Un modèle sémantique en vue d'améliorer la FAIRisation des données météorologiques (PFIA 2021)

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

Un modèle sémantique en vue d'améliorer la FAIRisation des données météorologiques

Amina Annane[1], Mouna Kamel[1], Nathalie Aussenac-Gilles[1], Cassia Trojahn[1], Catherine Comparot[1], Christophe Baehr[2]

[1] Université de Toulouse, IRIT
[2] Météo-France, CNRM

Résumé

Rendre les données météorologiques FAIR pour faciliter leur réutilisation est un enjeu stratégique car ce sont des données essentielles à la recherche scientifique dans de nombreux domaines. Cet article propose un modèle sémantique associant un modèle de métadonnées et un modèle de données pour décrire les données météorologiques d'observation. En effet, la modélisation des (méta)données est une étape essentielle vers leur FAIRisation. Nous utilisons le jeu de données "SYNOP" de Météo-France pour illustrer les difficultés liées à l'accès et à la compréhension de ce type de données, et pour montrer comment le modèle proposé améliore leur adhésion aux principes "F", "I", et "R".

Mots-clés

Données météorologiques, principes FAIR, métadonnées sémantiques.

Abstract

Making meteorological data FAIR in order to ease its reuse is a strategic issue because this data is essential to advance research in many fields. This work proposes a semantic model which combines a metadata model and a data model for describing meteorological observation data. Indeed, modeling (meta)data is an essential step towards their FAIRification. We use the SYNOP open dataset made available by Météo-France to illustrate how difficult data access and understanding can be, and how the use of the proposed model to represent meteorological data improves their compliance with the "F", "I" and "R" principles.

Keywords

FAIR principles, meteorological data, semantic metadata, ontology.

1 Introduction

La météorologie s'appuie sur des modèles mathématiques qui agrègent des données provenant de nombreuses sources, essentiellement de capteurs disposés sur les stations, de satellites ou de radars météorologiques. Les données météorologiques sont nécessaires au développement de bon nombre d'applications, dans différents domaines tels que la météorologie, les transports, l'agriculture, la médecine, etc. Partager ces données est donc devenu un enjeu majeur pour faire des avancées scientifiques dans tous ces domaines.
Or la réutilisation des données météorologiques est difficile car initialement, ces données n'étaient utilisées que par les services météorologiques qui les produisaient : leur codification et interprétation étaient destinées aux météorologues, les usages restant contraints et limités à leurs pratiques. Si l'on veut désormais que des utilisateurs non experts en météorologie puissent les réutiliser, il faut partager, en plus des données elles-mêmes, toutes les métadonnées nécessaires pour les retrouver, y accéder, interpréter correctement, intégrer et analyser (e.g., format, signification, droits d'accès).
Pour répondre à cet enjeu, des initiatives européennes importantes ont été entreprises ces dernières années telles que la directive INSPIRE [1] ou le programme Copernicus [2]. La directive INSPIRE [6], élaborée par la Direction générale de l'environnement de la Commission européenne, impose aux autorités publiques produisant des données géolocalisées, y compris météorologiques, de les rendre publiques. Le programme Copernicus, quant à lui, met à disposition un catalogue de jeux de données d'observation de la terre et météorologiques de différentes origines : satellites, capteurs ou au sol, modèles de prévision météorologiques, etc.
Les principes **FAIR** ont été proposés pour répondre de façon plus globale à la problématique de partage des données en vue de leur réutilisation [21]. Ils consistent en un ensemble de 15 recommandations pour rendre les données faciles à (re)trouver (**F**indable), accessibles (**A**ccessible), intéropérables (**I**nteroperable) et réutilisables (**R**eusable) (Fig. 1). Selon [10], le processus de FAIRisation des données comprend trois phases (1) Pré-FAIRisation : identifier l'objectif de FAIRisation, et analyser les (méta)données (i.e., données et métadonnées), (2) FAIRisation comportant trois étapes (i) développer un modèle sémantique pour représenter les (méta)données , (ii) transformer les (méta)données en une représentation exploitable par la machine (i.e., machine-readable) en utilisant le modèle sémantique développé précédemment, et (iii) rendre les (méta)données disponibles pour les humains et les machines. Enfin, (3) Post-FAIRisation : évaluer si l'objectif fixé dans la phase (1) a été atteint.

1. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008R1205&from=EN
2. https://atmosphere.copernicus.eu/catalogue#/

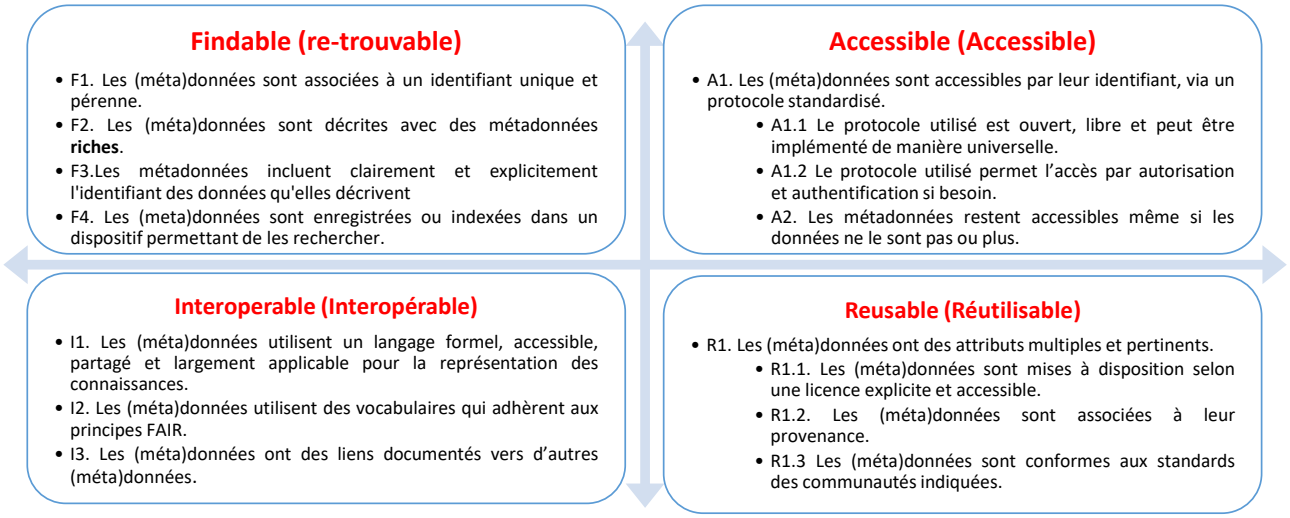

FIGURE 1 – Les principes FAIR (selon [21]).

Dans ce travail, nous nous intéressons à la FAIRisation des données météorologiques d'observation dites "in situ". Il s'agit de mesures directes de différents paramètres (température, vent, humidité, rayonnement, etc.) effectuées par des instruments au sol ou en altitude à partir de lieux prédéfinis (stations d'observation). Après avoir analysé ce type de données (phase 1), nous avons développé un modèle sémantique générique pour représenter les (méta)données (étape (i) de la phase (2)) du processus de FAIRisation. Comme souligné par les auteurs [10], le développement du modèle est l'étape la plus difficile qui prend le plus de temps. Cela vient du fait que cette modélisation nécessite à la fois une expertise métier (expertise en météorologie dans notre cas) et une expertise en modélisation sémantique de données. D'où l'intérêt d'avoir des modèles génériques pour accélérer le processus de FAIRisation. Pour développer notre modèle, nous avons bénéficié de l'expertise métier des météorologues travaillant chez Météo-France, le service officiel de météorologie en France.

Comme de nombreux travaux qui ont traité la FAIRisation des données [22], nous avons opté pour les technologies du web sémantique pour implémenter notre modèle et représenter les (méta)données. Les principes FAIR n'imposent pas l'utilisation de RDF ou de tout autre technologie du web sémantique. Néanmoins, RDF, ainsi que les ontologies formelles, constituent actuellement une solution populaire au problème du partage des connaissances qui répond également aux exigences de FAIR [17].

Le modèle proposé (section 2) est composé de deux sous-modèles pour la représentation des métadonnées et des données, respectivement. Pour que notre modèle soit à son tour FAIR, nous avons réutilisé des vocabulaires FAIR existants. Contrairement aux travaux de [16] et [19], et vu les caractéristiques des données météorologiques (voir section 2.1), nous proposons de ne pas transformer toutes les données en RDF, mais de décrire finement le schéma et la structure des données, et d'expliciter les entités sémantiques incluses dans ces données à l'aide des ontologies de domaines pour permettre une recherche de données plus efficace sans avoir à manipuler un graphe RDF immense [12].
En collaboration avec Météo-France et dans le cadre du projet ANR Semantics4FAIR [3], nous montrons comment la description du jeu de données "DONNÉES SYNOP ESSENTIELLES OMM" (dit SYNOP) par des métadonnées conformes au modèle proposé améliore son adéquation aux principes FAIR. Les données SYNOP et leur évaluation sont présentées en section 3. La section 4 situe notre contribution par rapport aux travaux similaires, avant de conclure par des perspectives en section 5.

2 Modèle Sémantique

Une des implémentations possibles du principes I2 est de développer des ontologies de représentation des métadonnées et de données [10]. Le développement de ces ontologies doit se baser autant que possible sur des ontologies FAIR existantes pour favoriser l'intéropérabilité.
Nous avons développé notre ontologie en suivant quatre étapes principales : (i) Spécification : identifier les besoins qui doivent être couverts par l'ontologie à développer. Pour ce faire, nous avons interviewé trois profils d'utilisateurs de données météorologiques (chercheur en météorologie, biologiste et chercheur en informatique) afin d'identifier un ensemble de questions (i.e., competency questions) auxquelles l'ontologie doit répondre. (ii) Sélection d'ontologies existantes : étudier les ontologies déjà disponibles, les comparer et les confronter aux besoins de représentation identifiés afin de sélectionner celles qui sont les plus pertinentes à réutiliser. (iii) Intégration d'ontologies : préciser les fragments d'ontologies à réutiliser et à combiner afin d'obtenir le modèle final, définir la manière de réutiliser ces fragments (i.e., directe ou indirecte), et vérifier le besoin d'enrichir ou non avec de nouvelles entités. (iv) Evaluation et maintenance : vérifier si le modèle obtenu répond

3. https ://www.irit.fr/semantics4fair/

aux besoin identifiés et l'entretenir. Pour plus d'informations sur la méthodologie adoptée, un rapport détaillé est disponible [4].

Dans ce qui suit, nous commençons par décrire brièvement les caractéristiques des données météorologiques d'observation (un des résultats de l'étape spécification), ensuite nous présentons notre modèle qui est une combinaison de vocabulaires/ontologies de référence complémentaires, en précisant les besoins couverts par chacun d'eux.

2.1 Propriétés des données météorologiques

Données géospatiales. Pour être exploitables, les valeurs de mesures météorologiques doivent être localisées dans l'espace. La localisation est généralement renseignée à l'aide de coordonnées géospatiales (latitude, longitude et altitude). L'interprétation de ces coordonnées dépend du système de coordonnées de référence utilisé (CRS).

Données temporelles. Chaque mesure est effectuée à un moment précis qui doit être noté avec le résultat de la mesure (c'est-à-dire la valeur de la mesure). Comme pour la localisation géospatiale, la localisation temporelle est essentielle à la bonne interprétation des mesures.

Données volumineuses. Les données météorologiques sont produites en continu. Au sein de chaque station, plusieurs capteurs sont installés (thermomètre, baromètre, etc.). Chaque capteur génère plusieurs valeurs de mesure avec une fréquence qui diffère d'une mesure à l'autre selon le besoin (horaire, journalière, trihoraire, etc.).

Données tabulaires. Les données d'observation sont généralement publiées sous forme de tableaux dans lesquels les valeurs de mesure sont organisées selon des dimensions spatio-temporelles. Selon une étude récente de Google, le format tabulaire est le format le plus répandu pour publier des données sur le Web (37 % des jeux de données indexés par Google sont en CSV ou XLS) [1].

2.2 Représentation des métadonnées

Rendre les données FAIR passe par la génération de métadonnées sémantiques qui les décrivent. En effet, 12 des 15 principes FAIR font référence aux métadonnées (Fig. 1) qui doivent de surcroît rester accessibles même si les données ne le sont plus (A2). Ces principes donnent des indications sur les catégories de métadonnées requises : (i) métadonnées descriptives pour l'indexation et la découverte des données (titre, mots-clé, etc.) ; (ii) métadonnées sur la provenance des données (R1.2) ; (iii) métadonnées sur les droits d'accès et les licences d'usage (R1.1).

Notre objectif est donc d'avoir un vocabulaire de métadonnées qui couvre ces différentes catégories, assurant ainsi l'adhésion au principe F2 (métadonnées riches), bien qu'aucune mesure ne permette de quantifier la richesse d'un ensemble de métadonnées.

GeoDCAT-AP. Nous avons choisi le vocabulaire GeoDCAT-AP [5] pour représenter les métadonnées. GeoDCAT-AP est un vocabulaire RDF permettant de décrire les catalogues de données géospatiales sur le WEB. Il a fait l'objet d'une spécification de la recommandation W3C DCAT qui a été développée par Joinup (une plateforme collaborative créée par la Commission européenne et financée par l'Union européenne dans le but de promouvoir l'intéropérabilité des données). Le choix de GeoDCAT-AP est motivé par la richesse des éléments de vocabulaire qu'il inclut. Cette richesse vient du fait que GeoDCAT-AP combine plusieurs vocabulaires/ontologies de référence PROV-O pour représenter la provenance, DCAT, Time, GeoSPARQL, DQV pour représenter des mesures sur la qualité des données, etc. Ainsi, GeoDCAT-AP couvre toutes les catégories de métadonnées citées précédemment. De plus, il offre des propriétés spécifiques et requises pour interpréter correctement des données spatiales [20] comme `dct:spatial` pour décrire la zone géographique concernée par les données, `dct:conformsTo` pour spécifier le CRS utilisé et à choisir dans une liste définie par l'OGC [6], ainsi que `dcat:spatialResolutionInMeters` pour préciser la résolution spatiale des données. De plus, GeoDCAT-AP permet de distinguer la description du jeu de données de la description de ses distributions [7] grâce à la propriété et la classe `dcat:distribution` et `dcat:Distribution`, respectivement. Ce qui permet de préciser par exemple la licence d'utilisation ou de décrire la structure interne d'une distribution spécifique.

2.3 Représentation des données

Comme discuté précédemment, les données météorologiques sont volumineuses et ne cessent de croître. Transformer toutes les archives en RDF ne nous semble pas pertinent pour deux raisons principales :

1. Coût important : transformer toutes les archives de données météorologiques en RDF nécessiterait des moyens conséquents (humains et matériels), ce qui générerait un coût important. Or, selon les principes FAIR, les données doivent être accessibles gratuitement, autant que faire se peut.
2. Efficacité : transformer les archives des données d'observation en RDF générerait un immense graphe RDF qui ne favoriserait pas l'interrogation et l'accès aux données [12].

Cependant, il est essentiel de décrire finement et de manière sémantique le schéma des données et la structure de leurs distributions, pour permettre aux humains d'interpréter et d'explorer correctement les données, et aux machines de les traiter et les interroger automatiquement [15].

Nous avons choisi le vocabulaire RDF data cube pour représenter le schéma des données indépendemment de tout format physique, et le vocabulaire csvw pour représenter la structure des distributions tabulaires car c'est le format le plus populaire. Le modèle peut être enrichi avec d'autres vocabulaires pour d'autres formats tels que JSON-LD [8] et

4. https ://www.irit.fr/semantics4fair/files/onto_report.pdf
5. `https://semiceu.github.io/GeoDCAT-AP/releases/2.0.0/`
6. http ://www.opengis.net/def/crs/EPSG/
7. Une distribution est une représentation spécifique ou une sérialisation du jeu de données
8. https ://www.w3.org/TR/json-ld/

XML [9] si besoin.
De plus, nous réutilisons des ontologies de domaine afin d'expliciter les entités incluses dans les données.

RDF Data Cube (qb). Les données météorologiques sont des données multidimensionnelles, organisées selon les dimensions espace et temps. qb est un vocabulaire dédié à la représentation de ce type de données. Il est une recommandation W3C depuis 2014. Il peut être vu comme un méta-modèle qui permet dans un premier temps de représenter le schéma des données principalement à l'aide des trois sous-classes de `qb:ComponentProperty` : (i) `qb:DimensionProperty` pour spécifier les dimensions, (ii) `qb:MeasureProperty` pour les mesures, et (iii) `qb:AttributeProperty` pour documenter les artefacts comme l'unité de mesure d'une `qb:MeasureProperty` par exemple. La représentation des données se fait dans un deuxième temps par l'instanciation du concept `qb:Observation` en affectant des valeurs aux différents dimensions, mesures, et éventuellement attributs. Comme discuté précédemment, dans notre travail, on se limite à la représentation du schéma de données, d'ailleurs la classe `qb:Observation` n'appartient pas à notre modèle. qb offre la possibilité de représenter les fragments (`qb:Slice`) appartenant au même jeu de données. Par ailleurs, la propriété `qb:concept` permet d'expliciter la sémantique des composants (mesure ou dimension) en les associant aux concepts qui leur correspondent. Ces concepts appartiennent aux ontologies de domaine. En effet, le codomaine de cette propriété est un `skos:Concept`, un type générique que peut avoir tout concept.
L'extension qb4st [10] de qb permet de décrire plus finement les aspects spatiaux et temporels, en spécialisant les classes de qb par exemple `qb4st:SpatialDimension` et `qb4st:TemporalProperty` sont des sous-classes de `qb:ComponentProperty`.

csvw. Comme souligné dans [14], il est essentiel, notamment pour la réutilisation des données, de représenter la structure interne des jeux de données. Le vocabulaire `csvw` [11] proposé par le W3C répond à ce besoin pour les données tabulaires. Dans notre modèle, une distribution dans un format tabulaire sera ainsi instance de `csvw:Table`, dont les colonnes `csvw:Column` seront spécifiées à partir du schéma (`csvw:Schema`) de la table. csvw permet aussi de représenter les relations pouvant exister entre deux distributions (une distribution correspond à un seul fichier), notamment grâce aux notions de clés primaire et étrangère (`csvw:Foreignkey`).

Ontologies de domaine. Les données météorologiques font référence à des concepts du domaine météorologique tels que la température, la vitesse du vent, l'humidité ou tout autre paramètre atmosphérique, les capteurs (e.g., thermomètre, baromètre, etc.), etc. Pour décrire au mieux les jeux de données, nous utilisons les ontologies suivantes :

- **SOSA** (Sensor, Observation, Sample, and Actuator) : ontologie de référence pour la représentation des données issues de capteurs.
- **ENVO** (ontologie de l'environnement) [5] et SWEET (Semantic Web Earth and Environment Technology ontology) [18] : elles incluent les concepts qui représentent les paramètres atmosphériques. Nous les utilisons pour les associer aux mesures représentées avec RDF Data Cube.
- **aws** : ontologie représentant les types de capteurs météorologiques selon les paramètres atmosphériques.
- **QUDT** : ontologie représentant les unités de mesure.

La Fig. 5 liste les acronymes des vocabulaire utilisés.

2.4 Nouvelles propriétés

Lorsqu'un jeu de données X dépend d'un autre jeu de données Y pour être exploité, il est essentiel pour la réutilisation de X d'expliciter cette dépendance. GeoDCAT-AP n'offrant pas cette possibilité, nous avons défini une nouvelle propriété récursive `:requires` concernant le concept `dcat:dataset` pour représenter cette relation.

Par ailleurs, il nous a semblé important de lier la représentation de la structure des données avec csvw à la représentation du schéma de données avec qb. Pour cela, nous avons rajouté la relation `:references` qui permet d'associer à chaque colonne du vocabulaire csvw, une dimension ou une mesure représentée avec qb.

La Fig. 2 présente une vue globale du modèle proposé, sans toutefois, pour des questions de lisibilité, reporter toutes les classes et propriétés utilisées.

3 Cas d'étude : données SYNOP

Aujourd'hui, l'adhésion aux principes FAIR devient un enjeu stratégique pour tout producteur de données voulant promouvoir la réutilisation de ses données. C'est le cas de Météo-France, dont les données sont difficiles à trouver et à réutiliser. En effet, il est surprenant de constater (début 2021) qu'à la requête "normales climatiques en France", les trois premiers résultats fournis par Google ne pointent pas vers le portail de Météo-France, mais vers des sites web concurrents tels que *lameteo.org*, *meteocontact.fr*, qui re-publient les données de Météo-France. Pourtant Météo-France a mis en place un portail web de données.

Afin d'évaluer le degré de FAIRisation de ces jeux de données, nous avons choisi un jeu de données représentatif : le jeu de données SYNOP ("SYNOP ESSENTIELLES OMM"). Il s'agit de données d'observations issues des messages internationaux d'observation en surface circulant sur le Système Mondial de Télécommunication de l'Organisation Météorologique Mondiale (OMM). Ce choix se justifie par le fait que ces données sont ouvertes et gratuites, et concernent plusieurs paramètres atmosphériques mesurés (température, humidité, direction et force du vent, etc.). Ces paramètres sont importants pour de nombreuses études scientifiques. À titre d'exemple et dans le cadre du projet Semantics4FAIR, des chercheurs biologistes du laboratoire

9. https ://www.w3.org/XML/Schema
10. RDF Data Cube extensions for spatio-temporal components
11. https ://www.w3.org/ns/csvw

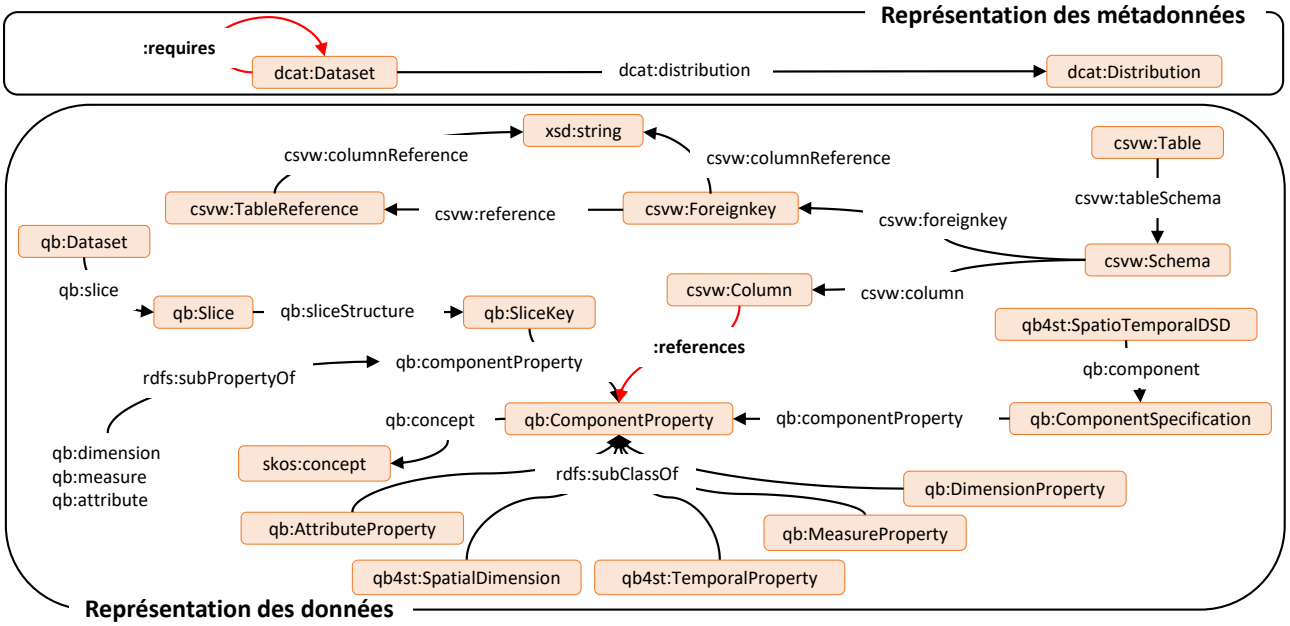

FIGURE 2 – Extrait du modèle montrant les concepts principaux.

numer_sta	date	pmer	ff	t	...
7005	20200201000000	100710	3.200000	285.450000	...
7015	20200201000000	100710	7.700000	284.950000	...
7020	20200201000000	100630	8.400000	284.150000	...
7027	20200201000000	100770	5.500000	285.650000	...
...	...	...	...	...	...

FIGURE 3 – Extrait d'un fichier de données SYNOP.

GET de l'OMP [12] qui étudient la corrélation entre les conditions météorologiques (température et humidité), et la propagation (germination et floraison) d'une plante très allergisante "l'ambroisie", affirment que les données SYNOP auraient pu répondre à leurs besoins, s'ils en avaient eu connaissance et avaient pu accéder à ces données et à leur documentation. Ces données présentent d'autant plus d'intérêt qu'elles sont publiées par tous les états membres de l'OMM, ce qui permettrait d'élargir leur étude.

3.1 Description des données SYNOP

Les données SYNOP publiées en open data [13] sont décrites par sept items : (i) *description* : description résumée du contenu en langage naturel, (ii) *conditions d'accès* : licence Etalab [14], (iii) *moyens d'accès* : téléchargement direct, (iv) *téléchargement* : téléchargement au format csv pour une date donnée, (v) *téléchargement de données archivées* : semblable à l'item précédent, mais pour un mois donné, (vi) *Informations sur les stations* : liste des stations (id_station, nom) accompagnée d'une carte affichant la localisation de ces stations, et (vii) *documentation* : trois liens qui référencent respectivement (a) un fichier *pdf* qui explicite les acronymes (libellé, type, unité de mesure) présents dans l'entête des fichiers de données SYNOP, (b) un fichier *csv* qui fournit les localisations des différentes stations météorologiques de Météo-France (id_station, nom, latitude, longitude, altitude), et (c) un fichier *json* contenant les mêmes informations que le fichier *csv* précédent. Fig. 3 montre un extrait des données SYNOP. Le fichier contient 59 colonnes, les deux premières correspondent au numéro de la station et à la date des mesures effectuées, les 57 autres colonnes aux mesures météorologiques.

Plusieurs problèmes sont alors constatés au vu de tous ces fichiers pourtant disponibles. Chaque observation, pour être exploitable, doit être localisée ; or la localisation de la station est enregistrée dans un fichier de documentation. Les noms des colonnes (acronymes) du fichier de données (Fig. 3) ne sont pas significatifs pour un utilisateur non expert en météorologie. De plus, le fichier de documentation se limite à éclater les acronymes sans donner aucune définition ou information sur la manière dont la mesure a été effectuée (type de capteur ou méthode de calcul utilisée), ou sur le type précis de la mesure (e.g., pour les températures mesurées, s'il s'agit de température de l'air, au sol, à l'abri, etc.). Enfin, la documentation en langage naturel et au format *pdf* (format difficile à traiter automatiquement) ne comporte pas les métadonnées sémantiques conformes exploitables par la machine, pour que ce jeu de données puisse être indexé par des moteurs de recherche de données comme Google dataset search [4].

12. Observatoire Midi-Pyrénées
13. https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=90&id_rubrique=32
14. https://www.etalab.gouv.fr/wp-content/uploads/2014/05/Licence_Ouverte.pdf

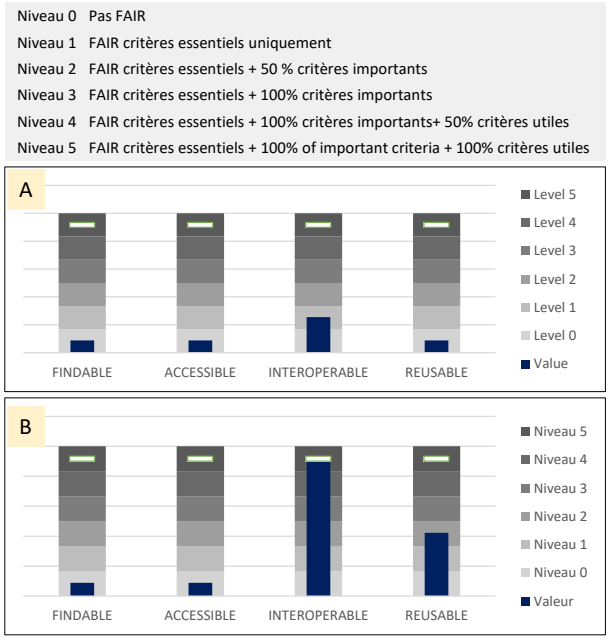

FIGURE 4 – Évaluation du jeu de données SYNOP selon le modèle de maturité de la RDA : (A) avant l'ajout des métadonnées sémantiques, (B) après l'ajout de ces métadonnées.

3.2 Évaluation du degré de FAIRisation des données SYNOP

Plusieurs travaux se sont intéressés à l'évaluation du degré de FAIRisation d'un jeu de données. Parmi ceux-ci, nous avons choisi le modèle *FAIR data maturity model* (Modèle de maturité des données FAIR) proposé par la RDA [7]. Ce modèle d'évaluation comporte : i) 41 indicateurs qui permettent de mesurer l'état ou le niveau d'une ressource digitale selon les principes FAIR ; ii) des priorités (*essentiel*, *important*, *utile*) associées aux indicateurs ; iii) 2 méthodes d'évaluation : la première dédiée aux fournisseurs de données consiste à attribuer à chaque indicateur un niveau de maturité compris entre 0 et 4 (indication permettant d'améliorer le degré de FAIRisation des données côté producteur), la seconde dédiée aux évaluateurs externes, consiste à vérifier si le critère porté par l'indicateur est vrai ou faux.

L'évaluation du jeu de données SYNOP selon ce modèle et en utilisant la seconde méthode d'évaluation a fourni les résultats suivants : (i) le niveau 0 pour les principes F, A et R car au moins un indicateur essentiel n'est pas satisfait pour ces 3 principes, et (ii) le niveau 1 pour le principe I car aucun indicateur n'est essentiel pour ce principe, le niveau 1 est le niveau minimum. Ces résultats nous permettent de conclure que, bien qu'ouvertes et publiées, les données SYNOP ne sont pas FAIR (voir Fig. 4 (A), ainsi que le rapport détaillé de l'évaluation [15]).

15. `https://hal.archives-ouvertes.fr/hal-03197115`

3.3 Instanciation du modèle proposé

Nous avons implémenté le modèle proposé [16] et vérifié sa consistance grâce aux différents raisonneurs implémentés dans Protégé (Hermit, ELK, et Pellet). De plus, en guise d'une première évaluation de sa capacité à représenter les métadonnées de jeux de données météorologiques, nous l'avons instancié avec le jeu de données SYNOP [17]. Pour plus de lisibilité, les figures 6, 7 et 8 ne décrivent qu'une partie des métadonnées qui y sont associées. De plus, les types des instances sont représentés entre parenthèse au dessous de leurs identifiants.

L'archive des données SYNOP se compose d'un ensemble de fichiers mensuels (le plus ancien date de janvier 1996), chaque fichier mensuel ne contenant que les observations réalisées durant ce mois. Nous représentons ici le jeu de données SYNOP du mois de février 2020 (voir Fig. 3).

Fig. 6 montre les différentes métadonnées associées à `:synop_dataset_feb20`, instance des deux classes `dcat:Dataset` et `qb:Slice`. En effet, chaque jeu de données mensuel est représenté par une instance de `dcat:Dataset` conformément aux bonnes pratiques de la publication des données sur le web *"...each dataset covers a different set of observations about the world should be treated as a new dataset."* [18]. On observe que le producteur de données est Météo-France (`dct:Creator`), les données proviennent des Stations Météo-France (`dct:provenance`), les données couvrent la période du 1er février 2020 au 29 février 2020 (`dcat:startDate` et `dcat:endDate`), le système de coordonnées de référence est le `crs:4326` (`dct:conformsTo`), la couverture spatiale est la France (`dct:spatial`), etc. Notons l'utilisation de la nouvelle propriété `:requires` pour représenter que `:synop_dataset_feb20` nécessite le jeu de données `:station_dataset` pour être exploité. Le jeu de données `:synop_dataset_feb20` possède une distribution au format csv représenté par `:synop_distribution_feb20`, instance de `dcat:distribution`. Les métadonnées associées à cette distribution sont décrites sur la Fig. 8. Notons l'utilisation des vocabulaires contrôlés (rectangles bleus) spécifiés par GeoDCAT-AP pour une meilleure intéropérabilité et intégration de données.

Fig. 7 décrit la représentation de la structure de données SYNOP avec qb et les ontologies de domaines. Les rectangles vides correspondent à des noeuds anonymes "blank nodes". Les données SYNOP (tous les fichiers mensuels) ont exactement la même structure, d'où la définition d'une seule structure de données `:synop_dataset_structure` instance de `qb4st:SpatioTemporalDSD`.

Toutes les données SYNOP sont alors représentées par une instance de `qb:Dataset`, et chaque fragment mensuel par une instance de `qb:Slice`, en plus d'être une instance de `dcat:Dataset`. Cette modélisation permet de regrouper toutes les parties d'un même jeu de données.

16. https ://www.irit.fr/semantics4fair/files/MeteOnto.owl
17. https ://www.irit.fr/semantics4fair/files/synop_feb_2020.ttl
18. https ://www.w3.org/TR/dwbp/#dataVersioning

@prefix	URI
base	https://synop-example.ttl#
aws	http://purl.oclc.org/NET/ssnx/meteo/aws#
cat	http://inspire.ec.europa.eu/metadata-codelist/TopicCategory/
country	http://publications.europa.eu/resource/authority/country/
crs	http://www.opengis.net/def/crs/EPSG/0/
csvw	http://www.w3.org/ns/csvw
dcat	http://www.w3.org/ns/dcat#
dct	http://purl.org/dc/terms/
foaf	http://xmlns.com/foaf/0.1/
freq	http://publications.europa.eu/resource/authority/frequency/
geodcatap	http://data.europa.eu/930/
gsp	http://www.opengis.net/ont/geosparql#
lang	http://publications.europa.eu/resource/authority/language/
owl	http://www.w3.org/2002/07/owl#
prov	http://www.w3.org/ns/prov#
qb	http://purl.org/linked-data/cube#
qb4st	http://www.w3.org/ns/qb4st/
qudt	http://qudt.org/1.1/vocab/unit#
qudts	http://qudt.org/1.1/schema/qudt#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
sosa	http://www.w3.org/ns/sosa/
sweet	http://sweetontology.net/propPressure/
time	http://www.w3.org/2006/time#
type	http://inspire.ec.europa.eu/metadata-codelist/ResourceType/
vcard	http://www.w3.org/2006/vcard/ns#
xsd	http://www.w3.org/2001/XMLSchema#
timePeriod	http://publications.europa.eu/resource/authority/timeperiod/
year	http://reference.data.gov.uk/id/year/

FIGURE 5 – Liste des vocabulaires réutilisés

Nous avons défini une dimension spatiale `:station_dimension` et trois dimensions temporelles : `:year_dimension`, `:month_dimension`, et `:date_dimension`. La nature spatiale ou temporelle d'une dimension est précisée à l'aide des concepts du vocabulaire qb4st, respectivement `qb4st:SpatialDimension` et `qb4st:TemporalProperty`. Il est à noter que les dimensions `:year_dimension` et `:month_dimension` qui ne correspondent pas à des colonnes à proprement parler mais qui sont contenues dans la dimension `:date_dimension`, ont été rajoutées pour pouvoir instancier des `qb:Slices`. En effet, selon le vocabulaire qb, chaque instance de `qb:Slice` doit être associé à une instance de `qb:SliceKey` qui définit un sous-ensemble de dimensions à valeurs fixes via la propriété `qb:componentProperty`. Dans notre cas, pour l'instance `:synop_dataset_feb20` qui est un jeu de données mensuelles, les dimensions fixes sont l'année `:year_dimension` et le mois `:month_dimension` qui ont les valeurs `month:FEB` et `:2020`. Bien que la dimension `:station_dimension` ne soit pas directement une coordonnée géographique, elle est définie comme instance de `qb4st:spatialDimension` car elle permet d'accéder aux coordonnées géospatiales contenues dans le jeu de données des stations.

Chaque dimension ou mesure est associée à un concept de domaine, grâce à la propriété `qb:concept`. Ainsi, la mesure `:pmer_measure` (seule représentée ici alors que les 57 mesures ont été instanciées) est liée aux concepts `sosa:observableProperty` et `sweet:SeaLevelPressure` pour en expliciter le sens. L'attribut `:unit_of_measure_attribute`, correspondant à `qudts:physicalUnit` permet de documenter les unités de mesure des `qb:Measure`. Cela permet de spécifier que l'unité de mesure de `pmer_measure` est `qudt:Pascal`. Ainsi le contenu de ce jeu de données peut être indexé par ces concepts de domaines et pas uniquement des mots clés (chaîne de caractères).

En plus de la description du schéma de données avec qb, notre modèle permet de décrire d'une manière sémantique la structure de la distribution synop à l'aide des entités venant de csvw (Fig. 8). `:synop_distribution_feb20` est une instance à la fois de `dcat:distribution` et de `csvw:Table`. Cette distribution est accessible et téléchargeable via les URL (`dcat:accessURL` et `dcat:downloadURL`) reportées en Fig. 8; elle est soumise à une licence (`dct:license`). Enfin, les colonnes (ici `num_sta`, `date`, et `pmer`) de ce fichier sont caractérisées par leur nom (`csvw:name`), leur libellé (`csvw:title`), leur type (`csvw:datatype`) à partir du schéma `:synop_file_schema` (`csvw:tableSchema`), etc. Notons également la représentation de la clé étrangère `:fk` qui relie la colonne "num_sta" du fichier des données SYNOP, à la colonne "ID" du fichier des stations (`:station_distribution`) en passant par l'instance `:tr` de `csvw:TableReference`. Enfin, la nouvelle propriété `:references` est utilisée pour expliciter la sémantique de chaque colonne, en lui associant une mesure ou une dimension. Grâce à la représentation fine qui combine qb et csvw, les données peuvent être traitées et interrogées automatiquement sans être transformées en RDF.

Le jeu de données SYNOP a été réévalué une fois l'ensemble de ces métadonnées sémantiques générées. L'amélioration du degré de FAIRisation peut être observé Fig.4(B), particulièrement pour les principes "I" et "R" . Bien que l'évaluation du principe "F" n'ait pas montré d'amélioration, le modèle permet de représenter des métadonnées d'indexation "riches" qui sont essentielles pour le "F". Pour ce qui est du processus de FAIRisation dans sa totalité pour ce jeu de données SYNOP, les principes "F" et "A" nécessitent de générer des identifiants pérennes et uniques, et de publier les métadonnées générées sur le web. Ce sera la prochaine étape de notre travail.

3.4 Discussion

Dans ce travail nous nous sommes intéressés à une description sémantique des (méta)données météorologiques en vu de les rendre réutilisables, le but ultime des principes FAIR. Cette description est essentielle mais pas suffisante, elle doit être accompagnée d'autres efforts tels que la mise en place d'API d'accès à distance aux données, l'affectation d'identifiants pérennes, la publication et l'indexation des métadonnées générées sur le web, etc.

Certaines métadonnées pertinentes ne sont pas disponibles même si le modèle permet de les représenter, comme les instruments de mesure, les mesures de qualité, etc. Aussi,

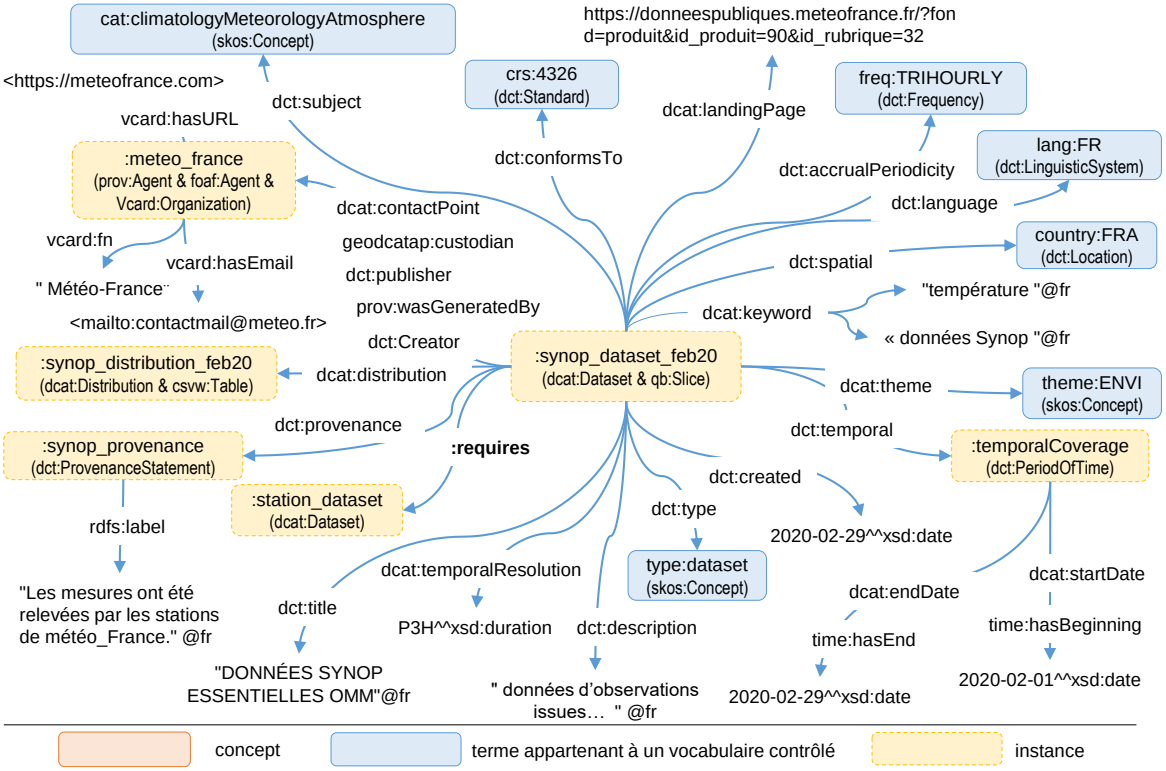

FIGURE 6 – Représentation des données SYNOP pour le mois de février 2020 avec GeoDCAT-AP.

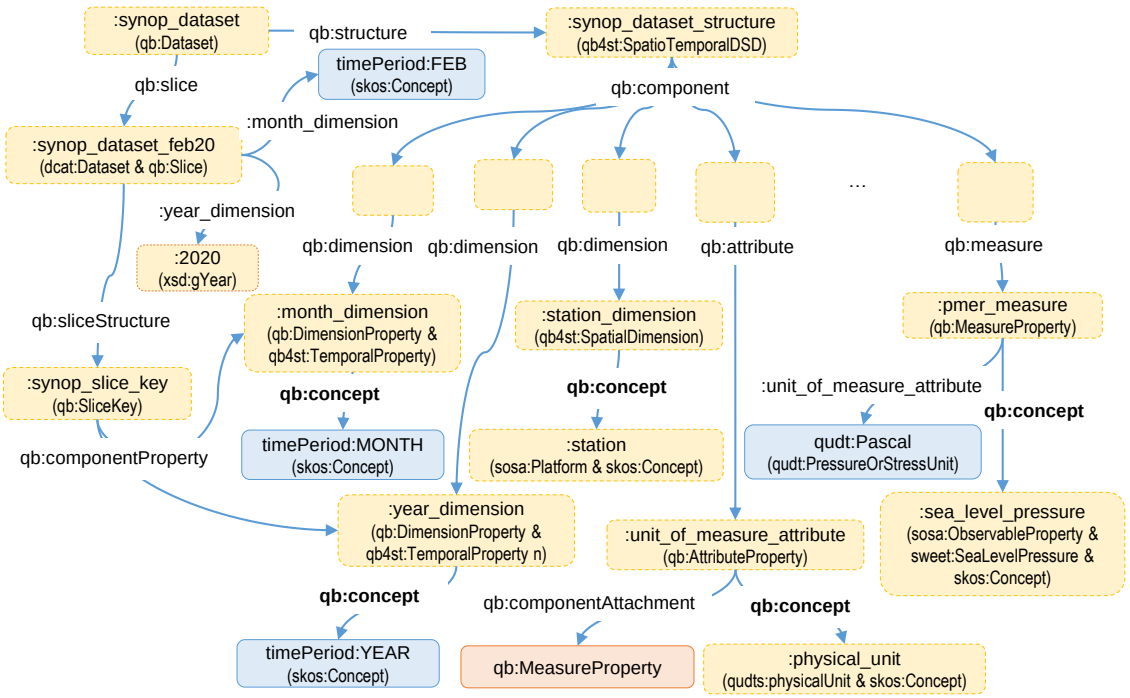

FIGURE 7 – Représentation sémantique des données SYNOP avec RDF data cube, QB4ST et les ontologies du domaine.

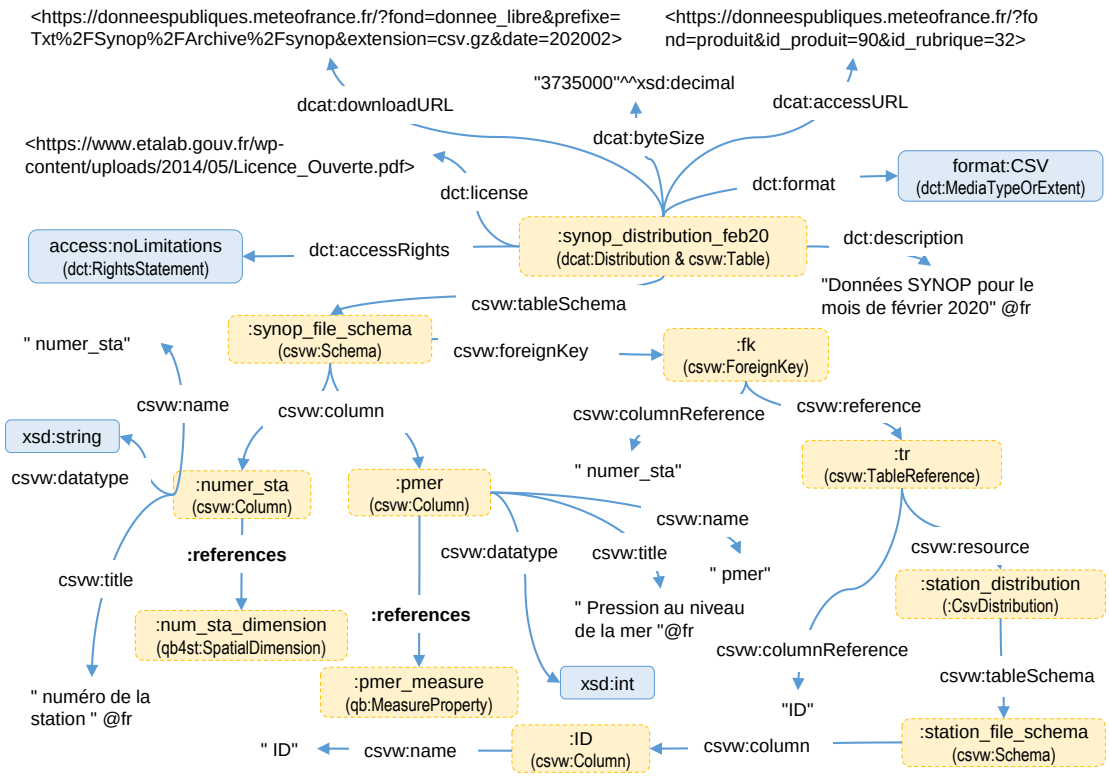

FIGURE 8 – Représentation de la distribution des données SYNOP de février 2020 avec GeoDCAT-AP et CSVW.

seule la dernière localisation (latitude, longitude, et altitude) de chaque station est fournie en documentation, alors que ces stations sont amenées à changer de localisation au fil du temps. Il est important de partager l'historique des localisations des stations dans un format facilement exploitable, pour permettre à l'utilisateur de connaître la localisation exacte de la station au moment de la mesure.
Il est aussi à noter qu'en météorologie, les procédures de mesure, les types de capteurs à utiliser, les normes de qualité, qui sont des métadonnées importantes pour la réutilisation, sont définis par l'OMM. Cette dernière fournit des guides détaillés, tel que "le guide des instruments et des méthodes d'observation météorologiques" [19]. Néanmoins, à notre connaissance, aucune version sémantique de ces guides n'existe. Nous pensons qu'il serait très intéressant de sémantiser ces guides pour une meilleure documentation des données météorologiques en faveur de leur réutilisation.

4 Travaux liés

Approches sémantiques et critères FAIR. Grâce à leur capacité à rendre explicites les types des données, dans un format manipulable par des services, et à produire des métadonnées riches, les ontologies contribuent doublement à rendre les données FAIR, les principes I et R étant sans doute ceux pour lesquels cette contribution est la plus immédiate. Les autres recommandations du I peuvent être assurées en liant les données à d'autres données formalisées (identifiants uniques d'auteurs, de journaux scientifiques, de lieux ou d'organismes) par des liens d'identité [8] ou grâce à l'alignement d'ontologies [23] [2]. Guizzardi redéfinit l'interopérabilité sémantique entre deux systèmes comme l'interconnexion non seulement des données, mais aussi de leur conceptualisation [9]. La réutilisabilité (R) fait référence non seulement à la richesse des métadonnées, aux standards utilisés pour les représenter et à leur accessibilité, mais également à la connaissance de la provenance des données. Pour cela, [8] utilisent l'ontologie PROV-O [20].
Comme recommandé par [11] et mis en oeuvre dans [3], nous enrichissons des vocabulaires standards selon les besoins du cas d'utilisation afin de décrire les métadonnées, y compris la provenance des données, et les types des données selon des concepts du domaine. L'originalité de notre contribution est d'y ajouter des métadonnées décrivant la structure de chaque jeu de données, et les liens entre les éléments de structure et les types de données.

Représentation de données météorologiques. En ce qui concerne le partage des données géolocalisées (comme les données météorologiques), plusieurs schémas de métadonnées existent. L'article [13] compare huit de ces schémas, et ainsi identifie sept critères cruciaux pour la description des données spatiales. Parmi les nombreux vocabulaires possibles pour représenter les données météorologiques et atmosphériques, SWEET [21] et SOSA [22] ainsi que les

19. https://library.wmo.int/doc_num.php?explnum_id=4148

20. https ://www.w3.org/TR/prov-o/
21. https ://www.github.com/ESIPFed/sweet
22. https ://www.w3.org/TR/vocab-ssn/

standards pour représenter des données spatio-temporelles (OWL-Time et Geo-SPARQL) sont parmi les plus utilisés.

5 Conclusion

Nous avons présenté un modèle sémantique générique pour décrire des jeux de données météorologiques d'observation. Ce modèle permet de représenter explicitement et formellement des informations jusque là non disponibles au sujet des jeux de données, à la fois sur la structure et leur contenu en des termes du domaine. Ainsi, les jeux de données vont être plus faciles à trouver et répondent mieux (mais encore incomplètement) aux critères FAIR. La prochaine étape consistera à étudier les spécificités des données issues des modèles statistiques pour enrichir le modèle actuel si besoin. Enfin, nous utiliserons le modèle final pour générer les métadonnées et les indexer dans des portails de données. Nous envisageons aussi de travailler sur la recherche sémantique des jeux de données.

Remerciements

Ce travail est financée par le projet ANR Flash Semantics4FAIR, contrat ANR-19-DATA-0014-01.

Références

[1] O. Benjelloun, S. Chen, and N. F. Noy. Google dataset search by the numbers. In *19th International Semantic Web Conf.*, pages 667–682, 2020.

[2] F. Beretta. A challenge for historical research : making data FAIR using a collaborative ontology management environment (OntoME). *Semantic Web – Interoperability, Usability, Applicability*, 2020.

[3] C. Brewster, B. Nouwt, S. Raaijmakers, and J. Verhoosel. Ontology-based access control for fair data. *Data Int.*, 2 :66–77, 11 2019.

[4] D. Brickley, M. Burgess, and N. F. Noy. Google dataset search : Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference, WWW2019*, pages 1365–1375. ACM, 2019.

[5] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, and S. E. Lewis. The environment ontology : contextualising biological and biomedical entities. *J. Biomed. Semant.*, 4 :43, 2013.

[6] Drafting Team Metadata and European Commission Joint Research Centre. INSPIRE Metadata Implementing Rules : Technical Guidelines based on EN ISO 19115 and EN ISO 19119 - V.1.3., 2013.

[7] FAIR Data Maturity Model Working Group RDA. FAIR Data Maturity Model. Specification and Guidelines, June 2020.

[8] J. D. Fernández, N. Lasierra, D. Clement, H. Mason, and I. O. Robinson. Enabling fair clinical data standards with linked data. In *ESWC*, 2020.

[9] G. Guizzardi. Ontology, Ontologies and the "I" of FAIR. *Data Int.*, 2(1-2) :181–191, nov 2020.

[10] A. Jacobsen and et al. FAIR principles : Interpretations and implementation considerations. *Data Intelligence*, 2(1-2) :10–29, 2020.

[11] A. Jacobsen, R. Kaliyaperumal, L. O. B. da Silva Santos, B. Mons, E. Schultes, M. Roos, and M. Thompson. A generic workflow for the data fairification process. *Data Intelligence*, 2(1-2) :56–65, 2020.

[12] F. Karim, M. Vidal, and S. Auer. Compact representations for efficient storage of semantic sensor data. *CoRR*, abs/2011.09748, 2020.

[13] T. J. Kim. Metadata for geo-spatial data sharing : A comparative analysis. *The Annals of Regional Science*, pages 33 :171–181, 1999.

[14] L. Koesten, E. Simperl, T. Blount, E. Kacprzak, and J. Tennison. Everything you always wanted to know about a dataset : Studies in data summarisation. *Int. J. Hum. Comput. Stud.*, 135, 2020.

[15] P. Kremen and M. Necaský. Improving discoverability of open government data with rich metadata descriptions using semantic government vocabulary. *J. Web Semant.*, 55 :1–20, 2019.

[16] L. Lefort, J. Bobruk, A. Haller, K. Taylor, and A. Woolf. A linked sensor data cube for a 100 year homogenised daily temperature dataset. In *5th Inter. Works. on Semantic Sensor Networks*, volume 904, pages 1–16, 2012.

[17] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. da Silva Santos, and M. D. Wilkinson. Cloudy, increasingly fair ; revisiting the FAIR data guiding principles for the european open science cloud. *Inf. Serv. Use*, 37(1) :49–56, 2017.

[18] R. Raskin. Development of ontologies for earth system science. In *Geoinformatics : Data to Knowledge*. Geological Society of America, 01 2006.

[19] C. Roussey, S. Bernard, G. André, and D. Boffety. Weather data publication on the LOD using SOSA /SSN ontology. *Semantic Web*, 11(4) :581–591, 2020.

[20] L. van den Brink, P. Barnaghi, J. Tandy, G. Atemezing, R. Atkinson, B. Cochrane, Y. Fathy, R. Castro, A. Haller, A. Harth, et al. Best practices for publishing, retrieving, and using spatial data on the web. *Semantic Web*, 10(1) :95–114, 2019.

[21] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1) :1–9, 2016.

[22] M. D. Wilkinson and et al. Interoperability and fairness through a novel combination of web technologies. *PeerJ Comput. Sci.*, 3 :e110, 2017.

[23] J. Wise, A. G. de Barron, A. Splendiani, B. Balali-Mood, D. Vasant, E. Little, G. Mellino, I. Harrow, I. Smith, J. Taubert, et al. Implementation and relevance of fair data principles in biopharmaceutical R&D. *Drug Discovery Today*, 24(4) :933–938, 2019.