A Model for Meteorological Knowledge Graphs: Application to Météo-France Data

Nadia Yacoubi Ayadi, Catherine Faron, Franck Michel, Fabien Gandon, Olivier Corby

A Model for Meteorological Knowledge Graphs: Application to Météo-France Data

Nadia Yacoubi Ayadi[1 [0000−0002−6132−8718]], Catherine Faron[1 [0000−0001−5959−5561]], Franck Michel[1 [0000−0001−9064−0463]], Fabien Gandon[1 [0000−0003−0543−1232]], and Olivier Corby[1]

University Côte d'Azur, Inria, CNRS, I3S (UMR 7271), France

Abstract. To study and predict meteorological phenomenons and to include them in broader studies, the ability to represent and exchange meteorological data is of paramount importance. A typical approach in integrating and publishing such data now is to formalize a knowledge graph relying on Linked Data and semantic Web standard models and practices. In this paper, we first discuss the semantic modelling issues related to spatio-temporal data such as meteorological observational data. We motivate the reuse of a network of existing ontologies to define a semantic model in which meteorological parameters are semantically defined, described and integrated. The model is generic enough to be adopted and extended by meteorological data providers to publish and integrate their sources while complying with Linked Data principles. Finally, we present a meteorological knowledge graph of weather observations based on our proposed model, published in the form of an RDF dataset, that we produced by transforming observation records made by Météo-France weather stations. It covers a large number of meteorological variables described through spatial and temporal dimensions and thus has the potential to serve several scientific case studies from different domains including agriculture, agronomy, environment, climate change and natural disasters.

Keywords: Knowledge Graph · Semantic Modelling · Observational Data · Linked Data · Meteorology.

1 Introduction

Meteorological data have attracted great interest in recent years since they are crucial for many application domains. Meteorological observations typically include measurements of several weather parameters such as wind direction and speed, air pressure, rainfall, humidity and temperature. However, these data are mostly collected and stored separately in different files using a tabular data format that lacks explicit semantics, which impedes their integration and sharing to serve researchers from different domains such as agriculture, climate change studies or natural disaster monitoring. Publishing such data on the Web using Linked Data (LD) principles would make them more accessible, easier to discover and reuse. However, integrating and interpreting weather data requires rich metadata about studied features of interest such as the air, observed properties such as the temperature or the humidity, the utilized sampling strategy, the specific location of a weather station and the time (instant or interval) at which the property was measured, and a variety of other information. Getting insights into these heterogeneous data motivates the need of a semantic model in which domain-specific ontologies play a central role by providing a coherent view over it.

In this paper, we propose a semantic model that relies on a network of modular ontologies and domain vocabularies that capture common and specific characteristics of observational meteorological data at a fine grained level, including time, location, provenance, units of measurement, etc. We paid specific attention to propose a model that adheres to LD best practices and standards, thereby allowing for its re-use and extension by other meteorological data producers, and making it accommodated for multiple application domains. To deal with the complexity of the domain knowledge to be modelled, we adopt the SAMOD agile methodology [8] for ontology development, consisting of small steps within an iterative process that focuses on creating well-developed and documented models by using significant exemplar data so as to produce semantic models that are always ready-to-use and easily-understandable by humans. Based on the early work of Uschold & Gruninger [12] the SAMOD process is initiated by a motivating scenario that leads to a set of competency questions that, in turn, provide requirements on the knowledge graph model. We build a self-contained semantic model reusing and extending standard ontologies, among which the GeoSPARQL ontology for spatial features and relations[3], the Time ontology [4] for temporal entities and relations, the Sensor, Observation, Sample, and Actuator (SOSA) [6] and Semantic Sensor Network (SSN) ontologies [5] for sensors and observations, and the RDF Data Cube ontology [10] for aggregation and multidimensionality features.

Furthermore, we implement and make available a software pipeline that is reproducible to generate knowledge graphs compliant with the proposed semantic model. We use the pipeline to generate the first release of the WeKG-MF RDF knowledge graph constructed according to this model from open weather observations published by Météo-France. It includes weather observations from January 2019 till December 2021. To demonstrate the interest of WeKG-MF and the underlying semantic model, competency questions identified in our use case were translated into SPARQL queries to retrieve data from the WeKG-MF knowledge graph in order to meet expert requirements.

The paper is structured as follows. Section 2 describes a motivating scenario that allows us to identify a set of competency questions. Section 3 details our semantic model and highlights its design principles. Section 4 presents the RDF-based knowledge graph constructed from the observational weather data archives of Météo-France. Section 5 presents a validation of the proposed model and the constructed knowledge graph through a set of SPARQL queries implementing the competency questions identified in our motivating scenario. Section 6 presents the related work on lifting meteorological data into RDF datasets. Finally, we conclude and present perspectives of our work in Section 7.

2 Motivating Scenario and Competency Questions

In this section, we present a motivating scenario [12, 8] inspired from requirements expressed by experts and collected in the context of the D2KAB French research project[1]. The primary objective of D2KAB is to create a framework to turn agriculture, agronomy and biodiversity data into semantically described, interoperable, actionable, and open knowledge. Experts in agronomy investigate the correlations between the development rate of plants and weather parameters. They are especially interested in comparing aggregated values of a weather

[1] `https://www.d2kab.org/`

parameter for the same period of time in the same geographic location across years, e.g. the Growing Daily Degrees (GDD) calculated from the daily average air temperature minus a certain threshold called base temperature. This motivating scenario already triggers competency questions that reflect the requirements on the knowledge that has to be represented in the proposed semantic model as well as the way of scoping and delimiting it [12, 8]. We present some of them in the following:

CQ1. What is the measurement unit of a given weather parameter?
Several parameters such as atmospheric pressure, air temperature, wind speed, relative humidity, sea surface temperature are measured using different sensors and procedures, and the resulting numeric/qualitative values are included in weather reports. Measurement units and possible values for qualitative parameters are not included in these reports and are usually documented in external sources (e.g., WMO documentations).

CQ2. At what time of the day was the highest value of a weather parameter measured (observed)?
Temporal features are crucial for observational data. Indeed, within a 24-hour time interval, sensors hosted by weather stations regularly produce different measurement values for the same weather parameter.

CQ3. What is the closest weather station to a specific spatial location?
This competency question points to the fact that the semantic model should encompass a spatial module to capture the geographic coordinates of stations by means of longitude and latitude values.

CQ4. For a specific location and given a calendar interval, provide time series of some aggregated (pre-computed) weather parameters.
Providing aggregated data over relevant time period and for a specific location/weather station is a recurrent need. For instance, daily minimum, maximum and mean temperature, cumulative rainfall during a period of time for each station are examples of significant aggregated parameters for different studies in agronomy or climate change studies.
According to *CQ1* and *CQ2*, weather parameters as well as their significance need to be clearly expressed and formalized. Metadata describing weather properties such as their possible lexical labels in different languages and their possible measurement units are required. *CQ4* is one example of competency questions that require the computation of aggregated values (sum of average temperatures, weekly average temperature). This motivated us to propose a semantic model presented in 3 that combines SSN/SOSA ontologies and RDF data cube vocabulary to represent inherent semantics of observations at different levels of semantic granularity.

3 Semantic Model

Our aim is to design a semantic model in which meteorological variables are semantically defined, described and integrated. The analysis of CQs presented in section 2 led us to select a set of state-of-the art ontologies and thesauri to be re-used. It includes:

- the SOSA/SSN ontologies [6, 5] designed for describing sensors and their observations, and that we extend with new classes to capture the semantics of meteorological observations and provide formal definitions of these new classes. The extension is motivated by the re-use of the Value Sets ontology design pattern;
- the Time Ontology [4] for describing the temporal properties of our data;
- the QUDT ontology and vocabulary [9] representing the various quantity and unit standards and supporting their processing such as conversion;
- the GeoSPARQL vocabulary [3] for representing spatial information in our data;
- the RDF data Cube Vocabulary [10] supporting the publication of multi-dimensional data, such as statistics. We use it to create spatio-temporal slices of meteorological observations by fixing time spans and geographic places as well as applying aggregation functions; SOSA/SSN ontologies only support the description of a single, atomic, observation.

The OWL formalization of our model as well as the related SKOS vocabulary are available in our Github repository[2]. The prefixes of ontologies and vocabularies reused or introduced in this paper are listed in the repository's README[3].

In the following we present in details our model according to four categories of features: features of interest, spatial features, temporal features, and aggregated features.

3.1 Features of Interest and Observable Properties: Describing Observations

In order to propose a self-contained model for representing and publishing meteorological data, we define three new classes. `weo:MeteorologicalObservation` is the core class of our model; it supports the description of a single, atomic observation. A meteorological observation is related to a particular feature of interest, instance of class `weo:MeteorologicalFeature`, and an observable property, instance of class `weo:WeatherProperty`. These three classes specialize classes from the SOSA/SSN ontologies as reflected by their formal definitions.

`weo:MeteorologicalFeature` is defined as a subclass of `sosa:FeatureOfInterest` and serves to represent meteorological features of interest, that is phenomena or events such as precipitations, gusts or storms. Formally, the class is defined as follows:

$$weo{:}MeteorologicalFeature \equiv sosa{:}FeatureOfInterest \cap \\ \forall\, ssn{:}hasProperty.weo{:}WeatherProperty \cap \geq 1\, ssn{:}hasProperty$$

`weo:WeatherProperty` is defined as a subclass of `sosa:ObservableProperty`. Its instances are observable properties of meteorological features. Precipitation amount, gust speed, air humidity are examples thereof. Formally, the class is defined as follows:

$$weo{:}WeatherProperty \equiv sosa{:}ObservableProperty \cap \\ \forall\, ssn{:}isPropertyOf.weo{:}MeteorologicalFeature \cap = 1\, ssn{:}isPropertyOf$$

Instances of `weo:MeterologicalObservation` are observations of a weather property of a certain feature of interest. The definition of `weo:MeterologicalObservation` expresses that only one weather property and one meteorological feature is used for a given meteorological observation:

[2] https://github.com/Wimmics/d2kab/tree/main/meteo/ontology

[3] https://github.com/Wimmics/d2kab/tree/main/meteo

$$weo:MeteorologicalObservation \equiv sosa:Observation \cap$$
$$\forall sosa:observedProperty.weo:WeatherProperty \cap =1\ sosa:observedProperty \cap$$
$$\forall sosa:hasFeatureOfInterest.weo:MeteorologicalFeature \cap =1\ sosa:hasFeatureOfInterest$$

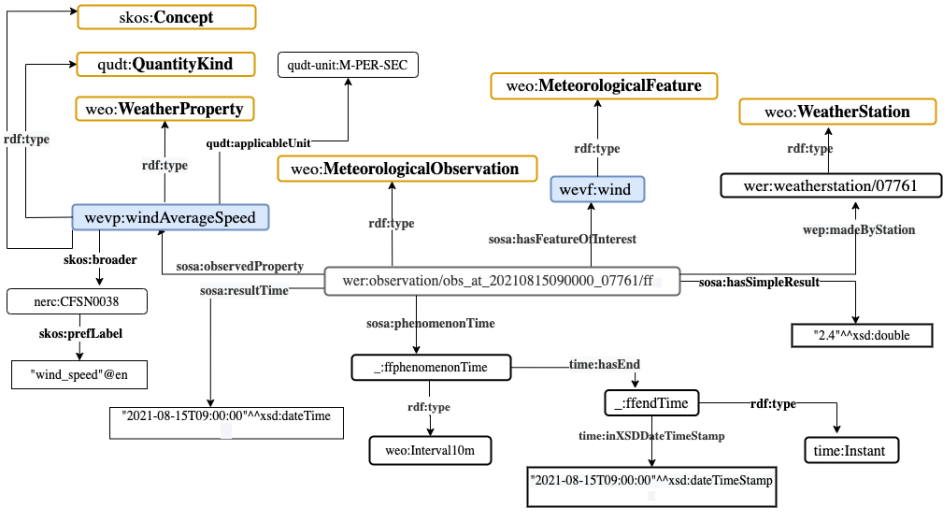

Fig. 1. Example meteorological observation of the `WindAverageSpeed` weather property

Figure 1 depicts the RDF graph representing an example meteorological observation relative to the wind feature of interest and reporting the average wind speed observable property. Although SOSA/SSN ontologies are commonly used to represent knowledge about sensor data across domains, the definition of observable properties and features of interest, as well as their alignment with existing controlled vocabularies, are delegated to the community of interest. Thus, we have reused the Value Sets[4] (VS) ontology design pattern and we defined a SKOS[5] vocabulary whose concepts are instances of `weo:WeatherProperty` and `weo:MeteorologicalFeature` and represent the possible observable properties and features of interest. This SKOS vocabulary is available on our Github repository [6].

An excerpt of it is given in Listing 1.1. The SKOS concepts representing weather properties are aligned with both terms from the NERC Climate and Forecast Standard Names vocabulary and terms from the QUDT Quantity Kind vocabulary that includes general concepts about quantifiable quantities such as `quantity-kind:Speed` or `quantity-kind:Temperature`. For instance, `wevp:averageWindSpeed` and `wevp:gustSpeed` are declared as narrower than `qudt-kind:Speed` (and instances of class `qudt:QuantityKind`). The vocabulary can be easily extended to include new observable properties and features as long as it is compliant with the proposed semantic model.

[4] `https://www.w3.org/TR/swbp-specified-values/`
[5] `https://www.w3.org/2004/02/skos/`
[6] `https://github.com/Wimmics/d2kab/blob/main/meteo/ontology/features-properties-vocabulaire.ttl`

```
wevf:wind a weo:MeteorologicalFeature, skos:Concept ;
 rdfs:label "wind"@en, "vent"@fr ;
 ssn:hasProperty wevp:windAverageSpeed, wevp:windAverageDirection.

wevp:windAverageSpeed a weo:WeatherProperty, qudt:QuantityKind, skos:Concept ;
 ssn:isPropertyOf wevf:wind ;
 skos:broader nerc:CFSN0038, <http://qudt.org/2.1/vocab/quantitykind/Speed>;
 qudt:applicableUnit <http://qudt.org/vocab/unit/M-PER-SEC> ;
 skos:prefLabel "Vitesse moyenne du vent 10mn"@fr,"Average wind speed 10mn"@en;
 wep:hasAbbreviation "ff".
```

Listing 1.1. SKOS representation of meteorological feature `wind` and related weather property `windAverageSpeed`.

```
@prefix : <http://ns.inria.fr/meteo/vocab/weatherproperty/wmocode/> .
:0901 a skos:Collection ;
  rdfs:label "State of ground without snow or ice cover"@en;
  skos:member :0901/0, :0901/1, ... ;

:0901/0 a skos:Concept; rdf:value 0 ;
  skos:definition "Surface of ground dry (without cracks and no appreciable
  amount of dust or loose sand)".

:0901/1 a skos:Concept; rdf:value 1;
  skos:definition "Surface of ground moist" .
```

Listing 1.2. SKOS collection representing the state of ground qualitative weather property (0901 WMO code).

Observation results are literals and an observation is linked to its result by a property `sosa:hasSimpleResult`. Instead of repeating the measurement units within each observation, we denote it at the level of the SKOS concept representing the observable property in our vocabulary (Listing 1.1). Furthermore, some qualitative weather properties require the use of standard encoded values defined by the WMO. For instance, the ground state is a weather property whose possible values (dry, moist, etc.) are in a predefined set of values of the WMO 0901 code[7]. For each qualitative weather properties, we created a `skos:Collection` whose members represent the possible values of the weather property as described in the WMO documentation. Listing 1.2 presents an excerpt of the `skos:Collection` of values for the *state of the ground* weather property.

3.2 Spatial Features: Locating the Weather Stations

A weather station typically hosts sensors and equipments for the purpose of measuring atmospheric conditions and providing information for weather forecasts. Whereas a description of sensors and equipment is not always made available by meteorological data providers, relevant metadata about weather stations generally include station identifier, name, latitude, longitude and altitude. Our model introduces the `weo:WeatherStation` class to represent any type of weather station. To capture stations' spatial location, `weo:WeatherStation` is introduced as a subclass of `geosparql:Feature`. Therefore, each instance of `weo:WeatherStation`

[7] `https://epic.awi.de/id/eprint/29967/1/WMO2011i.pdf`

has a geometry that is a point with specific coordinates. Following GeoSPARQL vocabulary, geo-coordinates of a weather station are defined as a Well-Known Text (WKT) literal (e.g., POINT(8.792667 41.918)). Our adoption of GeoSPARQL is motivated by the fact that it allows to efficiently query spatial data based on a set of spatial functions. It enables us to express spatial queries involving meteorological data, e.g. retrieving the closest station to a given location or the precipitations for a specific location. We also reused latitude, longitude, and altitude datatype properties from the WGS84 vocabulary since WKT literals do not integrate information about the altitude of a station.

3.3 Temporal Features: Defining Time Entities

In many cases, the observation of a given weather property is made over a period of time. The duration of a measurement varies depending on the property. For such cases, we reuse the `sosa:phenomenonTime` property to link an instance of `weo:Meteorological- Observation` to an instance of `time:Interval`. Since time durations are described in the documentation of weather observed properties, we defined different time interval classes by expressing an OWL restriction on their duration that may be declared in seconds, minutes or hours. The interest of doing this is that these time intervals are declared once in our model and are reused for all observations, and thus avoid substantial redundancy. For instance, in Figure 1 the `wevp:windAverageSpeed` weather property is measured during a period of 10 minutes. This is denoted by property `sosa:phenomenonTime` whose value is an instance of class `weo:Interval10m`, while the end time of the interval is an instance of class `time:Instant`.

3.4 Aggregated Features: Defining Observation Slices

Observations produced by sensors can rapidly reach enormous volumes. The *CQ4* competency question (see Section 2) stresses the need to create focused and homogeneous sets of observations that share some dimension. In particular, creating times series of air temperatures or other weather parameters is a recurrent need. In this respect, we reuse the RDF Data Cube vocabulary (DCV)[8] to describe multi-dimensional data according to a 'data cube' model. Each data cube is an instance of class `qb:DataSet` and is linked to instances of class `qb:DataStructureDefinition` by property `qb:structure` (Figure 2). A Data Structure Definition (DSD) defines the structure of a data cube and how observations are linked to the measures and dimensions of the data cube. Listing 1.3 presents an example of DSD `wes:annualTimeSeriesTemperature` that defines the structure of a data cube of air temperatures. According to this DSD, each observation contains three daily measures: the minimum, maximum and average temperatures. The `qb:Slice` class enables to represent a subset of observations that share the same dimensions. In our model, we declare spatio-temporal slices of observations by fixing the spatial and temporal dimensions: the spatial dimension may refer to the weather station, while the temporal dimension corresponds to a calendar interval. While the SOSA/SSN ontologies only support the description of a single, atomic, meteorological observation, an observation (instance of `qb:Observation`) in a spatio-temporal slice is represented by a set of measures (instances of `qb:MeasureProperty`) each linked to an observable property (declared in our SKOS vocabulary) with property `qb:concept`.

[8] `https://www.w3.org/TR/vocab-data-cube/`

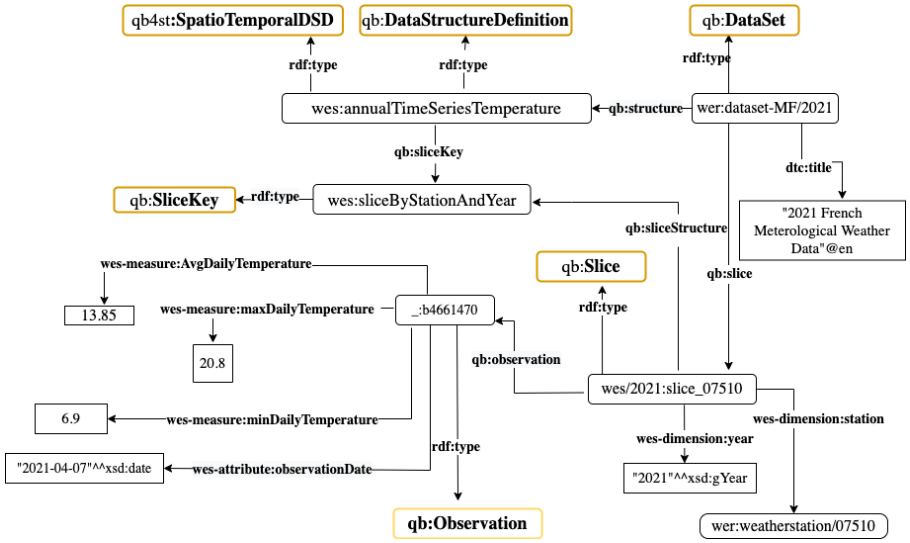

Fig. 2. Example of an RDF data Cube slice representing a TimeSeries of Air Temperatures

Furthermore, observations from the same `qb:Slice` have attributes such as the observation date that refers to a 24-hour interval during which a certain value of a certain parameter is selected with respect to a specific condition or aggregation (e.g. maximum daily temperature).

4 Météo-France Weather Observations RDF Dataset

This section presents the pipeline that we set up to lift the observation reports published by Météo-France into an RDF knowledge graph named *WeKG-MF* (Weather Knowledge Graph - Météo-France), that complies with the model presented in Section 3.

4.1 Météo-France Dataset

In France, the primary source of weather data and forecasting is the Météo-France[9] organisation. As a member of World Meteorological Organization (WMO)[10], Météo-France provides access to daily meteorological observations. These data are the result of measurements performed by 62 weather stations located in different regions in metropolitan France and overseas departments. Measurements are generated by different sensors/equipments hosted by weather stations, collected in daily tabular data files such as the table presented in Figure 3. Each line corresponds to the values of meteorological parameters measured or observed at a given weather station (column 1) at a specific date and time (column 2). For instance, column *u* denotes the values of "relative air humidity" measured at different times of the day at different

[9] `https://www.meteofrance.com/`
[10] `https://public.wmo.int/en/`

```
<http://ns.inria.fr/meteo/dataset-MF/2021> a qb:DataSet ;
 qd:structure wes:annualTimeSeriesTemperature ;
 dct:title "French Meterological Weather Data of 2021"@en ;
 dct:description "Daily min/max/avg temperature in 2021"@en .

wes:annualTimeSeriesTemperature
 a qb:DataStructureDefinition, qb4st:SpatioTemporalDSD ;
 qb:component
  [qb:dimension wes-dimension:year ; qb:componentAttachment qb:DataSet],
  [qb:dimension wes-dimension:station ; qb:componentAttachment qb:Slice],
  [qb:measure wes-measure:minDailyTemperature],
  [qb:measure wes-measure:maxDailyTemperature],
  [qb:measure wes-measure:avgDailyTemperature],
  [qb:attribute wes-attribute:observationDate] ;
 qb:sliceKey wes:SliceByStationAndYear.

wes-dimension:station a rdf:Property, qb:DimensionProperty ;
    rdfs:range weo:WeatherStation.

wes-dimension:year a rdf:Property, qb:DimensionProperty ;
    rdfs:range xsd:gYear.

wes-measure:minDailyTemp a  rdf:Property, qb:MeasureProperty ;
    rdfs:label "Daily Minimum Temperature"@en;
    rdfs:range xsd:decimal ;
    qb:concept wevp:minAirTemperature .
```

Listing 1.3. The `wes:annualTimeSeriesTemperature` Structure Definition.

numer_sta	date	pmer	tend	cod_tend	dd	ff	t	td	u	vv	ww	w1	w2	n	nbas	hbas	cl	cm	ch
07005	20210212030000	102710	-90	6	90	6.300000	269.350000	264.350000	68	20000	0	mq	mq	mq	mq	mq	mq	mq	mq
07015	20210212030000	102960	-10	6	50	3.000000	267.850000	265.150000	81	8940	0	mq	mq	mq	0	mq	mq	mq	mq
07020	20210212030000	102080	-20	6	100	13.100000	274.450000	267.950000	62	12000	2	2	mq	100	8	1250	35	mq	mq
07027	20210212030000	102300	-30	6	110	7.300000	271.350000	265.250000	63	41450	0	mq	mq	100	8	2580	mq	mq	mq

Fig. 3. Snapshot of a CSV file of meteorological parameters

location. However, presented in a tabular-delimited structure and stored separately in different files, weather measurements are hardly exploitable.

4.2 Lifting Process

We downloaded from Météo-France's portal[11] the list of SYNOP[12] weather stations in GeoJSON format[13], and the monthly observation reports generated by these stations as CSV files. Measurement are generated every 3 hours and disseminated into the WMO network in less than 15 minutes. Then, we implemented a reproducible software pipeline to generate WeKG-MF in compliance with the proposed model. The core of the pipeline is the mapping task that is performed with Morph-xR2RML tool[14], an implementation of the xR2RML

[11] `https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=90&id_rubrique=32`

[12] SYNOP: surface synoptic observations, a numerical code used for reporting observations made by weather stations.

[13] `https://donneespubliques.meteofrance.fr/donnees_libres/Txt/Synop/postesSynop.json`

[14] `https://github.com/frmichel/morph-xr2rml/`

mapping language [7] for MongoDB databases. Pipeline scripts as well as xr2RML mapping triples are available in our github repository[15].

Additionally, we enriched the weather stations' descriptions by linking each station to the closest Wikidata entity, based on its geographic coordinates, using property `dct:spatial`. This allows us to get further information about the regions, departments and municipalities in which weather stations are located using simple SPARQL queries. Furthermore, leveraging Wikidata allows to benefit from its many links to other data sources, in particular the French national institute for statistics and economic studies (INSEE) which is highly used and trusted by French organisms.

WeKG-MF is published under an open licence, is assigned a DOI[16] and can be downloaded from Zenodo. In the short term, we intend to make it available through a public SPARQL endpoint.The current version of WeKG-MF covers the period from January 2019 to November 2021. Statistics about its content are provided in Table 1.

Category	**Resources**
Total Nr. of triples	60.601.248
Nr. of classes	9
Nr. of weather stations	62
Nr. of Observations for 2019	2.788.528
Nr. of Observations for 2020	2.789.574
Nr. of Observations for 2021 (till November 2021)	2.528.467
Nr. of weather properties	22
Nr. of meteorological features	6
Nr. of Observations per observed property	$\approx$ 405.328
Nr. of Air Temperatures slices	183

Table 1. key statistics of the WeKG-MF dataset

5 Validation: Implementing the Competency Questions

The validation process is intended to check the consistency of the model and its ability to address requirements and cover the domain [8]. In Section 2, we have presented an example motivating scenario that pointed to a set of competency questions which reflect requirements that potential users may want to get answers for. In this section, we evaluate the proposed semantic model by demonstrating how CQs can be translated into SPARQL queries. Note that the model and the WeKG-MF dataset were loaded in a Virtuoso triple store deployed as a Docker image.

5.1 Querying Low-Level Observations

Let us first address *CQ2* "*At what time was the highest value of a weather parameter measured (observed)?*". It points to the need to query the exact time at which a given parameter reaches

[15] `https://github.com/Wimmics/d2kab/tree/main/meteo/Lifting-dataset`

[16] `https://doi.org/10.5281/zenodo.5925413`

```
SELECT ?date ?hour ?station ?temp_max WHERE {
   {
   SELECT ?date  ?s  (MAX(?v) as  ?temp_max)
     WHERE {
       ?obs a  weo:MeteorologicalObservation;
         sosa:observedProperty wevp:airTemperature ;
         sosa:hasSimpleResult  ?v;
         wep:madeByStation ?s ;
         sosa:resultTime ?t .
       BIND(xsd:date("2020-08-01") as ?date)
       FILTER(xsd:date(?t) = ?date) }
     GROUP BY ?s ?date
   }
   ?obs a  weo:MeteorologicalObservation;
     sosa:observedProperty wevp:airTemperature ;
     sosa:hasSimpleResult  ?temp_max ;
     wep:madeByStation ?s ;
     sosa:resultTime ?t .
   ?s rdfs:label ?station  .
   FILTER(xsd:date(?t)= ?date)
   BIND(HOURS(?t) as ?hour) }
```

Listing 1.4. SPARQL query implementing CQ2

```
SELECT  ?label ?lat ?long ?coordinates WHERE {
        ?x rdfs:label ?label ;
           geosparql:hasGeometry [ geosparql:asWKT ?coordinates].
           geo:lat ?lat; geo:long ?long .
BIND("Point(0.1413499 45.1423348)"^^geosparql:wktLiteral as ?Currentposition)
BIND (geof:distance(?coordinates,?Currentposition , uom:metre) as ?distance)

}
ORDER BY ?distance
LIMIT 1
```

Listing 1.5. SPARQL query implementing CQ3

its peak. Our model captures the importance of temporal features surrounding observational data by capturing the exact time at which each and every observation is generated. The SPARQL query, presented in Listing 1.5, is a formal translation of CQ2 that allows us to retrieve, for each station available in the WeKG-MF dataset, at what time the maximum air temperature was reached on August 1st, 2021. It shows that *CQ2* can be successfully converted and executed as a SPARQL query over the dataset. Another set of SPARQL queries leveraging spatial GeoSPARQL functions demonstrate how end-users can query meteorological observations based on geospatial coordinates of weather stations. For instance, *CQ3* expresses the need to query the closest weather station given specific geospatial coordinates.

5.2 Querying Observation Slices

Let us now address the *CQ4* "*For a specific location and given a calendar interval, provide time series of some aggregated weather parameters?*". This question motivates our adoption of the RDF Data Cube Vocabulary to represent pre-calculated time series of aggregated weather parameters. For example, in agronomy, experts are interested in calculating GDD values that are calculated based on the average daily temperature minus a base temperature

```
SELECT ?date ?station ?temp_avg ?GDD  WHERE {
    BIND(URI("http://ns.inria.fr/meteo/weatherstation/07510") as ?station)
    ?s  a qb:Slice ;
      wes-dimension:station ?station ;
      wes-dimension:year "2021"^^xsd:gYear ;
      qb:observation [
        a qb:Observation ;
        wes-attribute:observationDate ?date ;
        ?p ?temp_avg ] .
    ?p a qb:MeasureProperty ; qb:concept wevp:airTemperature .
    BIND((?temp_avg - 10) as ?GDD) }
ORDER BY ?date
```

Listing 1.6. SPARQL query implementing CQ4

which varies from a crop to another. Note that daily average temperature corresponds to the average of the minimum and maximum temperatures measured during a 24-hour interval. Listing 1.6 shows the SPARQL query formalizing competency question CQ4 and shows it can easily calculate GDD values based on pre-calculated slices corresponding to a specific weather station and by selecting beginning date of a calendar interval. Note that the value of 10 in the query denotes an example of base temperature. Without pre-calculated slices, *CQ4* could be implemented by a SPARQL query that computes min/max/avg temperatures for a specific weather station on the fly. However, the complexity of the writing of the query as well as its execution time would be significantly higher. The generation of spatio-temporal slices is done once and they can be reused for the calculation of any new aggregated parameters and facilitates their implementation.

5.3 Implemented Notebook and Visualizations

We developed a set of SPARQL queries available on the Github repository of our project[17], together with a Jupyter Notebook that demonstrate how the results of SPARQL queries can be used to generate visualizations from the WeKG-MF knowledge graph. As an example, Figure 4 presents different types of data visualisations. The first plot (on the top-left) shows daily cumulative precipitations measured at the "Bordeaux-Merignac" station and the second one (on the top-right) shows the evolution of daily average temperature collected from weather stations located in the French region of "Nouvelle Aquitaine". Both plots show a comparison of aggregated values calculated based on two weather parameters (precipitation and air temperatures) available in the WeKG-MF knowledge graph. The third visualisation (on the bottom-center) shows the different weather stations located in Metropolitan France.

6 Related Work

In this section, we present existing research works on the publication of meteorological data as LOD datasets. First, the AEMET meteorological dataset [2] makes available some data sources from the Spanish Meteorological Office through a SPARQL endpoint. The dataset is based on the AEMET ontology network which follows a modular structure: a central ontology

[17] `https://github.com/Wimmics/d2kab/tree/main/meteo/sparql-examples`

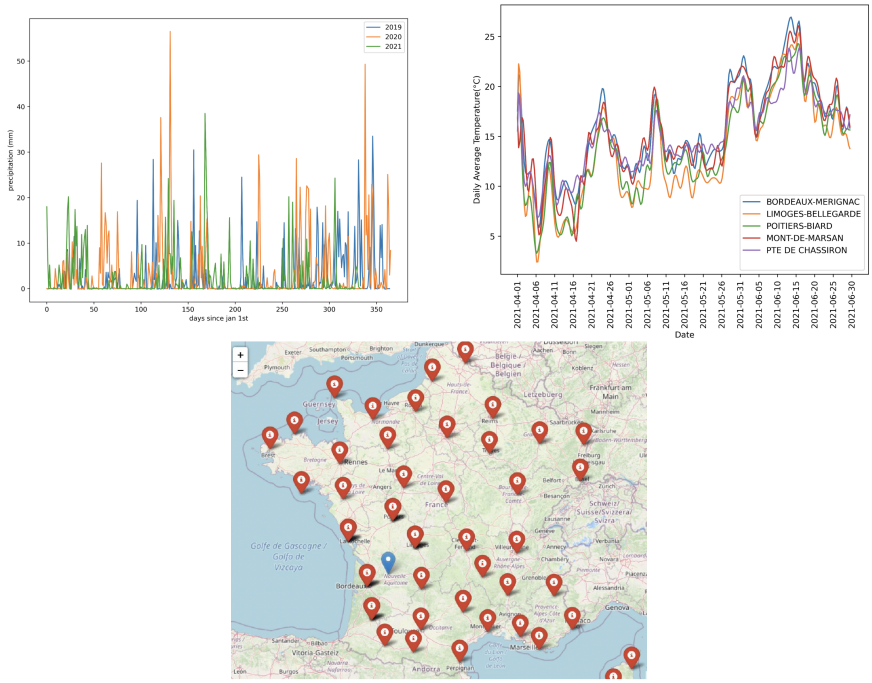

Fig. 4. Examples of Visualisation of Daily Precipitations, Average Temperature and Weather Station Locations in Metropolitan France

relates a set of ontologies that describe different sub-domains involved in the modeling of meteorological measurements. These sub-domains are: (meteorological) Measurements, Sensors, Time and Location. As an attempt to access to the dataset, we tried to quey the AEMET SPARQL endpoint[18], however, we noticed that the endpoint is no longer available[19]. The authors of [11] present an RDF dataset of meteorological measurements made by a weather station located at the Irstea experimental farm. Our proposition is in line with their work as we rely on most of the ontologies that they used (SOSA/SSN, GeoSPARQL, QUDT, OWL-Time ontology). Yet, we adopt somehow different design principles to propose a minimal yet extensible semantic model for meteorological data. Furthermore, we extend their work to support the description and dynamic generation of homogeneous slices of observations pre-calculated using aggregation functions over temporal and spatial dimensions. Thus, we are able to represent annual times series of daily min, max and average temperatures for each weather stations in our dataset.

The authors of [1] propose an ontological model to represent metadata and data schema of meteorological observation data from the Météo-France archives. The focus of this work is to enable access and understanding of the data sources (weather reports) with adherence to FAIR principles, yet without actually transforming the observational data included in weather

[18] `http://aemet.linkeddata.es/sparql`
[19] Last attempt on February, 7th 2022

reports into RDF data. In our work, we are interested not only in describing observational data but also in transforming them into semantically-enriched observations accessible via SPARQL queries in order to enable their integration in a wide range of applications from different domains such as agronomy or natural disaster monitoring.

7 Conclusion and Future Works

Meteorological observations refer to values of different weather observable properties measured across space and time by means of different sensors and equipment available in weather stations. Transforming these data into RDF knowledge graphs bridges the semantic gap between observational data and other resources also published on the Web as Linked Open Data, thus enabling their re-use in different domain applications. In terms of sustainability, we provide a fully automatic pipeline that enables us the update of the WeKG-MF graph over time with new weather data downloaded from Météo-France.

Towards this goal, in this paper we proposed a reusable and extensible model that semantically describes the multiple dimensions behind meteorological data. Our semantic model reuses the SOSA/SSN ontologies and extends it with new classes about specific feature of interest entities. These classes are rigorously defined and aligned with third-party vocabularies and ontologies. We rely on Time Ontology and GeoSPARQL to capture the spatio-temporal context surrounding observational data, as well as the QUDT schema and vocabulary to include metadata about measurement units of observed weather properties. We leverage the RDF Data Cube vocabulary to create slices of weather parameters that are the result of aggregation functions over spatial and temporal dimensions. This is typically needed to represent time series of min/max/average temperatures or precipitations in a given spatial area. We also propose a SKOS vocabulary of observable properties and features aligned with existing controlled vocabularies. In addition, we generated and published WeKG-MF, an RDF knowledge graph complying with this semantic model, from Météo-France meteorological data observations. To the best of our knowledge, our research work is the first that proposes a meteorological RDF-based knowledge graph.

This work was started in the context of the D2KAB French project[20]. Within this project, a use case concerns the design and development of a reading interface for the Plant Health Bulletins (PHB) that are meant to inform bio-vigilance stakeholders about the status of plant diseases and crop pests in French regions. This interface shall be able to augment reading experience by integrating related information likely to provide the reader with enriched context and insights into the data they are currently reading. Various related information may be involved, such as phenological stages of crops and pests, phenotyping information, taxonomic resources, geographic references and meteorological observations record history. In the latter, we typically expect the aggregated data (such as max/min/avg temperature or precipitation and the measure of Growing Daily Degrees) to be of utmost importance for experts to draw hypotheses about, e.g., the possible impact of weather conditions on the advent of crop pests at different periods or phenological stages.

[20] `https://www.d2kab.org/`

References

1. Amina Annane, Mouna Kamel, Nathalie Aussenac-Gilles, Cássia Trojahn, Catherine Comparot, and Christophe Baehr. Un modèle sémantique en vue d'améliorer la fairisation des données météorologiques. In Maxime Lefrançois, editor, *IC 2021 : 32es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 32nd French Knowledge Engineering Conference), Bordeaux, France, June 30 - July 2, 2021*, pages 20–29, 2021.
2. Ghislain Auguste Atemezing, Óscar Corcho, Daniel Garijo, Jose Mora, María Poveda-Villalón, Pablo Rozas, Daniel Vila-Suero, and Boris Villazón-Terrazas. Transforming meteorological data into linked data. *Semantic Web*, 4(3):285–290, 2013.
3. Robert Battle and Dave Kolas. Enabling the geospatial semantic web with parliament and GeoSPARQL. *Semantic Web*, 3(4):355–370, 2012.
4. Simon Cox and Chris Little. Time ontology in OWL. W3C candidate recommendation 26 march 2020, W3C Organism, 2020.
5. Armin Haller, Krzysztof Janowicz, Simon J. D. Cox, Maxime Lefrançois, Kerry Taylor, Danh Le Phuoc, Joshua Lieberman, Raúl García-Castro, Rob Atkinson, and Claus Stadler. The modular SSN ontology: A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation. *Semantic Web*, 10(1):9–32, 2019.
6. Krzysztof Janowicz, Armin Haller, Simon J. D. Cox, Danh Le Phuoc, and Maxime Lefrançois. SOSA: A lightweight ontology for sensors, observations, samples, and actuators. *J. Web Semant.*, 56:1–10, 2019.
7. Franck Michel, Loïc Djimenou, Catherine Faron-Zucker, and Johan Montagnat. Translation of relational and non-relational databases into RDF with xR2RML. In Valérie Monfort, Karl-Heinz Krempels, Tim A. Majchrzak, and Ziga Turk, editors, *WEBIST 2015 - Proceedings of the 11th International Conference on Web Information Systems and Technologies, Lisbon, Portugal, 20-22 May, 2015*, pages 443–454. SciTePress, 2015.
8. Silvio Peroni. SAMOD: an agile methodology for the development of ontologies. In Mauro Dragoni, Maréa Poveda-Villalón, and Ernesto Jimenez-Ruiz, editors, *OWL: Experiences and Directions – Reasoner Evaluation*, pages 55–69. Springer, 2016.
9. Jack Hodges Ralph Hodgson, Paul J. Keller and Jack Spivak. QUDT quantities, units, dimensions and data types ontologies. Technical report, NASA, 2014.
10. Dave Reynolds and Richard Cyganiak. The RDF data cube vocabulary. W3C recommendation, W3C, January 2014. https://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/.
11. Catherine Roussey, Stephan Bernard, Géraldine André, and Daniel Boffety. Weather data publication on the LOD using SOSA/SSN ontology. *Semantic Web*, 11(4):581–591, 2020.
12. Mike Uschold and Michael Gruninger. Ontologies: Principles, methods and applications. *The knowledge engineering review*, 11(2):93–136, 1996.