



HAL
open science

Improving FAIRness of the SYNOP meteorological data set with semantic metadata

Amina Annane, Mouna Kamel, Cassia Trojahn, Nathalie Aussenac-Gilles,
Catherine Comparot, Christophe Baehr

► To cite this version:

Amina Annane, Mouna Kamel, Cassia Trojahn, Nathalie Aussenac-Gilles, Catherine Comparot, et al.. Improving FAIRness of the SYNOP meteorological data set with semantic metadata. *International Journal of Metadata, Semantics and Ontologies*, 2023, 16 (2), pp.118-137. 10.1504/IJMSO.2023.135332 . hal-04420766

HAL Id: hal-04420766

<https://hal.science/hal-04420766v1>

Submitted on 29 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Improving FAIRness of the SYNOP Meteorological Dataset with Semantic Metadata

Amina Annane

IRIT, CNRS, Université de Toulouse, France
E-mail: amina.annane@irit.fr

Mouna Kamel

IRIT, Université de Toulouse, France
E-mail: mouna.kamel@irit.fr

Cassia Trojahn

IRIT, Université de Toulouse, France
E-mail: cassia.trojahn@irit.fr

Nathalie Aussenac-Gilles

IRIT, CNRS, Université de Toulouse, France
E-mail: nathalie.aussenac-gilles@irit.fr

Catherine Comparot

IRIT, Université de Toulouse, France
E-mail: catherine.comparot@irit.fr

Christophe Baehr

CNRM, Météo-France, France
E-mail: christophe.baehr@meteo.fr

Abstract: Meteorological data, essential in a variety of applications, has been made available as open data through different portals, either governmental, associative or private ones. Making this data fully findable and reusable for experts from other domains than meteorology requires considerable efforts to guarantee compliance to the FAIR principles. Nowadays, most efforts in data FAIRification are limited to semantic metadata describing the overall features of datasets. However, such a description is not enough to fully address data interoperability and reusability by other scientific communities. This paper addresses this weakness by proposing a semantic model to represent different kinds of metadata, describing the data schema and the internal structure of a dataset distribution, together with domain-specific definitions. This model is used to provide a reusable schema of the SYNOP dataset, a largely used governmental meteorological dataset in France. The impact of using the proposed model for improving FAIRness was evaluated.

Keywords: Metadata; Ontologies; Meteorological data; FAIR principles.

1 Introduction

Meteorology data is essential in many applications, including weather forecasts, climate change, environmental studies, agriculture, health and risk management. It consists of different types of measurements of the Earth's atmosphere, such as air pressure, temperature or water vapour, including the interactions of those measurements or derivative physical quantities. Their production is based either on measurement tools and sensors, like those embedded

on weather stations, satellites and weather radars, or on mathematical models that assimilate the data from several of the previous sources. This data has been systematically captured or computed for many years, and it is still produced every day, in larger and larger volumes as new devices (IoT, cars or personal weather stations) measure weather features and new models (mathematics simulation models enriched with data-based learning) are developed.

This data was made available as open data and datasets through governmental portals like MeteoFrance

(<https://donneespubliques.meteofrance.fr/>) and worldweather (<http://worldweather.wmo.int/fr/home.html>), or associative or private ones (e.g., infoclimat (<https://www.infoclimat.fr>) or meteociel (<http://www.meteociel.fr>), under open licenses. Nevertheless, its exploitation from web portals is rather limited. On these portals, datasets and data are described and presented with properties that are relevant for meteorology domain experts (data producers) but that are not properly understood and reusable by other scientific communities. For the latter, one of the challenges is to find relevant data among the increasingly large amount of continuously generated data, by moving from the point of view of data producers to the point of view of users and usages.

One way to overcome these weaknesses is to guarantee compliance of data to the FAIR principles: Findability, Accessibility, Interoperability, and Reusability (Wilkinson et al. 2016). These principles correspond to a set of 15 recommendations that aims to facilitate data reuse by humans and machines. They are domain-independent and may be implemented principally by: (F), assigning unique and persistent identifiers to datasets, and describing them with rich metadata that enable their indexing and discovery; (A), using open and standard protocols for dataset access; (I), using formal languages, and FAIR vocabularies to represent (meta)data; and (R), documenting (meta)data with rich metadata about usage license, provenance and data quality. So the first step towards the fulfilment of FAIR principles is to assign metadata to the datasets, and to define precise metadata schemes. Indeed, 12 out of the 15 FAIR principles refer to metadata (Wilkinson et al. 2016). To go a step further in improving data FAIRness, several authors have shown that metadata schemes should be based on semantic models (i.e., ontologies) for a richer metadata representation (Guizzardi 2020). Thanks to their ability to make data types explicit, in a format that can be processed by machines, ontologies are essential to make data FAIR, even for data already published on the web (Jacobsen & et al. 2020).

While most efforts in data FAIRification are limited to specific kinds of metadata, mainly those describing the overall features of datasets, such a description is not enough to fully address all FAIR principles (Koesten et al. 2020), in particular for promoting reuse of this data by other scientific communities. With a focus on meteorology domain, we propose to address this weakness thanks to a rich representation of the meaning of meteorological data as well as the structure of the dataset in a formal model that allows semantic meaning to be shared with third-parties (Kremen & Necaský 2019). The proposed model is capable of representing different types of metadata, in particular i) those that describe the data schema and the internal structure of a dataset distribution and ii) those precisising the domain-specific definitions.

This effort comes from the need to go towards the FAIRisation of a large amount of meteorological data collected over more than 20 years by a public institution in France (Météo-France). The “SYNOP” (Synoptic dataset) dataset corresponds to a collection of files that share the same structure to represent the same types of data. Thank to the proposed model, the internal dataset structure is semantically schematised and made explicit, so that it can be reused to describe every SYNOP file with descriptive metadata. This is typically the case when dealing with observation data, which form a large volume of data exposed in a shared structure.

Contrary to existing works involving ontology population (Lefort et al. 2012, Roussey et al. 2020, Atezing et al. 2013, Patroumpas et al. 2019, Arenas et al. 2018), and due to the characteristics of meteorological data and to the data provider choices, we do not transform all data into RDF but rather represent in a fine-grained way the data schema and its distribution structure. The proposed model relies on existing FAIR vocabularies and ontologies and is itself compliant to the FAIR principles. To sum up, the contributions of this paper are the following:

- Proposing a semantic model for representing different kinds of metadata, in particular those describing the data schema and internal structure of a dataset distribution, together with domain-specific definitions.
- Reusing and integrating different existing (FAIR) vocabularies and ontologies.
- Proposing a schema that relies on the semantic model and that can be reused to semantically annotate the largely used meteorological dataset SYNOP provided by Météo-France – the official weather entity in France.
- Evaluating the FAIRness degree of this dataset without and with the proposed model, showing how the proposed model improves its FAIRness.

This paper extends the work done in (Annane et al. 2021) by providing (i) a detailed description of the methodology adopted in the construction of the proposed meteorological model based on the reuse of existing vocabularies and ontologies, including a specification of the ontology in terms of competency questions; (ii) a detailed integration of these different vocabularies and ontologies in the proposed model; (iii) a better distinction between the generic part of the model and the SYNOP structural schema model; and (iv) an evaluation of the FAIRness of the proposed model.

This work was carried out in the context of the Semantics4FAIR project which aims at facilitating the tasks of searching and accessing scientific data resulting from both the research and production of one scientific community, in order to support the development of new uses by other scientific communities. In this project,

biologist researchers aim to identify the meteorological conditions that favor the germination and flowering of ragweed. Hence, the need for accessing and reusing meteorological data is therefore driven by this use case.

The rest of this paper is organised as follows. Section 2 introduces the SYNOP dataset, followed by the model specification in terms of ontology requirements in Section 3. A discussion on the reuse of existing vocabularies and ontologies is presented in Section 4, followed by their integration in the proposed model Section 5. The annotation of a SYNOP dataset is presented in Section 6, followed by the evaluation of its FAIRness in Section 7. Section 8 discusses the related work and finally Section 9 concludes the paper.

2 Overview of the SYNOP datasets

SYNOP data represents observation data from international surface observation messages circulating on the Global Telecommunication System (GTS) of the World Meteorological Organization (WMO). On the Météo-France website, SYNOP is presented on a web page (https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=90&id_rubrique=32), with few metadata (in natural language): title, description, access rights, two files as documentation, and a form to select the date or the month for which the user wants to download SYNOP data.

Figure 1 shows an excerpt of the SYNOP data along with its documentation. SYNOP data is structured as tabular data with 59 columns (on the top of Figure 1). The two first columns are “numer_stat” (i.e., the identifier of the observation station where the measurement/observation was made) and “date” (i.e., the date when the measurement/observation was made). Then each column among the following 57 ones, represents a measure/observation. Another file lists all the observation stations (on the left bottom side of Figure 1) of Météo-France, where each row represents a station with their properties (identifier, name, and localisation). A PDF file (on the right bottom side of Figure 1) completes the dataset with a description of different SYNOP data columns.

This dataset however suffers from several weaknesses that restrict its exploitation by non-user experts and in an automated way:

- lack of rich metadata: metadata are provided in natural language which prevents crawlers of dataset search engines to exploit it. Therefore it negatively affects the capacity to discover the dataset.
- lack of parameter definitions: the documentation file does not provide definitions of the various parameters, which would be required for a user unfamiliar with meteorological vocabulary (for example, the acronym “td” is described as “point de rosée” (*dew point*), but no definition is given

to explain what a *dew point* is in meteorological terms.

- imprecise parameter descriptions: for instance, the acronym “t” is associated to the description “Temperature”. However, in meteorology different types of temperature can be measured such as: air temperature, soil temperature, etc. By convention, meteorologists know that “Temperature” refers to “air temperature”, however a biologist who is not familiar with such conventions will need a more precise description.
- missing documentation on coded values: some parameters have coded values such as the “cod_tend” parameter so that each value (e.g., 8, 7) has a specific interpretation. The code is defined by WMO. On the documentation file, a link to the manual documenting this code (a PDF file) is provided, however this link is broken. Therefore, the user cannot understand the coded values.
- no API for data access: the user has to download data day by day or month by month using the form on the dataset web page.

In order to address most of these shortcomings (as API data access has to be provided by the data provider), a semantic model for annotating the SYNOP datasets with rich metadata is proposed. The specification of this model is presented in the following.

3 Ontology specification

The essential activities for the development of the proposed semantic model include *specification*, *conceptualization*, *formalization*, and *implementation* (Figure 2), as defined in most ontology construction methodologies. The reader can refer to (Cristani & Cuel 2005) for an early review on them. As an extension of these activities, the NeOn methodology (Suárez-Figueroa et al. 2015) proposes different scenarios that can be combined to meet the needs of ontology developers, as opposed to a single, rigid scenario that would be implemented to build ontologies from scratch. In particular, the NeOn Scenario 3 “*Reusing ontological resources*” promotes the reuse of existing vocabularies and ontologies as a way to improve interoperability (Suárez-Figueroa et al. 2015). The sequence of these different steps is described in Figure 2. Reusing is also compatible with the construction of FAIR metadata models. Hence, the proposed approach relies on the reuse of (FAIR) existing vocabularies and ontologies. This section describes the ontology specification step while the conceptualisation guided by reuse is discussed in the next one.

The goal of the specification activity is to establish the purpose and scope of the ontology (why the ontology is being built, what are the intended uses and end-users, etc.). Here, this specification is based on four

numer_sta	date	pmer	tend	cod_tend	dd	ff	t	...
7005	20200201000000	100710	-200	8	200	3.200000	285.450000	...
7015	20200201000000	100710	-170	7	200	7.700000	284.950000	...
7020	20200201000000	100630	-40	5	210	8.400000	284.150000	...
7027	20200201000000	100770	-130	6	200	5.500000	285.650000	...
7037	20200201000000	100830	-230	6	200	7.000000	285.150000	...
7072	20200201000000	101140	-190	8	210	4.900000	285.450000	...
7110	20200201000000	100780	-60	8	230	4.500000	284.750000	...
...

Excerpt from Synop data

ID	Nom	Latitude	Longitude	Altitude
7005	ABBEVILLE	50.136000	1.834000	69
7015	LILLE-LESQUIN	50.570000	3.097500	47
7020	PTE DE LA HAGUE	49.725167	-1.939833	6
...

Documentation: Excerpt from the list of stations

Descriptif	Mnémonique	type	unité
Indicatif OMM station	numer_sta	car	
Date (UTC)	date	car	AAAAMDDHHMISS
Pression au niveau mer	pmer	int	Pa
Variation de pression en 3 heures	tend	int	Pa
Type de tendance barométrique	cod_tend	int	code (0200)
Direction du vent moyen 10 mn	dd	int	degré
Vitesse du vent moyen 10 mn	ff	réel	m/s
Température	t	réel	K
Point de rosée	td	réel	K

Documentation: Excerpt from the parameter description document

Figure 1 Excerpts of SYNOP data and its documentation as they are provided on Météo-France public data website.

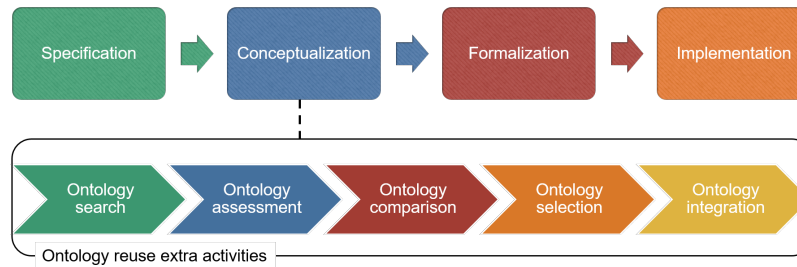


Figure 2 Scenario 3 of the NeOn methodology for reuse-oriented ontology construction.

dimensions: (i) what are the knowledge needs of users looking for meteorological data expressed as competency questions; (ii) what are the features of meteorological data; (iii) what are the metadata required for improving the FAIRness of the datasets; and (iv) what are the key concepts that should be covered by the semantic model. These specifications are summarised in the ORSD document (Section 3.5).

3.1 Competency questions

In ontology authoring, the knowledge needs of an ontology can be formalised thanks to *competency questions* (CQs) that were introduced as *ontology's requirements in the form of questions the ontology must be able to answer* (Grüninger & Fox 1995). For the specification of the proposed model, a set of competency questions were collected from the interviews with several biologist researchers involved in the project use case. The collected CQ are listed in Table 1.

3.2 Meteorological data characteristics

Another point of the ontology specification concerns the characterisation of the data the ontology will be used to describe. While there exist different types of meteorological data (satellite data, model data that are computed using statistical models such as weather forecast data, radar data, etc), the focus here is

on observation data referred to as “in situ” data. These are direct measurements of various parameters (temperature, wind, humidity, radiation, etc.) taken by instruments on the ground or at altitude from predefined locations (observation stations). Hence, the following characteristics of the data have to be taken into account:

- *Geospatial data*: the measure values must be localised, otherwise they are not fully exploitable. The localisation is usually defined using geospatial coordinates (latitude, longitude and altitude). The interpretation of these coordinates depends on the used Coordinate Reference System (CRS), hence the CRS has also to be specified.
- *Temporal data*: each measurement is made at a specific time that must be associated with the measurement result (i.e., value). As for the geospatial, the temporal localisation is essential to the right interpretation of measurements.
- *Observation data*: to be conform to the World Meteorological Organization (WMO) guidelines, many other parameters must be specified such as the measurement procedures, the types of sensors that captured the data, or the quality standards.
- *Tabular data*: observation data are usually published in tabular format where measure values are organized according to spatio-temporal

Table 1 List of competency questions.

#	Competency question
1	What is the temporal resolution of the data?
2	What is the license for data use?
3	What is the temperature (air/ground/etc.)?
4	In which format can data be exported/downloaded?
5	What is the precise date day/month/year/hour/minute/second) of a measurement?
6	How can we download the data?
7	How often are measures made (decadal/hourly)?
8	What are the atmospheric parameters related to precipitation?
9	What are the input parameters?
10	Who is the contact person of the dataset?
11	What is the spatial coverage of the dataset?
12	What are the title of each column of a distribution?
13	To what measure corresponds the "t" in the distribution?

dimensions. According to a recent study by (Benjelloun et al. 2020), the tabular format is the most widespread format for publishing data on the Web (37% of the datasets indexed by Google are in CSV or XLS).

- *Large volume of data:* meteorological data are produced continuously. In each weather station, several sensors are installed (thermometer, barometer, etc.). Each sensor generates multiple measurement values with a frequency that differs from one measurement to another (hourly, tri-hourly, daily, etc.).

3.3 Metadata ensuring FAIR principles

Making data FAIR requires first and foremost the generation of metadata. Indeed, 12 out of the 15 FAIR principles refer to metadata as explained in (Wilkinson et al. 2016). This metadata must remain accessible even if the data itself is no longer accessible. These 12 principles provide guidance on the categories of metadata: (i) descriptive metadata for data indexing and discovery (title, keywords, etc.); (ii) metadata about data provenance; (iii) metadata about access rights and usage licenses. Particularly for publishing data on the web, W3C recommends three other categories of metadata: (i) version history; (ii) quality; (iii) structure. Our goal is therefore to propose a metadata model that covers these different categories, thus ensuring adherence to the principle on rich metadata.

3.4 Key concepts

Based on the study of SYNOP datasets, on the set of CQs, on the observed data characteristics and on the FAIR principles, the following key concepts have been identified. A SYNOP **Dataset** is composed of different files (one per month), each of them corresponding to a fragment or **Slice**. Each fragment may be stored in different formats, each format giving rise to one **Distribution**. The dataset is created, published,

updated, etc. by an **Agent**. Data mainly correspond to **Spatial** and **Temporal Measures** (temperature, humidity, rain, sunshine, etc.), provided by **Sensors**, according to different **measure Units**. These measures are stored in a **Tabular Format**, each **Column** storing one kind of measure (which is described thanks to its semantics, its value, its unit, etc.).

At the end of this study, it appears as well that a SYNOP dataset (and many other types of datasets) must be described at different levels: at the dataset level with the general characteristics of the dataset (publisher, licence, versioning, etc.), at the structural level (how is data structured ? in tabular or multidimensional format, according to a schema, etc.), and at the data level (is data temporal, spatial, and/or domain specific, etc. ?).

3.5 Ontology Requirement Specification Document (ORSD)

As recommended by the NeOn methodology (Suárez-Figeroa et al. 2015), the ontology specification activity results can be summarized in the ORSD, as in Table2. This document includes the purpose and the scope of the ontology, the implementation language (OWL2 DL), the intended end users, the intended uses, the ontology requirements and the preliminary Glossary of terms.

4 Ontology reuse

As introduced below, the NeOn methodology includes activities related to the reuse of ontologies (ontology search, ontology assessment, ontology comparison, ontology selection and ontology integration) within the conceptualization step (Figure 2). Here, ontology reuse guides this step, which aims at structuring the acquired knowledge. Knowledge acquisition had brought to light the need for representing both metadata and fine-grained data schema and its distribution structure. Furthermore, in order to be compliant to the "I" principle, it is required to reuse FAIR ontologies (Guizzardi 2020, Poveda-Villalón et al. 2020). In this section, the steps

Purpose	Provide a semantic model of meteorological dataset metadata as a part of a larger process aiming at making FAIR SYNOP meteorological data
Scope	SYNOP Meteorological observation dataset, with the level of granularity related to the identified CQ and terms
Implementation language	OWL2
Intended end-users	User 1. meteorological data providers who want to publish their data User 2. meteorological data users who want to search for existing datasets and reuse them
Intended uses	Use 1. To publish dataset metadata Use 2. To index SYNOP meteorological dataset on international and national data portals (improve discoverability of the dataset) Use 3. To search for meteorological datasets
Ontology requirements	<i>Non-Functional Requirements</i> Data archives should not be transformed into RDF Reuse of reference ontologies The ontology must support a multilingual scenario (English and French) <i>Functional requirement</i> Representing meteorological dataset metadata Representing SYNOP dataset distribution structure Representing SYNOP dataset schema and semantics
Pre-glossary	dataset, dataset metadata, dataset schema, measure, station, measure localization, time, distribution, distribution structure spatial coverage, temporal coverage, measurement method measurement instrument, measuring documentation

Table 2 Ontology Requirements Specification Document (ORSD)

of ontology search, ontology assessment, comparison and selection are exposed, while ontology integration is discussed when presenting the proposed model.

4.1 Ontology search

According to the key concepts identified in Section 3.4, we searched for existing vocabularies and ontologies that would represent them, in several ontology repositories: Linked Open Vocabularies (LOV) (<https://lov.linkeddata.es/dataset/lov/>), vocab.org (<http://purl.org/vocab/>), ontologi.es (<http://ontologi.es/>), SOCoP+OOR (<https://ontohub.org/socop/>), BioPortal (<https://bioportal.bioontology.org/>), AgroPortal (<https://agroportal.lirmm.fr/>), OntoHub (<https://ontohub.org/>), COLORE (<http://stl.mie.utoronto.ca/colore/>), and Open Ontology Repository (OOR) Initiative (<http://www.oor.net/>), ONKI service (<https://onki.fi/>), as well as in academic papers. An ontology expert guided by the ontology requirement criteria performed a manual search with the support of search engines.

4.2 Ontology assessment and comparison

The selected ontologies were then assessed in terms of their relevance to the ontology specification (Table 3). The “Dataset description” column gathers 5 types of metadata (according to the classification proposed by (Greiner et al. 2017)), representing descriptive metadata (title, description, publisher, etc.), provenance,

access rights, versioning, and quality metadata. The number of stars in this column corresponds to the number of met criteria. To increase the ontology FAIRness, we paid attention to the FAIRness of the reused vocabularies and ontologies: whether they were recommended or candidate to recommendation, whether they corresponded to W3C working group notes or working drafts, and whether they had a fairsharing status (FS).

4.3 Ontology selection

From Table 3, the requirements of the ontology are covered by the following ontologies and vocabularies: GeoDcat-AP for describing the dataset, CSVW and RDF Data Cube for the dataset structure, OWL-Time for the time aspect, SOSA for observations, SWEET, ENVO, QUDT which are domain specific. All these vocabularies are recommended by W3C or FS, and are briefly described in the following subsections. The prefixes and namespaces for these ontologies and vocabulaires and for those used in the rest of the paper are listed in Table 4.

4.3.1 Dataset metadata

GeoDCAT-AP is the selected vocabulary for describing a SYNOP dataset at the dataset level.

GeoDCAT-AP. GeoDCAT-AP <http://data.europa.eu/930/> is a specification of the DCAT-AP vocabulary which is an application profile (AP) for the W3C

Vocabulary	Dataset description	Structure			Data				Status	
		Tabular	Multid.	Schema	Time	Space	Obser.	Domain	W3C	FS
AWS								*		
CF								*		
CSVW		*						*		
DCAT V2	****								*	R
DCAT-AP	****									
DQV WGN	*								WGN	R
ENVO								*		R
GeoDCAT-AP	*****							*		
GeoSPARQL						*				R
INSPIRE	****									R
JSON Schema				*					WD	
OA	*	*	*	*	*	*	*	*	*	
PROV-O	*								*	R
QUDT								*		R
RDF Data Cube		*	*						*	R
schema.org	****			*						R
SOSA							*		CR	R
SWEET					*		*	*	*	R
Time					*				CR	R
XML Schema				*					*	R

Table 3 Comparison of the reusable vocabularies and ontologies. For the “Dataset description” column the number of stars corresponds to the number of covered type of metadata (descriptive, provenance, access rights, versioning, and quality metadata).

DCAT (Data CAtalog vocabulary) recommendation. The choice of GeoDCAT-AP is motivated by the richness of this vocabulary for metadata representation. It allows to describe datasets and their distributions, using a large panel of metadata: descriptive metadata (title, language, source, user, etc.), access rights, quality, provenance and versioning (Figure 3). These metadata also offer specific properties required to correctly interpret spatial data such as the geographical area covered by the data (`dct:spatial`), the used reference coordinate system (`dct:conformsTo`) to be chosen from a list defined by the OGC (<http://www.opengis.net/def/crs/EPSS/>), as well as the spatial resolution (`dcat:spatialResolutionInMeters`) of the data. GeoDCAT-AP is also recommended by W3C/OGC to describe geospatial data on the Web (van den Brink & *et al.* 2019).

4.3.2 Structural metadata

As mentioned early in the paper, we chose not to transform all data into RDF because it would be (i) expensive: transforming the data archived for decades requires human and physical resources, and (ii) not effective: it would result in a huge RDF graph that would not be effective for querying and accessing the data (Karim et al. 2020). We chose CSVW to represent the syntactical structure of a tabular dataset distribution, while RDF data cube (qb) and domain ontologies are used to represent the semantics of the dataset structure, independently of any specific data format.

CSVW. As pointed out in (Koesten et al. 2020), it is essential for data reuse to represent the internal structure of the data. Since observation data are mostly tabular data, CSVW (<https://www.w3.org/ns/csvw>) is a suitable vocabulary. It results from the work of the

Table 4 Ontology namespaces.

prefix	namespace
dc	http://purl.org/dc/elements/1.1/
ns1	http://www.w3.org/2006/vcard/ns#
owl	http://www.w3.org/2002/07/owl#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
xml	http://www.w3.org/XML/1998/namespace
xsd	http://www.w3.org/2001/XMLSchema#
csvw	http://www.w3.org/ns/csvw#
qb	http://purl.org/linked-data/cube#
geodcatap	http://data.europa.eu/930/
dcat	http://www.w3.org/ns/dcat#
foaf	http://xmlns.com/foaf/0.1/
qudt	http://qudt.org/1.1/vocab/unit
rdfs	http://www.w3.org/2000/01/rdf-schema#
skos	http://www.w3.org/2004/02/skos/core#
dct	http://purl.org/dc/terms/
dmo	https://https://w3id.org/dmo#
dmo-synop	https://www.irit.fr/recherches/MELODI/ontologies/DMO/dmo-synop#
sosa	http://www.w3.org/ns/sosa/
sweetp	http://sweetontology.net/propPressure/
qb4st	http://www.w3.org/ns/qb4st/
envo	http://purl.obolibrary.org/obo/

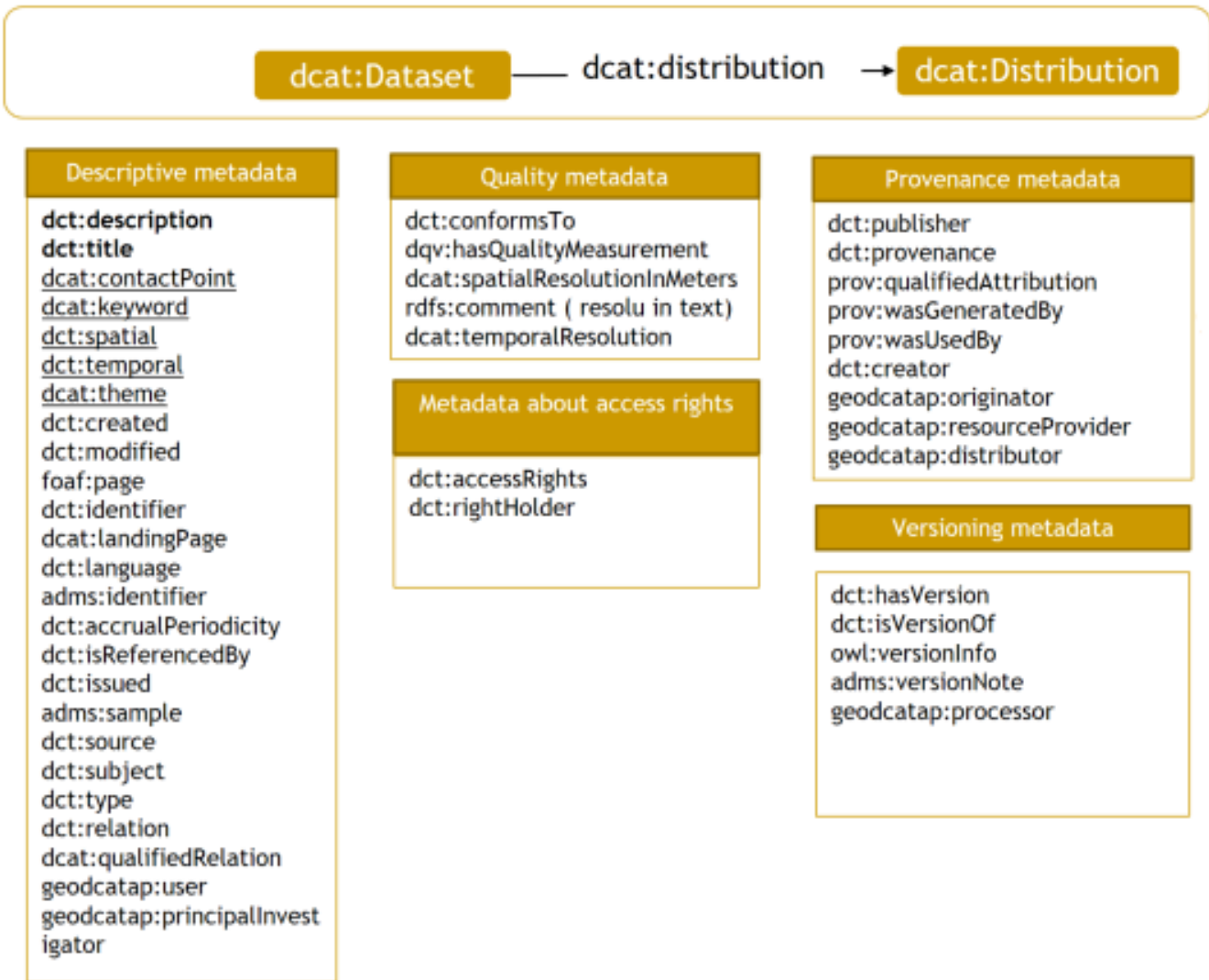


Figure 3 GeoDcat-AP vocabulary: main reused concepts and properties.

W3C group on publishing tabular data on the web. It allows to define the different columns `csvw:Column` of a given `csvw:Table` (i.e., csv file) via the `csvw:Schema` concept. Moreover, it represents the interdependence between two tables. Indeed, it allows to represent if a column (or a set of columns) in a given CSV file is a foreign key `csvw:ForeignKey` that references a column (or columns) of another CSV file. An overview of this CSVW is given in Figure 4.

RDF data cube (qb). qb (<https://www.w3.org/TR/2011/REC-qb/>) is a W3C vocabulary (van den Brink & *et al.* 2019) dedicated to the representation of multi-dimensional data. qb is suitable in our case since observation data is multi-dimensional and organized according to spatio-temporal dimensions (Figure 5). A dataset (`qb:Dataset`) is related to its fragments via the `qb:slice` property, a slice being linked to a Slicekey (a subset of the component properties of a Dataset which are fixed in the corresponding slices) by the `qb:sliceStructure` property. A dataset is associated to a structure (`qb:DataStructureDefinition`)

via the property `qb:structure`. Multidimensional data schema is then described using three subclasses of `qb:ComponentProperty`: (i) measures (`qb:MeasureProperty`), (ii) dimensions (`qb:DimensionProperty`) according to which the measures are organized, and (iii) attributes to represent additional information (e.g. unit of measurement) (`qb:AttributeProperty`). The `qb:concept` property allows to link a `qb:ComponentProperty` (i.e., measure, dimension or attribute) to the corresponding domain concept to make its semantics explicit. We use this property to associate component properties to domain concepts. The RDF Data Cube extensions for spatio-temporal components (`qb4st`) <https://www.w3.org/TR/2011/REC-qb4st/> emphasizes the spatio-temporal aspects by specializing qb concepts. In particular, the `qb:DataStructureDefinition` class has been specialized to cover spatio-temporal dimensions (`qb4st:SpatioTemporalDSD`), as indicated in Figure 5.

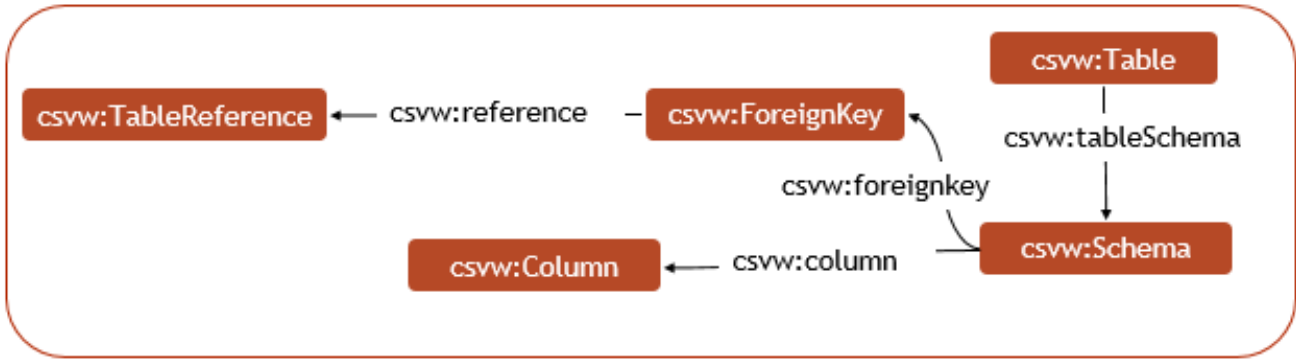


Figure 4 csvw vocabulary: main reused concepts and properties.

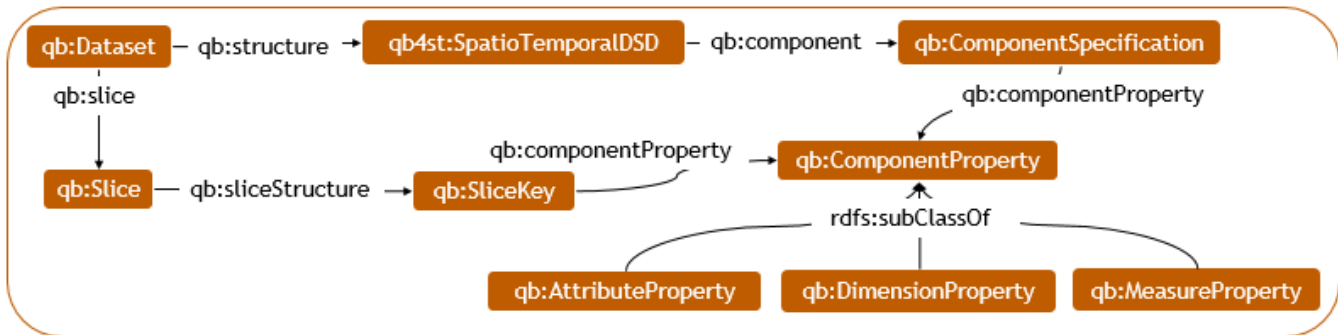


Figure 5 RDF data cube: main reused concepts and properties.

4.3.3 Domain specific metadata

In addition to qb, several domain and cross-domains ontologies have been reused for making explicit the semantics of measures and dimensions thanks to concepts that belong to the meteorological domain, such as atmospheric parameters (e.g., temperature, wind speed) or sensors (e.g. thermometer, barometer).

SWEET (Semantic Web Earth and Environment Technology ontology). SWEET (Raskin 2006) is a collection of ontologies conceptualizing a knowledge space for Earth system science, including both orthogonal concepts (space, time, Earth realms, physical quantities, etc.) and integrative science knowledge concepts (phenomena, events, etc.). We are interested by the part of SWEET that models meteorological parameters such as humidity, wind speed, pressure at sea level or rainfall.

ENVO (Environment ontology). ENVO (Buttigieg et al. 2013) is a knowledge representation of environmental entities, allowing the description of ecosystems, entire planets and other astronomical bodies, their parts, or environmental processes. It helps (meta)data records to achieve demonstrable FAIRness. ENVO can be used in addition to SWEET to better describe environmental processes, offering for example the possibility to specify the extremes of a temperature (minimum and maximum).

SOSA (Sensor, Observation, Sample, and Actuator). SOSA (Janowicz et al. 2019) is an ontology for describing sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties, as well as actuators. It is the reference ontology for representing observations (measures) from sensors (such as thermometer, barometer, etc.).

QUDT (Quantities, Units, Dimensions and Data Types). QUDT (<http://www.qudt.org/>) defines the base classes, properties, and restrictions used for modelling physical quantities, units of measure, and their dimensions in various measurement systems. For our purpose, QUDT allows to specify the unit of measurement of each measurement.

OWL-Time. OWL-Time (<https://www.w3.org/TR/owl-time/>) is an ontology of temporal concepts, for describing the temporal properties of resources in the world. This vocabulary allows to express facts about topological (ordering) relations among instants and intervals, together with information about duration, and about temporal position including date-time information. OWL-Time provides the means to model meteorological processes in time.

4.4 Ontologies reuse implementation

According to (Carriero et al. 2020), ontology reuse can be implemented in three different ways: (i) *direct reuse*: the reused ontologies or a selection of their terms are imported into a new ontology; (ii) *indirect reuse*: terms from external ontologies are reused as templates, or just aligned to the terms and their semantics (axioms) naively described in the new ontology; and (iii) *hybrid reuse*: it is a design choice where ontology terms are selected either for direct reuse or for being indirectly reused as templates, according to characteristics of reused ontologies and requirements of the project. We built the *dmo* model following a *hybrid reuse* approach:

- Direct reuse: by means of the annotation `owl:imports`, the main ontologies that compose the proposed model and that are reused in a large amount, are imported: GeoDCAT-AP, qb and QB4st, and CSVW. It is the case also for DCAT-AP and DCAT since they are required to get a full definition of GeoDCAT-AP terms. For instance, all concepts/properties of GeoDCAT-AP are reused, 81% concepts (18 out of 22), and 84% properties (16 out of 19) of qb, etc. In addition, SKOS vocabulary is indirectly imported since it is already reused by DCAT.
- Indirect Reuse: each data schema component represented with qb vocabulary is linked to a domain concept via `qb:concept` property, when possible (if the domain concept exists). Hence, the domain (meteorological) ontologies are not imported, but rather referenced.

5 A Model for Annotating SYNOP Datasets

The ontologies and vocabularies presented just before were integrated to provide a semantic model for annotating SYNOP datasets. We distinguish two modules of this model: i) *dmo*, the part that refers to the vocabulary representing the descriptive metadata along with the introduction of new concepts allowing to harmonise the semantics of the reused vocabularies. This part is generic enough for accommodating different kinds of domain datasets; ii) *dmo-synop*, the part of the model that relies on *dmo* and that refers to the structural schema of the SYNOP dataset collection (see prefixes in Table 4). This part of the model is the one connected to domain ontologies so that the dataset content be described with domain concepts.

5.1 Modelling principles

Before introducing *dmo* and *dmo-synop*, the adopted modelling principles for constructing them are discussed. It is (strongly) assumed here that all SYNOP datasets share exactly the same structure.

In *dmo*, new concepts have been created in order to accommodate the semantics of the different reused models, as presented in the next section.

For *dmo-synop*, the dataset structure is instead represented as instances (“facts”). Concerning qb, instances of the Data Structure Definition (DSD) (`qb:DataSetDefinition`) have been created to define the structure of one or more datasets. This choice is motivated by the fact that creating a DSD instance allows to define this structure once and then reuse it for each SYNOP file. Users can then be confident that the data structure has not changed and that it is consistent for the entire collection of SYNOP datasets. The same is true for CSVW, because again, the structure of CSVW files is the same for all the SYNOP datasets, which means that the tables and their columns store the same parameters. The meaning of the table columns is represented using instances of `csvw:Column` which refer (via the `dmo:references` property, as introduced in the following) to instances of `qb:ComponentProperty` that, in turn, are linked to domain concepts.

It is worth noting that most of the data in the SYNOP distribution (stored in the tables) are valued as dates, reals, integers or words. But the *dmo-synop* model is not designed to represent this data as semantic entities. Instead, the semantic model provides classes or types to represent which domain concepts or parameters are labelling these columns of these tables. This is why they are represented as instances of `qb:ComponentProperty`.

5.2 Overview

Figure 6 presents an overview of the integration of reused vocabularies and ontologies. The notion of dataset is represented in GeoDCAT-AP and RDF Data Cube with the `dcat:Dataset` and the `qb:Dataset` classes respectively. However, RDF Data Cube distinguishes between a dataset and its fragment `qb:Slice`, while GeoDCAT-AP does not. Considering that meteorological data are continuously produced, they are archived per fragment (slice). For example, the SYNOP dataset is archived as monthly fragments. Following the best practices of web data publishing (Greiner et al. 2017) and data versioning (<https://www.w3.org/TR/dwbp/#dataVersioning>), each fragment *covers a different set of observations about the world and should be treated as a new dataset*, and thus it defines a `dcat:Dataset`. Hence, a new concept was introduced: `dmo:DatasetSlice` is a subclass of both `qb:Slice` (a slice corresponding to a fragment) and `dcat:Dataset`. The `qb:Dataset` concept is rather used to represent the whole dataset.

In many cases, including SYNOP, different datasets have to be integrated, because they describe complementary information. For instance, the station dataset is required because it includes the spatial coordinates of stations (i.e., the spatial localisation of measurements). Since GeoDCAT-AP does not offer the

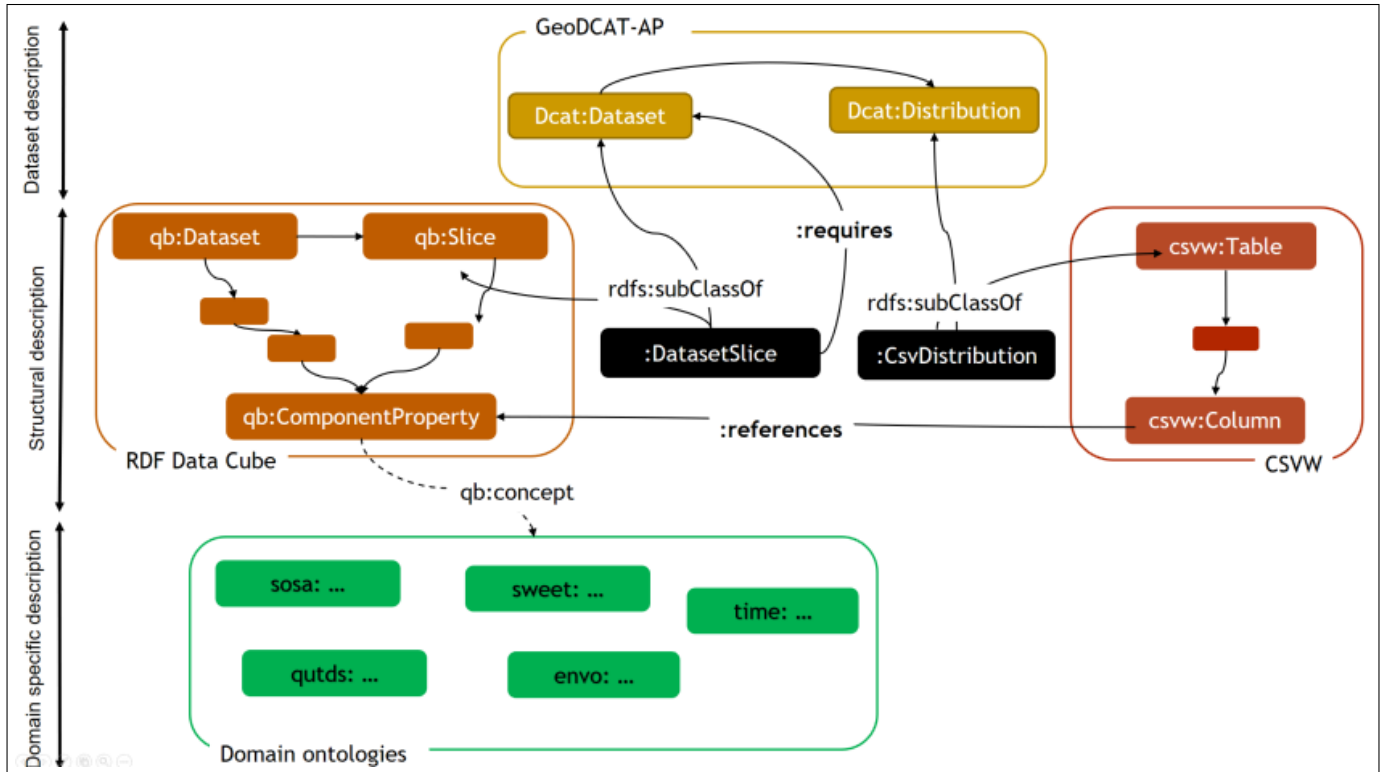


Figure 6 An overview of the integration of the reused vocabularies and ontologies.

possibility to represent such a relationship between two datasets, the `dmo:requires` property is introduced.

In addition, the concept `dmo:CsvDistribution` was introduced, specialising both `dcat:Distribution` and `csvw:Table` as a way to represent distributions with in a CSV format. The relation `dmo:references` was also introduced between a `csvw:Column` and `qb:ComponentProperty` (i.e., `qb:MeasureProperty` or a `qb:DimensionProperty`) to associate columns in the distribution schema to components in the dataset schema (i.e., dimensions and measures). Thus, we make explicit the relationship between structural components (i.e., columns) and data schema components (i.e., measures and dimensions). Furthermore, the data schema components are associated with concepts from the domain ontology, which also makes explicit the semantics of each column, using the `qb:concept` property, as required by the *dmo-synop* presented below.

5.3 *dmo-synop* schema model

Relying on the integration of the different ontologies and vocabulaires (Figure 6), the *dmo-synop* schema represents the dataset structure and its relations with domain ontologies. This model applies to all SYNOP datasets that conform to this structure.

5.3.1 Representing SYNOP dataset structure

Describing structure using RDF data cube (Figure 7). The SYNOP dataset structure is represented via the `dmo-synop:synop_dataset_structure` entity

which is an instance of `qb4st:SpatioTemporalDSD`. `dmo-synop:synop_dataset` represents the SYNOP archive that is related to all its slices (included `:synop_dataset_feb20`) via the `qb:slice` property. `dmo-synop:synop_dataset_structure` includes one spatial dimension `dmo-synop:station_dimension` and three temporal dimensions: `dmo-synop:year_dimension`, `dmo-synop:month_dimension`, and `dmo-synop:date_dimension`. The spatial or temporal nature of a dimension is specified using `qb4st:SpatialDimension` and `qb4st:TemporalDimension`, respectively. Although the station dimension is not directly a geographic coordinate, it is defined as an instance of `qb4st:SpatialDimension` because it provides access to geospatial coordinates contained in the station file. In addition of dimensions, 57 measures (one for each SYNOP file column) are represented as a `qb:MeasureProperty`. Each measurement is associated to its unit of measure (`dmo-synop:unitOfMeasure`). In Figure 7, a fragment of the definition of a measure `:pmer_measure` and an attribute is presented.

Describing structure using CSVW (Figure 8).

The distribution schema is represented by `:synop_file_schema` an instance of `csvw:Schema`. It includes the different columns of the CSV file (e.g., `numer_sta` and `pmer`). For each column, its name (`csvw:name`), its label (`csvw:title`), its data type (`csvw:datatype`), etc. are represented. The foreign key is represented thanks to the instance `dmo-synop:fk` of the `csvw:ForeignKey` concept. It connects the column “`numer_sta`” of the

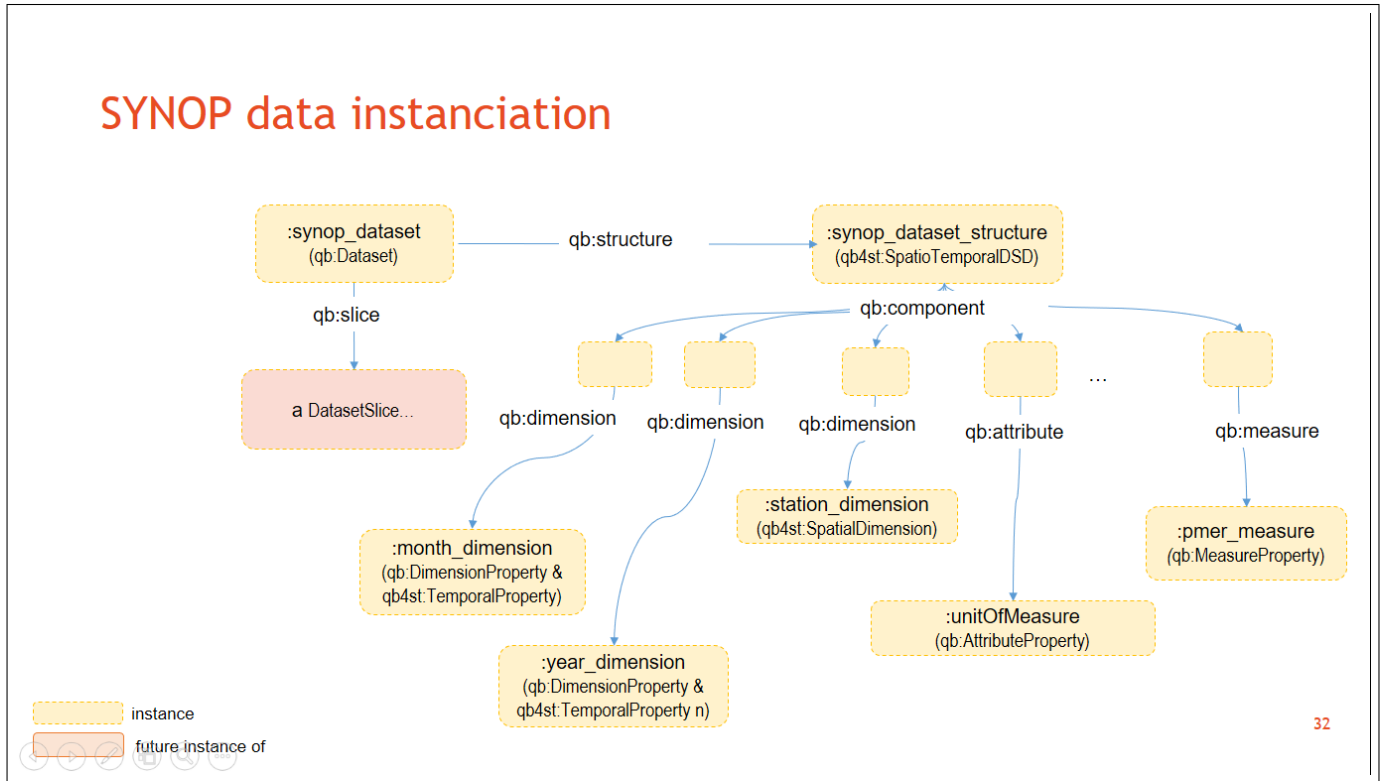


Figure 7 Representing SYNOP dataset schema using RDF data cube.

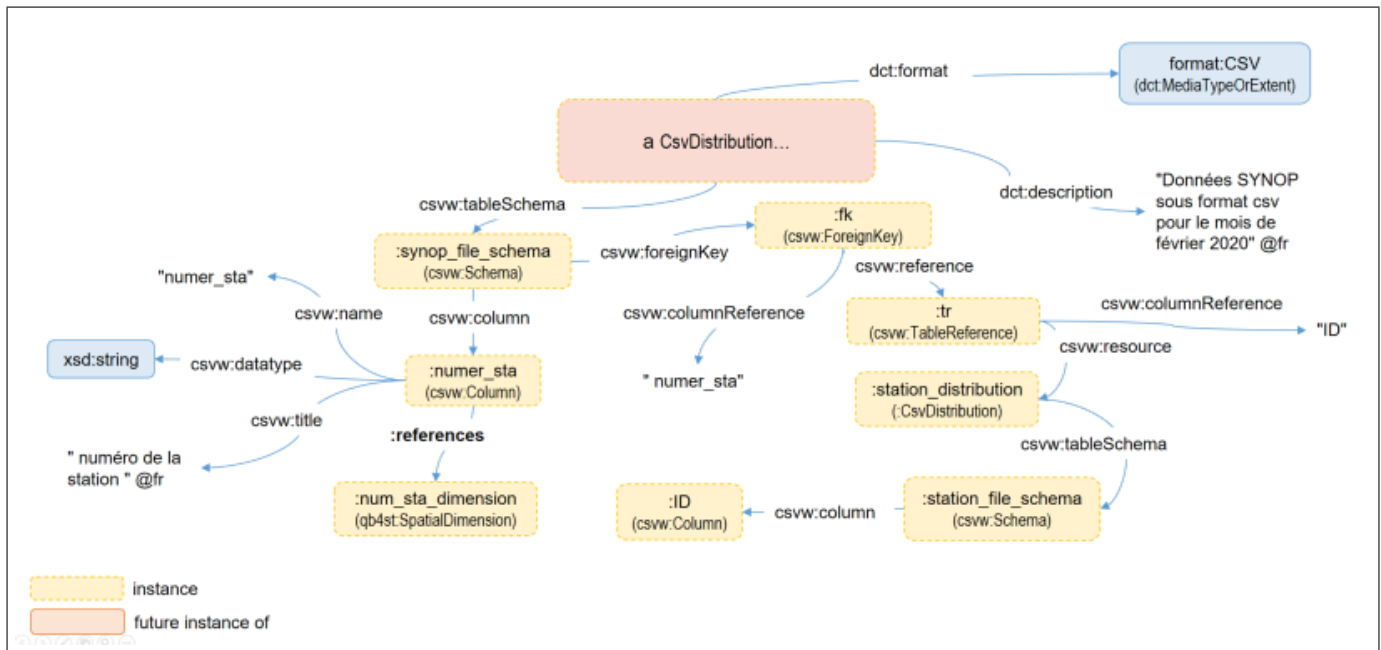


Figure 8 Representing Synop distribution using concepts from CSVW and GeoDCAT-AP.

SYNOP data, to the column “ID” of the station data (:station_distribution) using the instance `dmo-synop:tr` of `csvw:TableReference`. Each column is associated with its corresponding dataset schema component via the property `dmo:references`. Figure 8 presents a fragment where we associate the column `numer_sta` to the dimension `:num_sta_dimension`.

5.3.2 Representing domain concepts

Each dimension or measurement is associated with a concept from domain ontologies via the `qb:concept` property (Figure 9). Concretely, the measure `t` (Figure 1) is linked to the `ENVO:ENVO_09200001` concept which represents the `air temperature` (Figure 10). We have attached two attributes to `qb:Measure`: (i) `dmo-synop:unitOfMeasure` associated

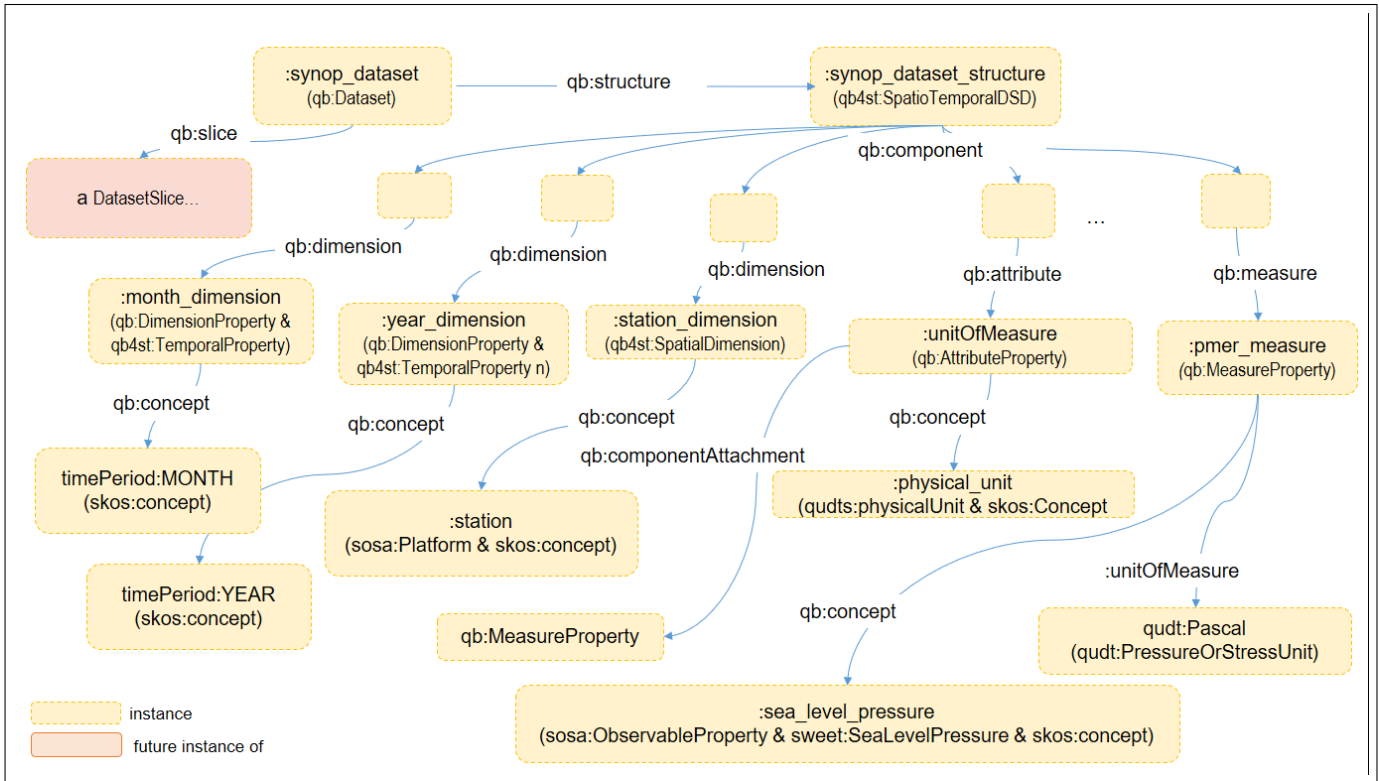


Figure 9 Representing SYNOP data using RDF data cube and domain ontologies.

to `qudts:physicalUnit` to represent the unit of measurement of each `qb:Measure`. This makes it possible to specify that the unit of measurement of `pmer_measure` is `qudt:Pascal`; (ii) `:method_of_measure_attribute` associated to `sosa procedure` to represent the capture procedure.

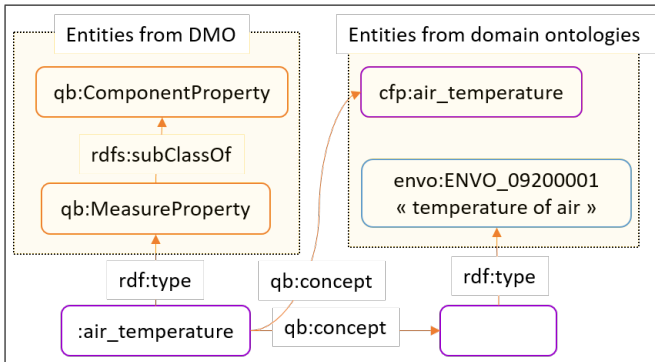


Figure 10 Example of linking a data schema component (a measure) to its domain entities: an individual from CFP (CF standard names ontology) and a concept from ENVO.

6 Annotation of the `synop_feb20` CSV file

The `dmo-synop` model has been used to annotate a SYNOP file, the `synop_feb20` CSV file (identified as `dmo-synop:synop_distribution_feb20`) which is a distribution (instance of `dmo:CsvDistribution`) of

the `dmo-synop:synop_dataset_feb20` dataset, which is itself a `qb:slice` of `dmo-synop:synop_dataset`. Annotating this file with `dmo-synop` means generating metadata at both the dataset and distribution levels. Tables 5 and 6 describe these metadata. At the dataset level, metadata gives information on the dataset, more specifically about the topics of the resource which are here Environment, Climatology/Meteorology/Atmosphere and Geoscientific Information. These Topic Categories are selected in accordance with EN ISO 19115 (`dc:subject property`). Additional metadata at the dataset level are the title in French (“Données SYNOP essentielles OMM pour le mois de février 2020 (France)”) and in English (“WMO SYNOP data for the month of February 2020 (France)”), the creator given as an instance `dmo-synop:meteo_france`, the dataset type according to the INSPIRE standard and referenced by its URI <http://inspire.ec.europa.eu/metadata-codelist/ResourceType/dataset>, a description in natural language (“*Observation data from international surface observation messages (SYNOP) circulating on the global telecommunications system (GTS) of the World Meteorological Organization (WMO). Atmospheric parameters are either measured (temperature, humidity, wind direction and force, atmospheric pressure, amount of precipitation) or observed (sensitive weather, description of clouds, visibility) from the Earth’s surface. Depending on the instrumentation and local features, other parameters may be available (snow depth, ground condition, etc.)*”), the languages used for

that description (French and English), the provenance given as the `dmo-synop:synop_provenance` instance, etc. Furthermore, a reference to the description of its file structure (`dmo-synop:synop_dataset_structure`) is ensured by the `qb:sliceStructure` property. The temporal coverage of this data is given by the `dmo-synop:temporal_coverage` instance related to the start (`dcat:startDate`) and end (`dcat:endDate`) dates of the covered period, which in this case are “2020-02-01” and “2020-02-29” respectively. The W3C representation specifies that this dataset metadata representation can be supplemented by the weather station dataset representation thanks to the `dmo:requires dmo-synop:station_dataset` property.

At the distribution level, metadata give information mainly about the URL https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=90&id_rubrique=32 from which data is accessible (using the `dcat:accessURL` property), the URL https://donneespubliques.meteofrance.fr/?fond=donnee_libre&prefixe=Txt%2FSynop%2FArchive%2FSynop&extension=csv.gz&date=202002 from which data is downloadable (using the `dcat:downloadURL` property), access rights which in that case are “no limitations to public access” according to the INSPIRE standard (`dct:accessRight` property), the kind of license which is here an open license according to the specifications of the French government (`dct:license` property), the description “Données SYNOP pour le mois de février 2020” (`dct:description` property), the format described by an RDF document from Publications Office of the European Union (`dc:format` property), and the file size which is 3735000 bytes (`dcat:byteSize` property).

6.1 Querying the data

To query the instantiated data, the competency questions introduced in Section 3 have been translated into SPARQL queries. All material (competency questions, SPARQL queries, instantiated data) are publicly available at <https://gitlab.irit.fr/melodi/semantics4fair/synop>. In the following, the SPARQL query corresponding to the competency question #13 is presented. It allows for retrieving the information on the column “t” of the distribution file together with its definition from linked domain ontologies.

```
SELECT *
FROM <http://www.openrdf.org/schema/sesame#nil>
FROM <http://purl.obolibrary.org/obo/envo>
WHERE
{
  dmo-synop:synop_dataset_feb20
  dcat:distribution ?distribution .
  ?distribution csvw:tableSchema ?schema .
  ?schema csvw:column ?column .
  ?column csvw:name "t" .
  ?column dmo:correspondsTo ?measure .
  ?measure qb:concept ?concept .
  ?concept rdf:type ?domain .
```

```
}
```

A fragment of the result of the above SPARQL query, serialized in JSON, is presented below:

```
"schema" : {
  "type" : "uri",
  "value" : "dmo-synop#synop_file_schema"
},
"measure" : {
  "type" : "uri",
  "value" : "dmo-synop#air_temperature"
},
"concept" : {
  "type" : "bnode",
  "value" : "node814"
},
"domain" : {
  "type" : "uri",
  "value" : "envo#ENVO_09200001"
},
"column" : {
  "type" : "uri",
  "value" : "dmo-synop#air_temperature_col"
},
"distribution" : {
  "type" : "uri",
  "value" : "dmo-synop#synop_distribution_feb20"
}
}
```

As in the query results, an ENVO domain concept (Figure 10) is associated to column “t”. Such a link allows for further exploring the ENVO ontology in order to obtain additional information on the definition of “t”:

```
envo:ENVO_09200001
  envo:IAO_0000115 "The temperature of some air.",
  oboInOwl:hasExactSynonym "air temperature",
  rdf:type owl:Class,
  rdfs:label "temperature of air",
  rdfs:subClassOf _:node583,
  rdfs:subClassOf envo:ENVO_09200000,
  owl:#equivalentClass _:node579 .
```

7 Evaluation

The evaluation is carried out on the proposed models and on the impact of using it to improve the FAIRness of SYNOP data.

7.1 Model evaluation

The `dmo` and `dmo-synop` were implemented in OWL2 and their consistency was verified with the different reasoners available in Protégé (Hermit, ELK, and Pellet). While several metrics such as OntoMetrics (Lantow 2016) (<https://ontometrics.informatik.uni-rostock.de/ontologymetrics/>), and tools such as Ontology Pitfall Scanner! (Poveda-Villalón et al. 2014) (<http://oops.linkeddata.es/catalogue.jsp>) can be used to evaluate ontology quality, the proposed models rather highly rely on existing (reference) models. Hence, the (content) quality measure here is the consistency when putting together these existing models.

To assess the model compliance to the FAIR principles, the FOOPS (Garijo et al. 2021) online tool

dmo-synop:synop_dataset_feb20	rdf:type	owl:NamedIndividual , dmo:DatasetSlice ;
	qb:sliceStructure	dmo-synop:synop_dataset_structure;
	geodcatap:custodian	dmo-synop:meteo_france ;
	dcat:contactPoint	dmo-synop:meteo_france ;
	dcat:distribution	dmo-synop:synop_distribution_feb20 ;
	dmo:requires	dmo-synop:station_dataset ; "données Synop"@fr , "humidité"@fr , "température"@fr , "vitesse du vent"@fr ;
	dc:creator	dmo-synop:meteo_france ;
	dc:subject	<http://inspire.ec.europa.eu/metadata-codelist/TopicCategory/climatologyMeteorologyAtmosphere>, <http://inspire.ec.europa.eu/metadata-codelist/TopicCategory/environment>, <http://inspire.ec.europa.eu/metadata-codelist/TopicCategory/geoscientificInformation>, <https://www.irit.fr/recherches/MELODI/ontologies/dmo-synop>;
	dc:type	<http://inspire.ec.europa.eu/metadata-codelist/ResourceType/dataset>;
	dct:accrualPeriodicity	<http://publications.europa.eu/resource/authority/frequency/TRIHOURLY>;
	dct:conformsTo	<http://www.opengis.net/def/crs/EPSG/0/4326> ;
	dct:description	"Observation data from international surface observation messages (SYNOP) circulating on the ..."
	dct:language	<http://publications.europa.eu/resource/authority/language/ENG>, <http://publications.europa.eu/resource/authority/language/FRA>;
	dct:provenance	dmo-synop:synop_provenance ;
	dct:spatial	<https://www.geonames.org/countries/FR/>;
	dct:temporal	dmo-synop:temporal_coverage ;
	dct:title	"Données SYNOP essentielles OMM pour le mois de février 2020 (France)."@fr , "WMO SYNOP data for the month of February 2020 (France)"@en ;
	foaf:page	<https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=90&id_rubrique=32>.
dmo-synop:synop_provenance	rdf:type	owl:NamedIndividual , dct:ProvenanceStatement ;
	rdfs:label	"All the measures have been stored/saved by the Météo-France weather stations. The dataset covers ..."@en .
dmo-synop:temporal_coverage	rdf:type	owl:NamedIndividual , dct:PeriodOfTime ;
	dcat:startDate	"2020-02-01"8sd:date .

Table 5 A subset of Synop_feb20 dataset metadata.

dmo-synop:synop_distribution_feb20	rdf:type	owl:NamedIndividual , dmo:CsvDistribution ;
	dcat:accessURL	<https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=90&id_rubrique=32>;
	dcat:downloadURL	<https://donneespubliques.meteofrance.fr/?fond=donnee_libre&prefixe=Txt%2FSynop%2FArchive'%2FSyno&extension=csv.gz&date=202002>;
	dcat:byteSize	3735000 ;
	dc:format	<http://publications.europa.eu/resource/authority/file-type/CSV>;
	dct:accessRights	<http://inspire.ec.europa.eu/metadata-codelist/LimitationsOnPublicAccess/noLimitations>;
	dct:description	"Données SYNOP pour le mois de février 2020"@fr ;
	dct:license	<https://www.etalab.gouv.fr/wp-content/uploads/2014/05/Licence_Ouverte.pdf>;
	csvw:tableSchema	dmo-synop:synop_file_schema .
dmo-synop:synop_license	rdf:type	owl:NamedIndividual , dct:LicenseDocument ;
	rdfs:label	"Synop data is available for public access without any restriction"@en.

Table 6 A subset of Synop_feb20 CSV distribution metadata.

		<i>dmo</i>	<i>dmo-synop</i>
Findable			
F1 (5)	(meta)data are assigned a globally unique and persistent identifier	4/5	2/5
F2 (1)	data are described with rich metadata (R1)	1/1	1/1
F3 (1)	metadata include the described data identifier	1/1	1/1
F4 (2)	(meta)data are registered or indexed in a searchable resource	0/2	0/2
Accessible			
A1 (2)	(meta)data are retrievable by their identifier via standardized protocol	2/2	2/2
A2 (1)	metadata are accessible, even when the data are no longer available	0/1	0/1
Interoperable			
I1 (3)	(meta)data use a formal, accessible, shared, language for KR	3/3	3/3
Reusable			
R1 (5)	meta(data) are richly described with a plurality of relevant attributes	4/5	4/5
R1.1 (2)	(meta)data are released with a clear and accessible data usage license	2/2	2/2
R1.2 (2)	(meta)data are associated with detailed provenance	2/2	2/2

Table 7 FAIR evaluation of the proposed models using the FOOPS! criteria.

has been used. It takes as input an OWL ontology and runs 24 different checks distributed across the 4 FAIR dimensions (Table 7): 9 checks on F (unique, persistent and resolvable URI and version IRI, minimum descriptive metadata, namespace and prefix found in external registries); 3 checks on A (content negotiation, serialization in RDF, open URI protocol); 3 checks on I (references to pre-existing vocabularies); and 9 checks on R (human-readable documentation, provenance metadata, license, ontology terms properly described with labels and definitions). Following these criteria, a score of 79% of FAIRness is obtained for *dmo* against 70% for *dmo-synop*. These score can be further improved by indexing the models in a searchable resource (LOV, for instance). A permanent and unique identifier was created for *dmo* (the generic part of the proposed model) using a web service proposed by the W3C Permanent Identifier Community Group (<https://w3id.org/>): <https://w3id.org/dmo>. For the SYNOP schema model, which must less likely to be reused because of its specificity, we rather made the model accessible using a dereferenced HTTP URI (<https://www.irit.fr/recherches/MELODI/ontologies/DMO/dmo-synop/index-en.html>).

7.2 Semantic data evaluation

In order to evaluate the degree of FAIRness, the framework *FAIR data maturity model* (FAIR data maturity model) proposed by the RDA FAIR Data Maturity Model Working Group RDA (2020) has been chosen. This model is based on three components: i) 41 indicators that measure the state or level of a digital resource according to a FAIR principle; ii) priorities (*essential, important, useful*) associated with the indicators; iii) two evaluation methods: the first consists of assigning each indicator a maturity level between 0 and 4; this method is recommended to data providers (potential indication to improve the FAIRness degree of the data on the provider’s side); the second consists of verifying whether the criterion carried by the indicator is true or false. The indicators were applied first considering the original description of the dataset,

and then considering the instantiation of the proposed model. The reader can refer to Zenodo¹ for a detailed evaluation report.

The first evaluation of the SYNOP dataset consisted in evaluating its original description (without the semantic model). This evaluation resulted in : i) level 0 for principles “F”, “A” and “R”, because at least one essential indicator was not satisfied for each of them; ii) level 1 for principle “I”, because no indicator is essential for this principle (Figure 11 (A)). As it stands, the SYNOP dataset is not FAIR. The data has been **re-evaluated** after generating the semantic metadata that describes it. This metadata significantly improves the FAIRness level, especially for the “I” and “R” principles (Figure 11 (B)).

In fact, one of main concerns when proposing the semantic annotation of this data was to improve their exploitation by non-experts from other scientific communities. With that respect, interoperability is crucial. For this, the proposal meets the main “I” criteria: metadata and data schemas are expressed using in standardised and machine-understandable format, using FAIR-compliant vocabularies; metadata and data refer to other (open) data (here, domain ontologies) and links with these files are made explicit.

Although the re-evaluation of the “F” principle did not show any improvement, the model does allow for the representation of “rich” indexing metadata that satisfy “F2” principle. However, improving the “F” and “A” degree requires satisfying essential indicators that are beyond the capabilities of any semantic model e.g., the generation of persistent and unique identifiers (“F1”), persistent metadata (“A2”), publication of metadata on searchable resources (“F4”), which must be managed by the data publisher (Meteo-FR).

8 Related work

Subsequently, we discuss work related to the main topics addressed in this study, with particular emphasis on work dealing with data characterised by geospatial and temporal dimensions, such as the meteorological data.

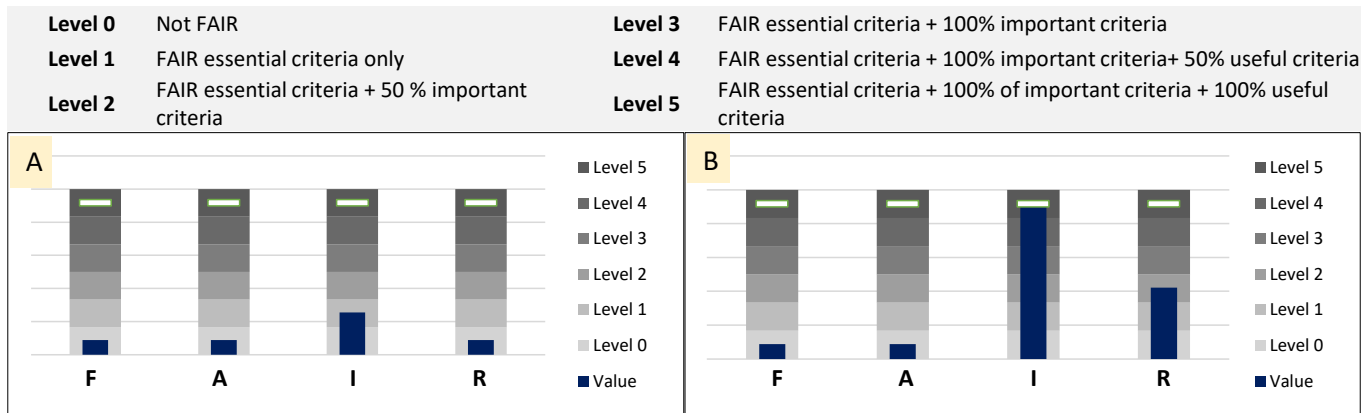


Figure 11 Synop data evaluation: (A) without and (B) with semantic annotation.

8.1 Metadata representation

The importance of sharing geospatial data and describing them with rich metadata has been recognised for decades. Several works have addressed the representation of dataset metadata, and specifically geospatial dataset metadata, even before the emergence of semantic web technologies. Indeed, Kim (1999) published a paper where he compares nine schema for metadata representation of geo-spatial data. The INSPIRE (2007) directive defined a metadata schema, mainly based on the previous standards for describing the European geospatial data on web portals. Later, with the emergence of the semantic web, semantic vocabularies were developed to describe dataset metadata such as Dublin core, VoID, schema.org and DCAT. DCAT-AP, a DCAT application profile, was designed to ensure interoperability between European data portals. GeoDCAT-AP was initially developed to enable interoperability between geospatial data portals implementing the INSPIRE directive, and those implementing DCAT-AP, by developing a set of mappings between the metadata schemes. In December 2020, a new version of GeoDCAT-AP was released, making this vocabulary a full-fledged specification for describing geospatial data catalogs on the Web (GeoDCAT-AP Working Group 2020). Besides these initiatives, several works have also proposed specific metadata vocabularies. Parekh et al. (2004) present a data model ontology and a mechanism for generating ontology-based semantic metadata for dataset publication. Instead of reusing existing vocabularies, the authors proposed their own way of representing metadata on spatial and temporal data identification, content, distribution and presentation forms. In a different way, Frosterus et al. (2011) proposed an extension of the existing VoID vocabulary to cover datasets that are not RDF ones. In our project, non-RDF datasets are limited to tabular datasets.

8.2 Data representation

Several works focused on the semantic representation of geospatial and meteorological data (Lefort et al. 2012, Roussey et al. 2020, Atezing et al. 2013, Patroumpas et al. 2019, Arenas et al. 2018). The proposed models generally are a combination of reference ontologies. In (Lefort et al. 2012) the authors combined qb and SOSA to represent 100 years of temperature data in RDF. Similarly, in (Roussey et al. 2020), the ontologies SOSA, GeoSPARQL, LOCN and QUDTS are reused to represent a meteorological dataset with several measures (temperature, wind speed, etc.). More recently, Yacoubi et al. (2022) proposed to represent in RDF some SYNOP data with a semantic model by reusing a network of existing ontologies (SOSA/SSN, Time, QUDT, GeoSPARQL, and RDF data Cube). In our case, given the characteristics of the meteorological data, the data are not transformed into RDF. Representing all the data in RDF generates a huge graph which is not effective for querying the data (Karim et al. 2020). Moreover, such a choice would require Météo-France to convert all its archives (some of them date back to 1872), which can turn out to be very expensive. Yacoubi et al. (2022) have also defined a set of competency questions (CQ). Contrary to us, those CQ refer to the data level, with the goal to help users find out precise data in the dataset. An example of such CQs is *At what time of the day was the highest value of a weather parameter measured (observed)?*. In our case, the semantic annotation helps the non-expert users to find the right distribution (slice), and in it, the right table and column where such data can be found. Finally, also close to our work, Kremen & Necaský (2019) propose the Semantic Government Vocabulary, based on the different ontological types of terms occurring in Open Government Data. They show how the vocabularies can be used to annotate Open Government Data on different levels of detail to improve “data discoverability”.

8.3 FAIR principles and FAIRness evaluation

As discussed in (Mons et al. 2017, Jacobsen & et al. 2020), semantic web technologies are most consistent with the implementation of FAIR principles. Since the FAIR principles emerged in 2016, several frameworks have been proposed to evaluate the FAIRness degree of a given digital object. The reader can refer to Sun et al. (2022) for a recent survey on the topic. In several of them, the evaluation is performed by answering a set of questions – also called metrics or indicators in some works – or fill in a checklist <https://fairassist.org/> such as the “FAIR Data Maturity Model” (FAIR Data Maturity Model Working Group RDA 2020) or “FAIRshake” (Clarke & et al. 2019). Other works proposed automated approaches for FAIRness evaluation (Wilkinson et al. 2019, Devaraju et al. 2020) based on small web applications that test digital resources against some predefined metrics. Recently, besides evaluating the degree of FAIRness of data, proposals addressed the evaluation of vocabularies and ontologies as well as best practices for implementing FAIR vocabularies and ontologies on the Web (Garijo & Poveda-Villalón 2020, Cox et al. 2021). However very few tools are available to support this evaluation process. One of these few online tools is FOOPS, introduced in Section 7. Another online evaluator, O’FAIRE (for Ontology FAIRness Evaluator) has been recently delivered. It is dedicated to the ontologies and RDF vocabularies accessible from the AgroPortal ontology repository. O’FAIRE is implemented within AgroPortal through 61 questions/tests, among 80% are based on the ontology metadata description (Amdouni et al. 2022).

9 Conclusion

This paper has presented a semantic model to represent different kinds of metadata: those describing a dataset and those on the internal structure of the dataset. This model has been used to represent the data schema of a large collection of datasets in meteorology in France, the SYNOP dataset. This work is part of an approach that aims to make Météo-France data FAIR. An evaluation on the FAIRness of a SYNOP dataset (February 2020) proves the relevance of the proposal in the FAIRisation process. In the future, we have plan for several improvements. First, while the strategy here relied on the fact that large volume of observational data share the same structure, as it is the case for the SYNOP collection, the structural metadata model proposed here fits the specific SYNOP structure. This strategy allowed for reusing this structural schema for semantically annotating the large collection of SYNOP datasets (20 years and 1 dataset per month). However the model could be generalised to represent any spatio-temporal tabular data that are associated to different dimensions and measures, independently of the domain.

Second, in order to facilitate the generation of metadata, the automatic generation of web forms, from SHACL files generated from the ontologies, is also planned. Third, the use of the model to generate metadata and index it on meteorological data portals should be addressed, as well as the possibility of providing a semantic search layer atop the annotated data.

Acknowledgement

This work is funded by the ANR (French National Research Agency) Semantics4FAIR project (<https://www.irit.fr/semantics4fair/>), contract ANR-19-DATA-0014-01.

References

- Amdouni, E., Bouazzouni, S. & Jonquet, C. (2022), O’FAIRE: Ontology FAIRness Evaluator in the AgroPortal semantic resource repository, *in* ‘ESWC 2022 - 19th Extended Semantic Web Conference, Poster and demonstration.’, Hersonissos, Greece.
- Annane, A., Kamel, M., Trojahn, C., Aussenac-Gilles, N., Comparot, C. & Baehr, C. (2021), Towards the fairification of meteorological data: a meteorological semantic model, *in* ‘15th International Conference on Metadata and Semantics Research (MTSR’21)’.
- Arenas, H., Trojahn, C., Comparot, C. & Aussenac-Gilles, N. (2018), ‘Un modèle pour l’intégration spatiale et temporelle de données géolocalisées’, *Revue Int. de Géomatique* **28**(2), 243.
- Atemezing, G., Corcho, O., Garijo, D., Mora, J., Poveda-Villalón, M., Rozas, P., Vila-Suero, D. & Villazón-Terrazas, B. (2013), ‘Transforming meteorological data into linked data’, *Semantic Web* **4**(3), 285–290.
- Benjelloun, O., Chen, S. & Noy, N. F. (2020), Google dataset search by the numbers, *in* ‘Proceedings of the 19th International Semantic Web Conference’, pp. 667–682.
- Buttigieg, P. L., Morrison, N., Smith, B. & et al. (2013), ‘The environment ontology: contextualising biological and biomedical entities’, *J. Biomed. Semant.* **4**, 43.
- Carriero, V. A., Daquino, M., Gangemi, A., Nuzzolese, A. G., Peroni, S., Presutti, V. & Tomasi, F. (2020), ‘The landscape of ontology reuse approaches’, *CoRR*.
- Clarke, D. & et al. (2019), ‘Fairshake: Toolkit to evaluate the fairness of research digital resources’, *Cell Systems* **9**(5), 417–421.
- Cox, S. J. D., Gonzalez-Beltran, A. N., Magagna, B. & Marinescu, M.-C. (2021), ‘Ten simple rules for making a vocabulary fair’, *PLOS Computational Biology* **17**(6), 1–15.

- Cristani, M. & Cuel, R. (2005), 'A survey on ontology creation methodologies', *International Journal on Semantic Web and Information Systems (IJSWIS)* 1(2), 49–69.
- Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J. & White, A. (2020), FAIRsFAIR Data Object Assessment Metrics 0.5, Technical report, Research Data Alliance (RDA). <https://zenodo.org/record/6461229> Accessed 3 May 2022.
- FAIR Data Maturity Model Working Group RDA (2020), 'FAIR Data Maturity Model. Specification and Guidelines'. <https://doi.org/10.15497/rda00050> Accessed 6 May 2022.
- Frosterus, M., Hyvönen, E. & Laitio, J. (2011), Datafinland - A semantic portal for open and linked datasets, in G. Antoniou, M. Grobelnik, E. P. B. Simperl, B. Parsia, D. Plexousakis, P. D. Leenheer & J. Z. Pan, eds, '8th Extended Semantic Web Conference, ESWC, Heraklion, Crete, Greece', Vol. 6644 of *Lecture Notes in Computer Science*, Springer, pp. 243–254.
- Garijo, D. & Poveda-Villalón, M. (2020), 'Best practices for implementing FAIR vocabularies and ontologies on the web', *CoRR*. <https://arxiv.org/abs/2003.13084> Accessed May 2022.
- Garijo, D., Corcho, Ó. & Poveda-Villalón, M. (2021), Foops!: An ontology pitfall scanner for the FAIR principles, in O. Seneviratne, C. Pesquita, J. Sequeda & L. Etcheverry, eds, 'Proceedings of the ISWC 2021 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24-28, 2021', Vol. 2980 of *CEUR Workshop Proceedings*, CEUR-WS.org.
- GeoDCAT-AP Working Group (2020), GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in Europe, Technical report, European Commission. Accessed 6 May 2022.
- Greiner, A., Isaac, A. & Iglesias, C. (2017), Data on the web best practices, Technical report, W3C. Accessed: 2021-09-30.
- Grüniger, M. & Fox, M. (1995), Methodology for the design and evaluation of ontologies, in 'Workshop on Basic Ontological Issues in Knowledge Sharing'.
- Guizzardi, G. (2020), 'Ontology, Ontologies and the "I" of FAIR', *Data Int.* 2(1-2), 181–191.
- Jacobsen, A. & *et al.* (2020), 'FAIR principles: Interpretations and implementation considerations', *Data Intelligence* 2(1-2), 10–29.
- Janowicz, K., Haller, A., Cox, S. J., Le Phuoc, D. & Lefrançois, M. (2019), 'Sosa: A lightweight ontology for sensors, observations, samples, and actuators', *Journal of Web Semantics* 56, 1–10.
- Karim, F., Vidal, M. & Auer, S. (2020), 'Compact representations for efficient storage of semantic sensor data', *CoRR*.
- Kim, T. J. (1999), 'Metadata for geo-spatial data sharing: A comparative analysis', *The Annals of Regional Science* pp. 33:171–181.
- Koesten, L., Simperl, E., Blount, T., Kacprzak, E. & Tennison, J. (2020), 'Everything you always wanted to know about a dataset: Studies in data summarisation', *Int. J. Hum. Comput. Stud.*
- Kremen, P. & Necaský, M. (2019), 'Improving discoverability of open government data with rich metadata descriptions using semantic government vocabulary', *J. Web Semant.* 55, 1–20.
- Lantow, B. (2016), Ontometrics: Putting metrics into use for ontology evaluation, in 'Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KEOD, (IC3K 2016)', INSTICC, SciTePress, pp. 186–191.
- Lefort, L., Bobruk, J., Haller, A., Taylor, K. & Woolf, A. (2012), A linked sensor data cube for a 100 year homogenised daily temperature dataset, in 'Proceedings of the 5th International Workshop on Semantic Sensor Networks', Vol. 904, pp. 1–16.
- Mons, B., Neylon, C., Velterop, J. & *et al.* (2017), 'Cloudy, increasingly fair; revisiting the FAIR data guiding principles for the european open science cloud', *Inf. Serv. Use* 37(1), 49–56.
- Parekh, V., Gwo, J. & Finin, T. W. (2004), Ontology based semantic metadata for geoscience data, in H. R. Arabnia, ed., 'Proceedings of the International Conference on Information and Knowledge Engineering, IKE'04, June 21-24, 2004, Las Vegas, Nevada, USA', CSREA Press, pp. 485–490.
- Patroumpas, K., Skoutas, D., Mandilaras, G. M., Giannopoulos, G. & Athanasiou, S. (2019), Exposing points of interest as linked geospatial data, in 'Proceedings of the 16th International Symposium on Spatial and Temporal Databases', pp. 21–30.
- Poveda-Villalón, M., Espinoza-Arias, P., Garijo, D. & Corcho, Ó. (2020), Coming to terms with FAIR ontologies, in '22nd International Conference Knowledge Engineering and Knowledge Management, EKAW', Vol. 12387 of *Lecture Notes in Computer Science*, Springer, Bolzano, Italy, pp. 255–270.

- Poveda-Villalón, M., Gómez-Pérez, A. & Suárez-Figueroa, M. C. (2014), ‘OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation’, *International Journal on Semantic Web and Information Systems (IJSWIS)* **10**(2), 7–34.
- Raskin, R. (2006), Development of ontologies for earth system science, *in* ‘Geoinformatics: Data to Knowledge’, Geological Society of America.
- Roussey, C., Bernard, S., André, G. & Boffety, D. (2020), ‘Weather data publication on the LOD using SOSA/SSN ontology’, *Semantic Web* **11**(4), 581–591.
- Suárez-Figueroa, M. C., Gómez-Pérez, A. & Fernández-López, M. (2015), ‘The neon methodology framework: A scenario-based methodology for ontology development’, *Appl. Ontology* **10**(2), 107–145.
- Sun, C., Emonet, V. & Dumontier, M. (2022), A comprehensive comparison of automated fairness evaluation tools, *in* K. Wolstencroft, A. Splendiani, M. S. Marshall, C. Baker, A. Waagmeester, M. Roos, R. A. Vos, R. Fijten & L. J. Castro, eds, ‘13th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, SWAT4HCLS 2022, Virtual Event, Leiden, The Netherlands, January 10th to 14th, 2022’, Vol. 3127 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 44–53.
- van den Brink, L. & *et al.* (2019), ‘Best practices for publishing, retrieving, and using spatial data on the web’, *Semantic Web* **10**(1), 95–114.
- Wilkinson, M., Dumontier, M. & Aalbersberg, e. (2016), ‘The FAIR Guiding Principles for scientific data management and stewardship’, *Scientific data* **3**(1), 1–9.
- Wilkinson, M., Dumontier, M. & Sansone, e. (2019), ‘Evaluating FAIR maturity through a scalable, automated, community-governed framework’, *Sc. Data* **6**(1), 1–12.
- Yacoubi, N., Faron, C., Michel, F., Gandon, F. & Corby, O. (2022), A Model for Meteorological Knowledge Graphs: Application to Météo-France Observational Data, *in* ‘22nd International Conference on Web Engineering, ICWE 2022’, 22nd International Conference on Web Engineering, ICWE 2022, Bari, Italy.