jena Fuseki's loop mechanism. We propose two spatial queries above, i.e., 1) the determination of land cover and soil type attributed to climate stations, and 2) the identification of the nearest weather station to farmyards. The former query relies on spatial processing according to Simple Features standards.[3] These standards are meticulously crafted to address topological relationships, such as the intersection of geometries used for discerning the land cover and soil type pertinent to climate stations in this article. The latter is an example of a nontopological query to determine the rankings of distances between geometries. When the loop mechanism coalesces with spatial queries, it becomes imperative to gauge the scalability of such spatial queries, particularly when executed across diverse KGs within the platform. This, in turn, stands as a quintessential benchmark for the query experience afforded to clients.

To evaluate the scalability of spatial queries, we adapt the aforementioned queries (Listings 3 and 4) by varying the size

[3][Online]. Available: https://www.ogc.org/standard/sfo/



(a) (b)

Fig. 10. Scalability measurements against the varying number of queried stations for topological intersection; The line plot representing the database without a spatial index is excluded from figure (b) due to its values being off the chart. (a) Scalability comparison between databases with and without spatial index. (b) Scalability measurement continues with an increasing number of stations.

of the queried entities in the queries and the total size of entities in the KGs. Since the virtual KGs are built upon PostgreSQL databases, we also evaluate the performance difference between databases with spatial indexes or not. In the context of this scalability evaluation, climate stations, CORINE land covers and EPA soil types, OpenStreetMap entities are assumed to be points, multipolygons, and points or linestrings (in alignment with the OpenStreetMap data model), respectively, within the domain of Simple Features. Currently, the queries are dedicated to only climate stations and their interconnections with other geographical features. Consequently, the present scalability assessment is applicable solely to relationships involving points and 1) multipolygons, and 2) points or linestrings. The topological intersection of stations and CORRINE land covers will guide the first relationship evaluation. The nontopological distance (i.e., the nearest) rankings of stations to farmyards will guide the second relationship evaluation. In our future research, more types of data enrichment will be researched on the LinkedGeoClimate platform. This will render an expanded array of relationships amenable to scalability assessment. The evaluation experiment settings are as follows.

1) Hardware: Intel Xeon W-1290P (10 cores @3.7 GHz), 64 GB of RAM, and 1 TB M.2 PCIe SSD.
2) Software:
- OS: Ubuntu 18.04 LTS
- Java: 11.0.13
- SQL database: PostgreSQL 14.4 with GiST (Generalized Search Trees) spatial index enabled
- VKG implementation: *Ontop* 4.2.1
- Triplestore: Apache Jena Fuseki 4.4.0 with 4 GB Memory of JVM

3) Launch type: Hot start, i.e., evalutations are conducted when the processes of VKGs and triplestores are already running in the background.
4) Cost measurement: arithmetic mean of the execution time; each query is run 20 times to get the mean value.
5) Datasets and equivalent geometry size:
- 2018 CORINE land cover (2375406 multi-polygons)
- Met Éireann climate stations (2083 points)
- LinkedGeoData@Ireland entities (varying number of linestrings in queries)

6) Queries used per scenario:
- Topological intersection between climate stations and CORRINE land cover
- Non-topological distance rankings from climate stations to LinkedGeoData@Ireland entities (e.g., OSM buildings)

1) Spatial Query on Topological Intersection: In this section, we undertake an assessment of the platform's scalability in relation to the processing of spatial SPARQL "loop" questions. These queries pertain to the identification of topological intersections between climate stations and CORINE land covers. The quantity of stations in the query will be augmented to 50 in order to assess the fluctuations in time expenditure, measured in seconds. The findings are shown in Fig. 10(a). The figure demonstrates that the use of the GiST geographic index significantly improves the response time of LinkedGeoClimate while processing linear loop queries, as compared to the scenario when no spatial index is employed. The use of the spatial index in the PostgreSQL database presents a significant improvement in performance. In order to assess the scalability of processing geographic queries on LinkedGeoClimate with spatial index capabilities, the study presents Fig. 10(b), which illustrates the time cost as it relates to the increasing number of stations up to 2000. In the experimental setup, the LinkedGeoClimate system was able to successfully check the intersection between 2000 points and 2375406 multipolygons in a time of 99.8 s. This is a significant improvement in efficiency, as it is about two orders of magnitude quicker compared to doing the same task without the use of a spatial index.

2) Spatial Query on Nontopological Distance Rankings: Nontopological distance queries are also essential in the proposed LinkedGeoClimate platform as they provide necessary calculations to identify climate stations located within a certain proximity of a designated point of interest on a map (e.g., OpenStreetMap). To the extent of our current understanding, there exists no topological resolution akin to that provided by spatial indexing in Postgres's PostGIS for ascertaining nearest neighbors within triplestores. In light of this, we make concessions

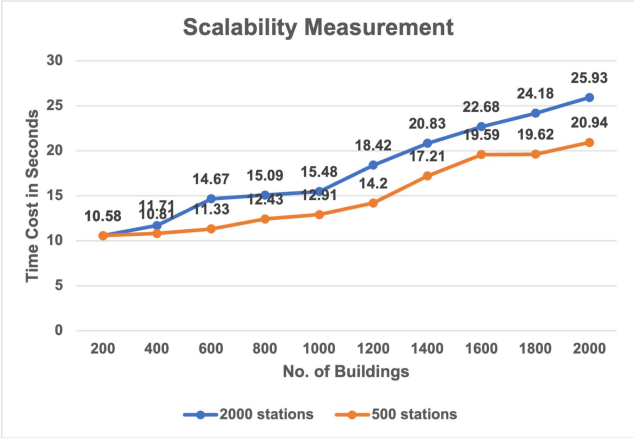


Fig. 11. Scalability measurements against the varying number of OSM buildings; the measurement is driven by determining the nearest station of OSM buildings (linestrings in geometry), and performance is also measured in the different total numbers of stations in a KG .

by using the existing nontopological distance function along with correlated queries to sort the distances in order to find the nearest neighbor across a number of KGs. The scalability results of processing nontopological distance rankings are shown in Fig. 11. We compare the time cost of the spatial query in two different total numbers of climate stations in a KG. According to Fig. 11, the size of buildings for a loop has a more pronounced detrimental impact on the time cost compared to the total number of stations in a KG for distance rankings. Even with four times the stations, time cost varies slightly in rate, i.e., the number of points for distance sorting is better than linear loop performance in scalability.

E. Visualization Aid for Spatial Operations of Climate Data

Access to climatic data is often contingent on the geographical location of stations and other possible places of interest (see Section III-C6). Visual assistance is necessary to see the distribution of the stations on a map. For instance, a data consumer may need to know the distribution of accessible climate data stations within a particular region on a map before consuming data from stations. Theoretically, this can be accomplished simply by utilizing a bounding box or a radius distance to identify the region of interest in GeoSPARQL queries. Yet, data consumers often lack precise coordinates to formulate an appropriate GeoSPARQL query to get the data. A useful way to help users determine the area of their interest is to allow them to sketch a spatial extent directly on a map. This has motivated us to enable a visual interface for the exploration of our RDF KGs from a geographical angle.

We incorporate Sextant [53] into the platform stack of LinkedGeoClimate. Sextant is a state-of-the-art linked geospatial data visualization tool. Importantly, it is compatible with GeoSPARQL standards and is able to recognize the `geos:Geometry` entities and display them on various online map APIs such as Bing Map [58] and OpenStreetMap [19]. We now present a scenario of how data consumers can combine Sextant with our RDF KGs to better access climate data. We assume data consumers are interested in the southside Dublin (divided by River Liffey) climate. The basic usage of RDBMS-based climate data access, such as ECA&D and Met Éireann, cannot precisely determine the southside of Dublin. ECA&D and Met Éireann only provide data at the county level which corresponds to Dublin in this scenario. They might have to find an additional map API to solve it manually or in programming, which is expensive in time. On the contrary, the use of Sextant in Fig. 12 quickly selects stations in southside Dublin by sketching a bounding box beneath River Liffey and adding this spatial constraint to the SPARQL query. Due to the limitation of the rectangular shape of the bounding box, it is presently not easy to precisely exclude all of River Liffey's northern stations. In the case, where a precise exclusion is needed, additional northern stations need to be identified from the map manually and then filtered by updating the SPARQL query. Until now, this approach still significantly reduces the work required by the former approach based on ECA&D or Met Éireann, since all geospatial operations are completed consistently with only the SPARQL query.

IV. Discussion

In this section, we discuss our work from the perspectives of the rationale of choosing the proposed approach, as well as some limitations that need to be dealt with in future potential research.

A. Rationale of Using LinkedGeoClimate

1) Interoperability in Data Integration and Access: LinkedGeoClimate integrates heterogeneous climate data and geographical data over the virtualized ontological layer (i.e., VKG) and exposes data in RDF on-the-fly. This keeps a natural way of ingesting tabular data into PostgreSQL, saving the extra cost (e.g., I/O, bandwidth) needed for semantic annotations required by triple stores. However, VKGs cannot easily be used for RDF data enrichment due to the schema-fixed nature of the underlying RDBMS. To improve the adaptability of the climate data with updates from external RDF KGs, we use a triple store that naturally supports SPARQL federation querying. Thus, data integration and access between LinkedGeoClimate and external RDF KGs is interoperable.

2) Interoperability in the Geographical Context of Climate Data: LinkedGeoClimate uses the open source Fuseki triple store for the extension of RDF data to climate stations, enabling us to achieve the enrichment of geographical metadata and the connections between climate data and the Semantic Web. The correlated query based on the "loop" mechanism of Fuseki can distribute computations on the filter conditions to the high-efficiency end, which breaks through the SPARQL 1.1 limits. For example, when geographical metadata enrichment is concerned, the literal geometries of stations are used in iteration to find the associated geometries in the underlying indexed PostGIS database of the VKG. The "loop" mechanism also provides LinkedGeoClimate with the power to achieve climate data access according to complex geospatial conditions (e.g., the "nearest") while still completely adhering to OGC GeoSPARQL standards.

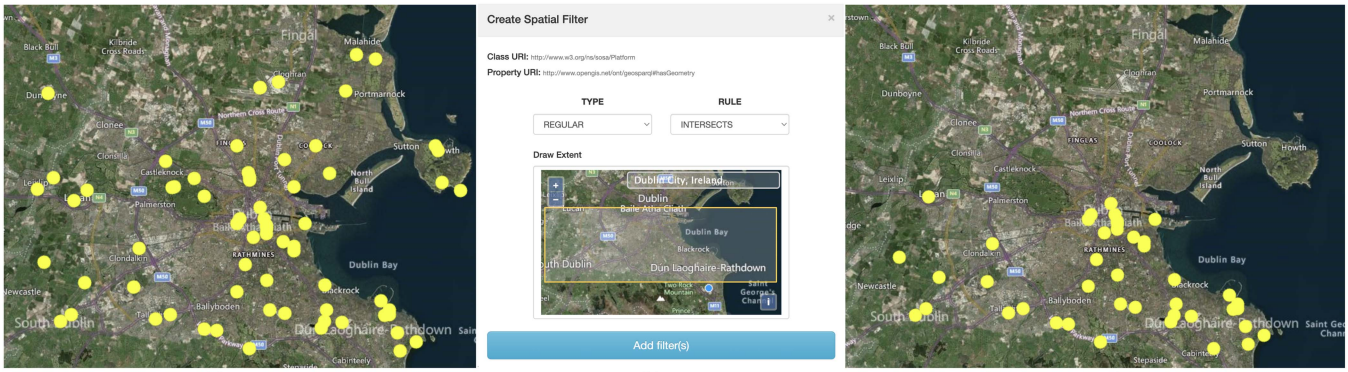


Fig. 12. Demonstration of drawing spatial extent in Sextant for selective stations located in southside Dublin. (a) Stations located in Dublin. (b) Spatial filter for southside Dublin. (c) Station located in southside Dublin.

3) New Knowledge Discovery for Environmental Sciences: The LinkedGeoClimate data is constructed using the semantic web technology stack, featuring triplestores, and virtual KGs. One of the primary advantages of using the semantic web technology stack is its ability to facilitate the creation of interconnected KGs. The discovery of new knowledge is significantly facilitated by the use of interoperable semantic models, namely ontologies, which are designed for KGs specializing in various areas. In this study, we illustrated the process of integrating data from various sources by using SPARQL queries after transforming them into RDF KGs. These examples are created to enhance the acquisition of additional geo-contextual knowledge for climate data. This includes enriching the data with information on land cover and soil type, as well as including comprehensive geographical context sourced from OpenStreetMap. Nevertheless, it is plausible to uncover further novel information using the existing KGs. As an example, it is feasible for users to ascertain that certain soil types exhibit a higher occurrence in regions characterized by specific land cover types, or that changes in land cover and soil types align with climatic data. By incorporating more pertinent data into the platform, individuals may use semantic inference techniques in KGs to derive novel insights from the linked geospatial data. As an instance, individuals may use predefined rules to ascertain the congruence between a certain crop and the soil type, temperature, and land cover characteristics of a given place. The process of uncovering novel insights can be facilitated by the use of SPARQL queries on KGs spanning several disciplines. This approach is further enhanced by the semantic interoperability offered by the platform.

4) Geographical Data Visualization: LinkedGeoClimate provides visualization of geographical data via Sextant which is compatible with the geospatial semantics defined in OGC GeoSPARQL standards. Significantly, the visualization can also help users, especially nonexpert users, define the geospatial constraints and show them with map APIs like OpenStreetMap. We present some examples in the article on filtering data within a bounding box and finding the nearest climate stations to farmyards presented on OpenStreetMap (i.e., LinkedGeoData@Ireland). With the inclusion of LinkedGeoData@Ireland, the use of Sextant can further boost interactive climate data access based on OpenStreetMap geographical data.

B. Current Limitations

1) Data Coverage: At the moment, the geographical context of LinkedGeoClimate is made up of vector datasets, such as the CORINE land cover, the EPA soil map, and OSM data. We have examined how vector data can be aligned with GeoSPARQL semantics via the Fuseki triple store in order to be interoperable and provide geospatial conditions for climate data access. However, geographical data can also come in gridded data format (i.e., raster data). Gridded data is usually produced by satellites, which can capture a complete coverage of the geographical features of the Earth. However, presenting gridded data in RDF for access purposes remains challenging due to the lack of defined standards for grid operations in RDF. LinkedGeoClimate is proposed on the basis of vector data as an interoperable geographical context for climate data access and, therefore, would be short of gridded satellite image data.

2) Nonexpert Usability: LinkedGeoClimate uses Fuseki as the administrative triple store for data integration (SPARQL federation) and access. Currently, all operations are performed at the expert level in the form of native SPARQL queries. The use of Sextant can support a certain degree of geospatial data filtering and visualization, which reduces the learning effort of formulating technical SPARQL queries. For some queries based on complex geospatial conditions, such as the "nearest" which includes another sorting operation of the distance calculations, Sextant does not support nonexpert input (e.g., the `HTML` form). More relevantly, the reuse of query results, i.e., subqueries, is a common demand during an analytical pipeline. For example, in Fig. 9, farmyards may be resolved first and then filtered within a specific area to execute the "finding nearest stations." Sextant uses different layers on the map to present associated query results, but they cannot be combined to create advanced SPARQL subqueries. Because Fuseki has provided `loop` and `cache` mechanisms, we will explore a potential nonexpert usability improvement to LinkedGeoClimate by implementing the subqueries formulation for the purposes of reusing results. A comprehensive usability evaluation is planned after the next iteration of visualization tool development has been concluded. The next iteration tool will overcome some limitations of the initial Sextant-based tool when it comes to advanced geospatial

analysis functionality that we envisage is required, such as the aforementioned subqueries formulation. The assessment of usability will encompass a hybrid approach, combining quantitative methodologies such as the Post-Study System Usability Questionnaire (PSSUQ) [59] with qualitative techniques such as the Think-Aloud protocol [60]. The selection of participants for this evaluation will be drawn from a user cohort comprising prospective consumers of climate data, specifically encompassing researchers affiliated with academic institutions and climate-focused communities.

V. Conclusion and Future Work

Geographical data is important to be used in climate studies associated with various environmental sectors. Most sophisticated climate data platforms, such as ECA&D, NOAA, Met Éireann, encode geographical context in plain metadata which has limited geographical coverage, and implement few geospatial constraints (mainly based on administrative regions) for climate data access. This article proposes LinkedGeoClimate which is an interoperable RDF KGs platform providing climate access within a geographical context. The metadata of climate datasets is modeled with the extended CA ontology and is enriched with the CORINE land cover and the EPA soil using SPARQL federation. In addition, the CORINE land cover, the EPA soil map, and OpenStreetMap data are provided as the interoperable geographical context for climate data. The use of Fuseki provides the necessary semantics (e.g., GeoSPARQL) to manage data integration and access in RDF KGs. Distinguishing from other studies, we explore the correlated queries in Fuseki for SPARQL federation and data access when geospatial constraints are defined. Using correlated queries, users can better exploit geospatial constraints, including complex ones such as "nearest," for interoperable climate data access. We also provide the scalability measurements of correlated spatial SPARQL queries, which demonstrate the efficacy improvement as a result of the spatial index in PostGIS. These measures serve as a baseline for informing the future architectural design of LinkedGeoClimate. Finally, we illustrate the use of Sextant as the visualization window of LinkedGeoClimate, which provides an interactive way to navigate the geographical data in RDF KGs and add spatial filters (bounding boxes) freely on the map.

In the future, we intend to address the current limitations on data coverage and nonexpert usability. For data coverage, we will explore an ideal solution to include gridded satellite imagery for querying. In terms of nonexpert usability, we plan to conduct additional research on relevant analytical pipelines that consume climate data in order to identify the commonly requested database operations. The results will be used to finalize the design and implement a dashboard to assist in the formulation of complex geospatial operations for climate access. In addition, upon the introduction of the new iteration, we shall expand our performance assessments beyond mere scalability considerations, encompassing a comprehensive evaluation of its processing efficiency tailored to the climate domain audience, including knowledge reasoning speed. In addition to implementing a dashboard to enhance usability for nonexperts, there is a current trend towards leveraging generative AI for translating natural language queries into technical queries, such as SQL. It is foreseeable that a solution will soon emerge to further simplify the formulation of SPARQL queries by utilizing prompt engineering, requiring only natural language input.

References

[1] "Global historical climatology network daily (GHCNd)," Accessed on: Jul. 01, 2023. [Online]. Available: https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily

[2] "Weather and climate change," Jul. 01, 2023. [Online]. Available: https://www.metoffice.gov.uk/

[3] 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.

[4] "Climexp data," Jun. 11, 2023. [Online]. Available: [Online]. Available: https://gitlab.com/KNMI-OSS/climexp/climexp_data

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

[6] J. Wu, H. Chen, F. Orlandi, Y. H. Lee, D. O'Sullivan, and S. Dev, "An interoperable open data portal for climate analysis," in *Proc. IEEE USNC-URSI Radio Sci. Meeting (Joint with AP-S Symp.)*, 2021, pp. 104–105.

[7] 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.

[8] F. Manola, E. Miller, B. McBride, and Others, "RDF primer," *W3C Recommendation*, vol. 10, no. 1-107, p. 6, 2004.

[9] J. Wu, F. Orlandi, D. O'Sullivan, and S. Dev, "Ontological modeling of climate data to improve climate analytics," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 2317–2320.

[10] 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.

[11] K. Janowicz, A. Haller, S. J. 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, 2019.

[12] 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, no. 15, pp. 4708–4718, May 2022.

[13] 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 (Joint with AP-S Symp.)*, 2021.

[14] J. Wu, F. Orlandi, T. AlSkaif, D. O'Sullivan, and S. Dev, "Ontology modeling for decentralized household energy systems," in *Proc. Int. Conf. Smart Energy Syst. Technol. (SEST)*, . 2021.

[15] E. D. López-Espinoza, J. Zavala-Hidalgo, R. Mahmood, and O. Gómez-Ramos, "Assessing the impact of land use and land cover data representation on weather forecast quality: A case study in central Mexico," *Atmosphere*, vol. 11, no. 11, 2020, Art. no. 1242.

[16] A. Golzio, S. Ferrarese, C. Cassardo, G. A. Diolaiuti, and M. Pelfini, "Land-use improvements in the weather research and forecasting model over complex mountainous terrain and comparison of different grid sizes," *Boundary-Layer Meteorol.*, vol. 180, no. 2, pp. 319–351, 2021.

[17] S. Lin, N. J. Moore, J. P. Messina, M. H. DeVisser, and J. Wu, "Evaluation of estimating daily maximum and minimum air temperature with MODIS data in east africa," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 18, pp. 128–140, Aug. 2012.

[18] 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.

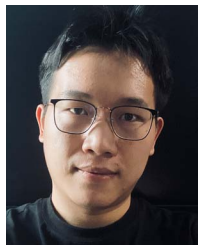
[19] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 12–18, Oct. 2008.

[20] G. Büttner, J. Feranec, G. Jaffrain, L. Mari, G. Maucha, and T. Soukup, "The corine land cover 2000 project," *EARSeL eProceedings*, vol. 3, no. 3, pp. 331–346, 2004.

[21] "Irish soil information system," Mar. 08, 2023. [Online]. Available: http://gis.teagasc.ie/soils/index.php

[22] P. S. Diouf, A. Boly, and S. Ndiaye, "Variety of data in the etl processes in the cloud: State of the art," in *Proc. IEEE Int. Conf. Innov. Res. Develop.*, 2018, pp. 1–5.

[23] "Home European climate assessment & dataset," Jul. 01, 2023. [Online]. Available: https://www.ecad.eu/
[24] I. Foster and C. Kesselman, *The Grid 2: Blueprint for a New Computing Infrastructure*. Amsterdam, The Netherlands: Elsevier, 2003.
[25] I. Foster et al., "The open grid services architecture," in *Proc. Glob. Grid Forum GFD-I*, vol. 30, 2005.
[26] M. Antonioletti et al., "The design and implementation of grid database services in OGSA-DAI," *Concurrency Comput.: Pract. Experience*, vol. 17, no. 2-4, pp. 357–376, 2005.
[27] K. Karasavvas et al., "Introduction to ogsa-dai services," in *Proc. Sci. Appl. Grid Computing: First Int. Workshop*, 2005, pp. 1–12.
[28] M. N. Alpdemir et al., "Using OGSA-DPQ to support scientific applications for the grid," in *Proc. SAG*, vol. 3458, 2004, pp. 13–24.
[29] J. Wu, F. Orlandi, I. Gollini, E. Pisoni, and S. Dev, "Uplifting air quality data using knowledge graph," in *Proc. Photon. Electromagn. Res. Symp.*, 2021.
[30] J. Wu, F. Orlandi, T. AlSkaif, D. O'Sullivan, and S. Dev, "A semantic web approach to uplift decentralized household energy data," *Sustain. Energy Grids Netw.*, vol. 32, no. 100891, Dec. 2022, Art. no. 100891.
[31] M. Compton et al., "The SSN ontology of the W3C semantic sensor network incubator group," *J. Web Semantics*, vol. 17, pp. 25–32, 2012.
[32] J. Wu, F. Orlandi, D. O'Sullivan, and S. Dev, "LinkClimate: An interoperable knowledge graph platform for climate data," *Comput. Geosci.*, vol. 169, Dec. 2022, Art. no. 105215.
[33] R. Catherine, B. Stephan, A. Geraldine, and B. Daniel, *Weather Data Publication on the LOD Using SOSA/SSN Ontology*. Boston, MA, USA: Semantic Web, 2019.
[34] 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.
[35] N. Y. Ayadi, C. Faron, F. Michel, F. Gandon, and O. Corby, "WeKG-MF: A knowledge graph of observational weather data," in *Proc. Semantic Web: ESWC 2022 Satellite Events*. Berlin, Germany: Springer International Publishing, 2022, pp. 101–106.
[36] D. Le-Phuoc, H. Nguyen Mau Quoc, H. Ngo Quoc, T. Tran Nhat, and M. Hauswirth, "The graph of things: A step towards the live knowledge graph of connected things," *J. Web Semantics*, vol. 37-38, pp. 25–35, Mar. 2016.
[37] M. Zárate, P. Rosales, G. Braun, M. Lewis, P. R. Fillottrani, and C. Delrieux, "OceanGraph: Some initial steps toward a oceanographic knowledge graph," in *Knowledge Graphs and Semantic Web*, ser. Communications in Computer and Information Science. Cham, Switzerland: Springer, Jun. 2019, pp. 33–40.
[38] K. Janowicz et al., "Know, know where, know where graph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence," *AIMag*, vol. 43, no. 1, pp. 30–39, Mar. 2022.
[39] B. Bondaruk, S. A. Roberts, and C. Robertson, "Assessing the state of the art in discrete global grid systems: OGC criteria and present functionality," *Geomatica*, vol. 74, no. 1, pp. 9–30, 2020.
[40] L. Lefort, A. Haller, K. Taylor, G. Squire, P. Taylor, and others, "The ACORN-SAT linked climate dataset," *Semant. Pragmat.*, 2017.
[41] M. Zárate, C. Buckle, R. Mazzanti, M. Lewis, P. Fillottrani, and C. Delrieux, "Harmonizing Big Data with a knowledge graph: OceanGraph KG uses case," in *Cloud Computing, Big Data & Emerging Topics*. Berlin, Germany: Springer International Publishing, 2020, pp. 81–92.
[42] "Indices," Mar. 02, 2023. [Online]. Available: https://www.climdex.org/learn/indices/
[43] T. R. Karl, N. Nicholls, and A. Ghazi, "Clivar/gcos/wmo workshop on indices and indicators for climate extremes workshop summary," *Clim. Change*, pp. 3–7, May 1999.
[44] B. DuCharme, *Learning SPARQL: Querying and Updating with SPARQL 1.1*. Sebastopol, CA, USA: O'Reilly Media, Jul. 2013.
[45] O. OGCI, "GeoSPARQL-a geographic query language for RDF data," 2010.
[46] M. Eireann and M. É. A. M. Unit, "Indications of climate change in Ireland. temperature changes at birr and mullingar," 2009.
[47] "CLC 2018 – copernicus land monitoring service," Accessed on: Mar. 14, 2023. [Online]. Available: https://land.copernicus.eu/pan-european/corine-land-cover/clc2018
[48] "Historical data," Accessed on: Mar. 14, 2023. [Online]. Available: https://www.met.ie/climate/available-data/historical-data
[49] D. Lewis, J. Keeney, D. O'Sullivan, and S. Guo, "Towards a managed extensible control plane for Knowledge-Based networking," in *Large Scale Management of Distributed Systems*. Berlin Heidelberg: Springer, 2006, pp. 98–111.
[50] N. F. Noy et al., "Protégé-2000: An open-source ontology-development and knowledge-acquisition environment," in *Proc. AMIA Annu. Symp. proceedings. AMIA Symp.*, 2003, pp. 953–953.
[51] D. Calvanese et al., "Ontop: Answering sparql queries over relational databases," *Semantic Web*, vol. 8, no. 3, pp. 471–487, 2017.
[52] D. Hernández, C. Gutierrez, and R. Angles, "The problem of correlation and substitution in sparql–extended version," 2018, *arXiv:1801.04387*.
[53] C. Nikolaou et al., "Sextant: Visualizing time-evolving linked geospatial data," *J. Web Semantics*, vol. 35, pp. 35–52, 2015.
[54] P. van Oort, B. Timmermans, R. Schils, and N. van Eekeren, "Recent weather extremes and their impact on crop yields of The Netherlands," *Eur. J. Agronomy*, vol. 142, 2023, Art. no. 126662.
[55] D. Bocchiola, L. Brunetti, A. Soncini, F. Polinelli, and M. Gianinetto, "Impact of climate change on agricultural productivity and food security in the Himalayas: A case study in Nepal," *Agricultural Syst.*, vol. 171, pp. 113–125, 2019.
[56] C. Stadler, J. Lehmann, K. Höffner, and S. Auer, "LinkedGeoData: A core for a web of spatial open data," *Semantic Web J.*, vol. 3, no. 4, pp. 333–354, 2012. [Online]. Available: http://jens-lehmann.org/files/2012/linkedgeodata2.pdf
[57] S. Magliacane, A. Bozzon, and E. Della Valle, "Efficient execution of Top-K SPARQL queries," in *Semantic Web–ISWC*, Berlin Heidelberg: Springer, 2012, pp. 344–360.
[58] R. Rischpater and C. Au, *Microsoft Mapping: Geospatial Development With Bing Maps and C*. New York, NY, USA: Apress, 2013.
[59] J. R. Lewis, "IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use," *Int. J. Human–Comput. Interaction*, vol. 7, no. 1, pp. 57–78, Jan. 1995.
[60] C. Lewis, "Using the thinking-aloud method in cognitive interface design," in *IBM TJ* Watson Research Center Yorktown Heights, NY, 1982.

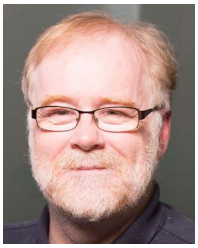


Jiantao Wu received the B.Eng. degree in resources recycling science and engineering and the M.Sc. degree in materials for energy and environment from the University College London, in 2016 and 2017, respectively. He is currently working toward the Ph.D. degree in computer science under the joint supervision of Prof. Soumyabrata Dev at the University College Dublin, Dr. Fabrizio Orlandi and Prof. Declan O'Sullivan at Trinity College Dublin.

He had been a Software Engineer for 2 years at China Electronics Technology Group Corporation until 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 currently a Senior Knowledge and Data Engineer with Inter IKEA Systems. He is also Visiting Research Fellow with the ADAPT research centre, Trinity College Dublin. Prior to this, he was a Marie Skłodowska-Curie EDGE fellow with the same research institute. In these areas, he has experience on foundational and applied research on both EU-funded and industry projects. Prior to joining ADAPT–at Fraunhofer IAIS in Germany–he was Coordinator with the EU H2020 OpenBudgets.eu project and contributed to several large European and industry research projects, such as BigDataOcean.eu and SLIPO.eu. His research activities are focused on 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.

He is currently a Professor in computer science with the School of Computer Science and Statistics, and is a co-applicant Principal Investigator with the ADAPT SFI Research Centre. Since joining TCD from industry in 2001, he has established himself as an international research leader in his field: authoring 300+ scientific peer-reviewed papers and international Journals; being a member of three journal editorial boards, and having undertaken 12+ chair roles in IEEE and IFIP conferences over the years.

Dr. O'Sullivan has won competitive research funding as PI and Co-PI of approximately €10 M. Funding has been won across a range of funding programmes: European Commission (ERC, H2020 and Marie Curie); Science Foundation Ireland (FAME, CNGL, ADAPT); HEA (NEMBES, TGI, SEAROBEND), and from industry: Huawei, Accenture, Ericsson, Nokia Bell Labs, Ordnance Survey Ireland, 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*) from the National Institute of Technology Silchar, India, in 2010, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2017.

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