Améliorer la FAIRisation des données météorologiques à l'aide de la ressource lexicale INMEVO

Mouna Kamel[1,2], Nathalie Aussenac-Gilles[1], Cassia Trojahn[1,3]

[1] IRIT, CNRS, Université de Toulouse
[2] Université de Perpignan
[3] Université Toulouse 2 Jean Jaurès

prenom.nom@irit.fr

Résumé

*Rendre les jeux de données météorologiques FAIR est un enjeu crucial pour la recherche scientifique. Le modèle **dmo-core** permet, pour les données tabulaires, d'expliciter la sémantique des colonnes en les reliant à des concepts d'ontologies de domaine. Ces concepts étant généralement peu ou pas documentés, nous proposons (1) d'enrichir **dmo-core** afin de pouvoir associer aux colonnes leurs définitions issues d'une ressource lexicale, et (2) de générer une telle ressource, à partir du Vocabulaire Météorologique International de l'Organisation Mondiale de Météorologie (INMEVO).*

Mots-clés

Ressource lexicale, données FAIR, métadonnées sémantiques, données météorologiques.

Abstract

*Making meteorological data sets FAIR is a key issue for scientific research. The **dmo-core** model allows, for meteorological tabular data, to make the semantics of columns explicit by linking them to concepts of domain ontologies. As these concepts are generally little or not documented, we propose (1) to enrich **dmo-core** in order to associate definitions from a lexical resource to columns, and (2) to produce such a resource from the International Meteorological Vocabulary produced by the World Meteorological Organization (INMEVO).*

Keywords

Lexical resource, FAIR dataset, Semantic Metadata, Meteorological data.

1 Introduction

Les données météorologiques sont essentielles pour plusieurs types d'applications, dans de nombreux domaines tels que la météorologie, la climatologie, les transports, l'agriculture, le tourisme ou la médecine. Leur production est le fruit de modèles mathématiques qui intègrent des mesures issues de différentes sources, notamment des stations météorologiques, des satellites ou encore des radars météorologiques. Bien que ces données aient été mises à disposition en tant que données ouvertes sur différents portails, tels que des portails gouvernementaux (e.g. MeteoFrance [1], worldweather [2]), ou des portails associatifs ou privés (e.g. infoclimat [3] ou meteociel [4]), sous licences ouvertes, leur exploitation est plutôt limitée. Une des raisons est qu'elles sont décrites et présentées avec des propriétés pertinentes pour les experts du domaine de la météorologie (producteurs de données), mais qui ne sont pas forcément comprises et réutilisables par d'autres communautés scientifiques. Un moyen de rendre ces données accessibles (non seulement au niveau physique mais également au niveau logique) à des utilisateurs non experts du domaine, est de garantir leur conformité aux principes FAIR (Findability/Faciles à trouver, Accessibility/Accessibilité, Interoperability/Interopérabilité, Reusability/Réutilisabilité) en suivant les 15 recommandations qui leur sont associées [15]. L'adhésion aux principes FAIR s'impose à tout producteur de données qui veut garantir la réutilisation de ses données. Ces recommandations insistent entre autres sur la représentation formelle et sémantique des méta-données à l'aide de vocabulaires ou d'ontologies standards [2].

Dans ce contexte, le modèle sémantique (i.e. ontologie) **dmo-core** présenté dans [12] permet de décrire à la fois les données et les schémas de jeux de données, notamment les jeux de données tabulaires qui représentent la grande majorité des données ouvertes et qui sont disponibles principalement dans les formats CSV ou JSON. Une des spécificités de **dmo-core** est de donner la possibilité aux producteurs de données de sémantiser les colonnes des jeux de données en les reliant à des concepts d'ontologies de domaine. Par exemple, si on considère les colonnes *t* et *pmer* du jeu de données tabulaire SYNOP de Météo-France (voir Figure 1), et qui, pour les producteurs de données et spécialistes du domaine, correspondent à 'température' et 'pression de

1. https://donneespubliques.meteofrance.fr/
2. https://worldweather.wmo.int/en/home.html
3. https://www.infoclimat.fr/
4. https:www.meteociel.fr

la mer' respectivement, **dmo-core** permet de relier les colonnes *t* et *pmer* aux concepts *Temperature* et *SeaLevelPressure* de l'ontologie de domaine SWEET [5] (Semantic Web for Earth and Environment Technology Ontology).

numer_sta	date	pmer	tend	cod_tend	dd	ff	t	td	...
7005	2,02E+13	103180	-80	8	120	1.800000	274.350000	272.750000	...
7015	2,02E+13	103320	0	5	80	4.700000	275.250000	275.150000	...
7020	2,02E+13	102870	-70	8	80	1.300000	280.550000	279.450000	...
7027	2,02E+13	103080	0	0	100	4.200000	275.750000	275.750000	...
7037	2,02E+13	103190	-30	8	130	2.200000	272.250000	272.250000	...
7072	2,02E+13	103320	-20	8	60	1.100000	270.650000	269.550000	...
7110	2,02E+13	102740	10	0	180	0.600000	282.750000	282.650000	...
7117	2,02E+13	102760	-20	8	130	0.500000	281.550000	280.950000	...
7130	2,02E+13	102940	-90	8	110	3.100000	278.350000	278.050000	...
...	...	...	...	...	...	...	...	...	...

FIGURE 1 – Extrait du jeu de données SYNOP de Météo-France.

Force est de constater que les concepts des ontologies de domaine sont le plus souvent peu ou pas documentés. Par exemple, dans l'ontologie SWEET, la définition du concept *SeaLevelPressure* n'est donnée qu'en langue anglaise, alors qu'aucune définition de la notion de *temperature* n'est fournie. De plus, les termes synonymes des labels ne sont pas spécifiés. Or des outils comme FOOPS ! [5] intègrent dans l'évaluation du degré de FAIRisation des ressources sémantiques, notamment au niveau de la Réutilisabilité, les critères de documentation lisible par un humain et l'accès aux définitions des termes utilisés dans la ressource.

Dans l'optique d'améliorer la réutilisabilité (R de FAIR), l'idée est d'offrir la possibilité d'associer, lors de la description des colonnes, des ressources de type thesaurus, dictionnaire, lexique, etc. pour mieux documenter les colonnes des jeux de données tabulaires. Ceci suppose de disposer d'une ressource lexicale dans un domaine donné, et de pouvoir relier cette ressource au modèle d'annotation **dmo-core**. Pour le domaine de la météorologie, des ressources terminologiques existent comme la CF Standard Name Table développée par le groupe de travail Climate and Forecast (CF) Metadata Conventions [6], mais le lexique qui paraît le plus complet et surtout fait office de référence mondiale à notre connaissance, est le Vocabulaire Météorological International (WMO) produit par l'Organisation Météorologique Mondiale (OMM). Si l'on reprend l'exemple de la température, ce lexique fournit la définition suivante, et en quatre langues : *Grandeur physique caractérisant l'agitation moyenne des molécules dans un corps physique.* Bien que ce vocabulaire de l'OMM ait été intégré à la base de données terminologique UNTERM [7] créée par l'Organisation des Nations Unies, la ressource n'est malheureusement pas accessible pour pouvoir l'utiliser dans notre processus de FAIRisation. Le seul format disponible à notre connaissance est le format PDF.

Notre contribution pour améliorer le processus de FAIRisation des données météorologiques est donc double : (1) enrichir le modèle **dmo-core** pour pouvoir intégrer une ressource lexicale, et (2) créer une telle ressource à partir du vocabulaire de l'OMM disponible aujourd'hui au format PDF.

Cet article est organisé de la façon suivante. Après un état de l'art sur les vocabulaires proposés pour représenter les métadonnées (section 2), nous décrivons brièvement section 3 le modèle sémantique **dmo-core** sur lequel s'appuie notre proposition, ainsi que le processus de FAIRisation. Nous montrons section 4 comment enrichir le modèle **dmo-core** en offrant la possibilité d'intégrer une ressource lexicale. La section 5 est dédiée à la construction de la ressource lexicale météorologique INMEVO. Le processus de FAIRisation de données météorologiques à l'aide de **dmo-core** enrichi et du lexique INMEVO est décrit en section 6, et illustré par un exemple d'instanciation du jeu de données SYNOP de Météo-France. La dernière section dresse un bilan de ce travail et en présente les perspectives.

2 Etat de l'art

2.1 Représentation sémantique de schéma de métadonnées

De nombreux vocabulaires ont été proposés pour représenter les métadonnées de jeux de données, dont plusieurs sont devenus des standards de fait : Dublin core [8], VoID, Schema.org, DCAT [9], DCAT-AP. Il en existe des extensions pour pouvoir les adapter à des jeux de données spécifiques, comme GeoDCAT-AP pour les jeux de données géographiques ou StatDCAT-AP pour des données statistiques. D'autres approches utilisent des ontologies pour construire un schéma de métadonnées particulier. Ainsi, Parekh *et al.* [10] présentent un modèle de données et un mécanisme pour générer un schéma de métadonnées basé sur des ontologies. Au lieu de réutiliser les vocabulaires existants, les auteurs proposent leur représentation des métadonnées qui comporte des informations spatiales et temporelles, le contenu, la distribution et la présentation du jeu de données. D'une manière différente, Frosterus *et al.* ont étendu le vocabulaire VoID pour prendre en compte les jeux de données dans un format autre que RDF [4]. Au delà de ces initiatives qui ciblent tout type de jeux de données, plusieurs travaux proposent aussi des vocabulaires spécifiques à des domaines. Dans le domaine des Sciences Sociales et des Humanités du European Open Science Cloud (EOSC), c'est la Data Documentation Initiative [10] (DDI) qui est identifiée comme le principal standard. Cette initiative propose deux schémas XML, DDI-C et DDI-L, particulièrement adaptés aux données d'enquêtes quantitatives en sciences so-

5. https://bioportal.bioontology.org/ontologies/SWEET
6. https://cfconventions.org/Data/cf-standard-names/current/build/cf-standard-name-table.html
7. https://unterm.un.org/unterm2/fr/
8. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
9. https://www.w3.org/TR/vocab-dcat/
10. https://ddialliance.org/learn/what-is-ddi

ciales et comportementales. DDI-L étend DDI-C pour décrire les jeux de données tout au long de leur cycle de vie. Plus récemment, la norme DDI-DCI (DDI Cross-Domain Integration) est adaptée à l'intégration de données provenant de différents domaines de recherche. DDI réutilise des vocabulaires tels que Prov-O, DC-terms, Data Cube ou CSVW et a des liens explicites avec des normes telles que DCAT. Cependant, il n'existe pas de sérialisation OWL de ces modèles.

Dans le domaine de la météorologie, les données étant de format tabulaire, plusieurs propositions ont utilisé RDF Data Cube [11] (qb) combiné à d'autres vocabulaires pour représenter les données d'observation. Dans [7], les auteurs combinent qb et l'ontologie de données de capteur SOSA [12] pour représenter 100 années de données relatives aux températures en RDF. Plus récemment, des données SYNOP ont été représentées en RDF à l'aide d'un modèle sémantique qui réutilise un ensemble d'ontologies existantes (SOSA/SSN, Time, QUDT, GeoSPARQL, et qb) [16].

Notre proposition consiste à enrichir le modèle sémantique **dmo-core** [12] capable déjà de représenter différents types de métadonnées au format tabulaire, dont le schéma de données et la structure du jeu de données.

2.2 Représentation de lexiques sous forme de graphes RDF

La représentation formelle de lexiques sous forme de données liées est considérée comme un enjeu pour leur diffusion, mais aussi pour améliorer l'interopérabilité d'autres ressources qui peuvent ainsi être enrichies de termes et de définitions [3]. Les recherches pour la représentation des lexiques sous forme de graphes RDF ont porté essentiellement sur trois aspects : (i) la définition de langages de représentation des lexiques permettant de les structurer et de les associer à des ontologies ; (ii) la construction de nouvelles ressources sous forme de graphes de connaissances ; ou encore (iii) la conversion de ressources existantes en RDF ou OWL [13]. Parmi les langages de représentation de lexiques, on trouve des recommandations du W3C comme SKOS [14] (Simple Knowledge Organization System) qui permet d'associer différents types de labels à des concepts ; des représentations en OWL de standards utilisés pour des lexiques, comme la représentation OWL du standard lexical SIMPLE [11] ; ou encore des propositions plus riches comme LEMON [8], fruit de plusieurs projets sur la représentation de ressources lexicales décrites par des connaissances linguistiques et liées à des ontologies. Parmi les ressources lexicales construites directement sous forme de graphes de connaissances, on peut citer BabelNet [15], une ressource multilingue basée sur une représentation riche des entrées lexicales, qui intègre WordNet et des éléments extraits de sources telles que Wikipedia [9]. Les premières ressources converties en langages du web sémantique sont WordNet et ses variantes locales (EuroWordnet etc.) traduits en OWL dès 2006 [14]. Depuis sa publication, le vocabulaire LEMON devient une norme pour une représentation riche de lexique. Ainsi, le lexique italien PAROLE SIMPLE CLIPS a été représenté à l'aide de LEMON [6] et publié sous forme de données liées, tout comme plus récemment, le dictionnaire Trésor de la Langue Française (TLF) [1].

Toutefois, peu d'articles traitent du processus de traduction ou de conversion du lexique vers un langage du web sémantique sous l'angle de la sémantisation d'un document pdf, comme c'est le cas dans cet article.

3 Travaux préalables

Nous rappelons brièvement ici les parties du modèle de **dmo-core** sur lesquelles repose notre proposition.

3.1 Le modèle dmo-core

Le modèle sémantique **dmo-core** a été élaboré dans le cadre du projet ANR Semantics4FAIR [16]. La Table 1 liste les espaces de nom et les prefixes associés utilisés dans ce modèle.

TABLE 1 – Espaces de nom des vocabulaires utilisés.

préfixe	espace de nom
dcat	http://www.w3.org/ns/dcat#
qb	http://purl.org/linked-data/cube#
skos	http://www.w3.org/2004/02/skos/core#
csvw	http://www.w3.org/ns/csvw#
prov	http://www.w3.org/ns/prov#
dmo-c	https://w3id.org/dmo#

Le modèle **dmo-core** [12] décrit, en OWL2, un schéma de métadonnées permettant la FAIRisation des jeux de données tabulaires, quel que soit le domaine. Il permet de décrire de manière fine le schéma des données et les structures de leurs distributions, et non de transformer les données en triplets RDF. Ce modèle est basé sur les vocabulaires DCAT [17] (`dcat`), CSVW [18]) et RDF Data Cube [19]), vocabulaires recommandés ou standardisés par le W3C. Tout jeu de données tabulaire, issu d'un catalogue (`dcat:Catalog`) et correspondant à une distribution (`dcat:Distribution`), peut être décrit selon ses différentes propriétés (`qb:MeasureProperty`, `qb:DimensionProperty` ou `qb:AttributProperty`), chacune correspondant à une colonne (`csvw:Column`) d'une table. Une colonne est reliée à une propriété par la relation `dmo-c:references`, et chaque propriété est reliée à un concept SKOS (`skos:Concept`) par la relation `qb:concept`. Ainsi, la sémantique de la colonne est fournie par ce concept qui peut lui-même correspondre ou être typé par un concept d'une ontologie du

11. https://www.w3.org/TR/eo-qb/
12. https://www.w3.org/2015/spatial/wiki/SOSA_Ontology
13. https://www.w3.org/TR/owl-guide/
14. https://www.w3.org/TR/skos-reference/
15. https://babelnet.org/
16. https://www.irit.fr/semantics4fair/index.html
17. https://www.w3.org/TR/vocab-dcat-2/
18. https://www.w3.org/ns/csvw/ (csvw
19. https://www.w3.org/TR/eo-qb/ (qb

domaine. Sur la Figure 7, le modèle **dmo-core** correspond aux concepts et relations en noir, bleu et orange.

3.2 dmo-core utilisé dans le domaine de la météorologie

L'utilisation de **dmo-core** pour un domaine particulier consiste à importer des ontologies de ce domaine. Dans le cas de la météorologie, nous avons importé les ontologies SWEET [20], ENVO [21] et SOSA [22]. Sur la Figure 7, ces ontologies de domaine sont représentées en violet. Le lien entre une colonne du jeu de données tabulaire et un concept d'ontologie de domaine devient effectif lors du processus de FAIRisation rappelé ci-dessous.

3.3 Processus de FAIRisation

Le processus de FAIRisation d'un jeu de données consiste à instancier **dmo-core** après importation des ontologies de domaine, par des propriétés propres à ce jeu de données. Il s'agit alors, pour chaque colonne Col_i de créer les instances suivantes :

1. l'instance INST_Col_i de `csvw:Column` ;
2. l'instance INST_i, instance à la fois de `skos:Concept` et du concept de l'ontologie de domaine ;
3. l'instance $\text{INST\_CompProp}_\text{i}$ (i.e., une dimension, un attribut ou une mesure) de `qb:ComponentProperty`.

Les instances $\text{INST\_CompProp}_\text{i}$ et INST_i sont reliées par la propriété `qb:concept`, et les instances $\text{INST\_CompProp}_\text{i}$ et INST_Col_i par la propriété `dmo-c:references`.

Afin d'illustrer le résultat de ce processus, nous donnons ci-dessous les métadonnées générées pour la colonne `pmer` du jeu de données SYNOP de Météo-France.

```
:sea_level_pressure_col
  rdf:type owl:NamedIndividual, csvw:Column ;
  dmo-c:references  :pmer_measure ;
  csvw:datatype "xsd:int" ;
  csvw:name "pmer" ;
  csvw:title "pression au niveau mer" .

:sea_level_pressure rdf:type
  owl:NamedIndividual ,
  sweet:SeaLevelPressure ,
  skos:Concept ,
  <http://www.w3.org/ns/sosa/ObservableProperty> .

:pmer_measure rdf:type owl:NamedIndividual ,
                       qb:MeasureProperty ;
  qb:concept :sea_level_pressure ;
  :unit_of_measures_attribute
        <http://qudt.org/vocab/unit#Pascal> .
```

4 Enrichissement de dmo-core

Bien que la sémantique des colonnes des jeux de données tabulaires soit explicitée par les concepts des ontologies de domaine, nombreuses sont les ontologies peu documentées, i.e. dont les concepts n'ont pas de définition en langage naturel. Dans ce cas, l'ontologie n'assure pas une bonne compréhension des données. Nous proposons donc d'enrichir **dmo-core** en offrant la possibilité de relier les colonnes des jeux de données tabulaires aux concepts définis dans une ressource sémantique de type thesaurus, dictionnaire ou lexique, qui complète ou apporte des définitions aux colonnes. Les vocabulaires SKOS [23] et OWL [24] (Web Ontology Language) permettent de publier et de rendre accessible tout vocabulaire. De plus, étant respectivement une recommandation et un standard du W3C, toute ressource formalisée à l'aide de ces vocabulaires adhère mieux aux principes FAIR.
Selon le vocabulaire SKOS, une instance de `skos:ConceptScheme` permet de représenter une ressource lexicale, une instance de `skos:Concept` une entrée lexicale, et la relation d'appartenance d'une entrée lexicale à la ressource par la relation `skos:inScheme`. Nous proposons alors d'intégrer ce modèle SKOS dans **dmo-core** en reliant le concept `dmoc:Dataset` (sous-classe du concept `dcat:Dataset`) au concept `skos:ConceptScheme` par la relation `dmo-c:isDocumentedBy`, pour indiquer le lien entre un jeu de données et la ressource lexicale qui peut être utilisée pour documenter les colonnes de ce jeu de données. Le modèle enrichi correspond à l'ensemble de la Figure 7, et l'intégration du lexique correspond aux relations et concepts en vert.

5 INMEVO : une ressource sémantique météorologique

Plusieurs étapes ont été nécessaires à la construction de la ressource lexicale INMEVO : analyse du document produit par l'OMM, identification du modèle des connaissances, puis extraction d'information et représentation des connaissances selon ce modèle.
Bien que le processus d'extraction des connaissances en français soit actuellement encore en cours de développement pour améliorer la qualité de la ressource, une version bêta de INMAVO est accessible à `https://gitlab.irit.fr/melodi/semantics4fair/inmevo`.

5.1 Analyse du document

Le *Vocabulaire météorologique international* [25] est une ressource terminologique publiée par l'*Organisation Météorologique Mondiale* (OMM) en 1966 dans une première version, puis en 1992 dans une version mise à jour et complétée. L'objectif était de normaliser la terminologie et de faciliter la communication entre experts de langues différentes. La version de 1992, la plus récente à ce jour, décrit environ 3500 termes météorologiques avec leurs définitions, dans 4 langues : anglais (EN), français (FR), russe (RU) et espagnol

20. `https://bioportal.bioontology.org/ontologies/SWEET`
21. `https://www.ebi.ac.uk/ols/ontologies/envo`
22. `https://www.w3.org/2015/spatial/wiki/SOSA_Ontology`
23. `https://www.w3.org/TR/swbp-skos-core-spec/`
24. `https://www.w3.org/TR/owl-guide/`
25. `https://public.wmo.int/fr/ressources/meteoterm`

(ES). Ces termes ont été approuvés par différentes organisations comme les membres de l'OMM, l'Aviation Civile Internationale ou la Commission Internationale de l'Eclairage. Bien que ce manuel manque de termes actuels, par exemple relatifs à la télédétection ou aux changements climatiques, il constitue un solide socle de connaissances pour comprendre les termes météorologiques ou interpréter les jeux de données du domaine.

Structure du document. Le document PDF est composé de 802 pages et comporte trois parties : les 19 premières pages sont consacrées à la préface, notice explicative, etc., les 694 pages suivantes sont propres à la terminologie, et les 89 dernières pages sont consacrées aux index (pour le français, l'espagnol et le russe). Chaque page est composée de 2 colonnes, une colonne par langue (pages anglais/français et pages russe/espagnol se succédant alternativement), la page pouvant contenir plusieurs entrées.

Dans ce lexique, chaque terme du vocabulaire météorologique possède une entrée lexicale, les entrées étant triées par ordre alphabétique des termes. Ces entrées lexicales bénéficient de propriétés lexicales, et de propriétés typographiques et dispositionnelles.

Propriétés lexicales. Deux types d'entrées sont à distinguer :

- Entrée de type 1 : c'est une entrée lexicale composée d'un identifiant, du terme défini que l'on appellera *terme descripteur* exprimé dans les 4 langues, d'une définition ou d'une liste de définitions exprimées dans les quatre langues, et pour chaque langue, d'une liste de synonymes en cours d'usage ou désuets s'il en existe. Le nombre de synonymes varie d'une langue à l'autre.
- Entrée de type 2 : c'est une entrée lexicale composée d'un identifiant, du terme défini que l'on appellera *terme descripteur* qui n'est exprimé qu'en anglais, et de la liste entre parenthèses des identifiants des entrées de type 1 ou 2 dont les termes descripteurs sont synonymes.

Dans l'exemple de la Figure 2, les entrées C0780 et C0790 sont de type 1, l'entrée C0800 est de type 2. L'entrée de type 2 C0800 (*clearance*) référence l'entrée C0820 de type 1 (Figure 3) qui possède le terme *clearance* parmi ses synonymes.

Par ailleurs, les entrées de type 2 ne concernent que les termes anglais. Pour les trois autres langues, la référence aux synonymes se fait via l'index qui liste les synonymes associés aux entrées terminologiques (Figure 4).

Il est important de noter que les entrées de type 2, dont le seul objectif est de permettre aux utilisateurs d'accéder à l'index alphabétique de tous les termes du lexique, y compris les synonymes, n'apportent pas de sémantique supplémentaire à celle exprimée au niveau des entrées de type 1. En effet, les termes descripteurs des entrées de type 2 existent déjà en tant que synonymes de termes descripteurs d'entrées de type 1. Par exemple, le terme descripteur *clearance* issu de l'entrée C0800 de type 2, référence l'entrée C0820 de type 1 dont le terme descripteur est *clearing*, qui a pour synonyme *clearance*, *clearing* et *clearance* partageant les mêmes définitions en tant que synonymes.

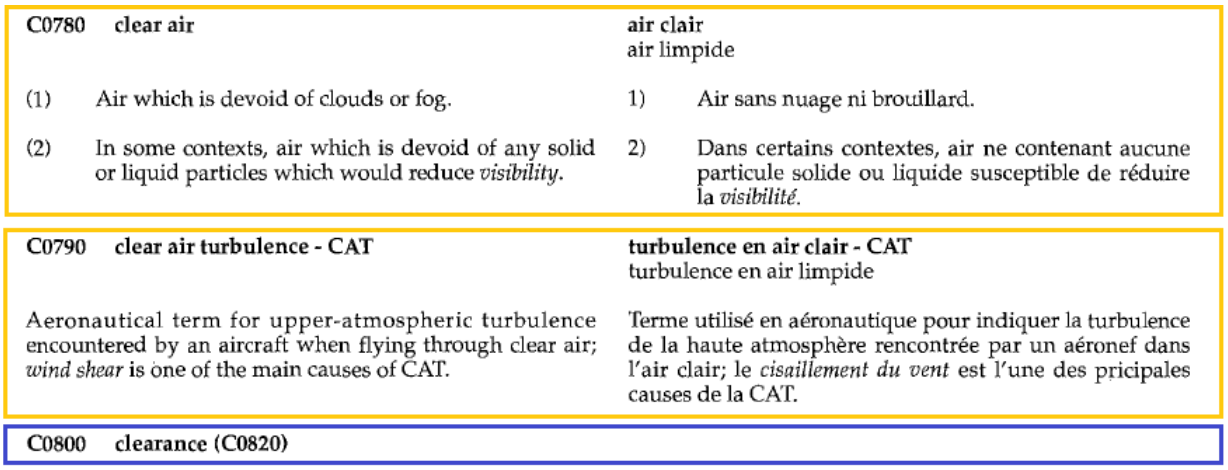
Propriétés typographiques et dispositionnelles. Des règles de mise en forme typographique et dispositionnelle appliquées sur la partie du document qui concerne la description des entrées lexicales permettent d'identifier les différents éléments (voir Figure 2) et d'envisager l'extraction automatique des connaissances exprimées dans cette ressource. Du point de vue typographique, les identifiants des entrées lexicales sont des chaînes de caractères gras répondant à un motif précis (une lettre majuscule suivie de 4 chiffres) ; les termes descripteurs sont en caractères minuscules et gras, les termes synonymes en caractères minuscules et séparés par des virgules, et entre crochets lorsqu'ils sont désuets ; les définitions commencent par une lettre majuscule et se terminent par un point, et sont numérotées lorsqu'un terme possède plusieurs définitions ; les termes apparaissant dans une définition et correspondant eux-mêmes à des entrées lexicales sont en italique ; les références aux concepts équivalents sont entre parenthèses.

Du point de vue dispositionnel, l'identifiant et le terme descripteur correspondant sont séparés par un caractère de tabulation ; les synonymes sont alignés verticalement aux termes descripteurs. À noter que la correspondance entre un identifiant et son terme descripteur en anglais est matérialisée par deux unités lexicales adjacentes dans le texte, alors que la correspondance avec les termes descripteurs dans les autres langues ne se fait que sur des critères dispositionnels.

Le processus d'extraction et de formalisation des connaissances que nous avons mis en œuvre est basé sur l'ensemble de ces règles. Dans cette étude, nous nous sommes limités à la représentation des connaissances exprimées en anglais et en français (présentes dans les pages paires) pour les raisons suivantes : la qualité des outils de conversion du format PDF au format texte est étroitement liée à la langue, et faute de connaissances en espagnol et en russe, nous n'aurions pas pu préjuger de la qualité des conversions. La méthodologie décrite ici pourra cependant être appliquée à la représentation des connaissances exprimées en russe et en espagnol une fois ces contraintes levées.

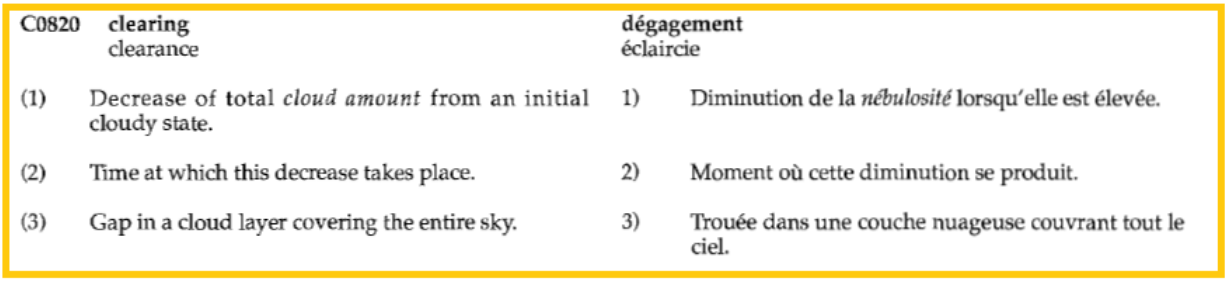
5.2 Représentation sémantique en SKOS : modèle IMV

Nous avons formalisé les entrées de type 1 du lexique de l'OMM et leurs propriétés à l'aide du modèle RDF et des vocabulaires SKOS et OWL. Chaque entrée lexicale de type 1 est représentée sous forme d'un concept SKOS, ayant pour URI l'identifiant de l'entrée lexicale. Ce concept SKOS est relié à plusieurs



C0780	**clear air**	**air clair** air limpide	
(1)	Air which is devoid of clouds or fog.	1)	Air sans nuage ni brouillard.
(2)	In some contexts, air which is devoid of any solid or liquid particles which would reduce *visibility*.	2)	Dans certains contextes, air ne contenant aucune particule solide ou liquide susceptible de réduire la *visibilité*.
C0790	**clear air turbulence - CAT**	**turbulence en air clair - CAT** turbulence en air limpide	
Aeronautical term for upper-atmospheric turbulence encountered by an aircraft when flying through clear air; *wind shear* is one of the main causes of CAT.		Terme utilisé en aéronautique pour indiquer la turbulence de la haute atmosphère rencontrée par un aéronef dans l'air clair; le *cisaillement du vent* est l'une des pricipales causes de la CAT.	
C0800	**clearance (C0820)**		

FIGURE 2 – Exemple d'entrées lexicales de types 1 et 2 (page 110 du document).



C0820	**clearing** clearance	**dégagement** éclaircie	
(1)	Decrease of total *cloud amount* from an initial cloudy state.	1)	Diminution de la *nébulosité* lorsqu'elle est élevée.
(2)	Time at which this decrease takes place.	2)	Moment où cette diminution se produit.
(3)	Gap in a cloud layer covering the entire sky.	3)	Trouée dans une couche nuageuse couvrant tout le ciel.

FIGURE 3 – Entrée lexicale C0820.

chaînes de caractères : le libellé de l'entrée (terme descripteur), ses éventuels synonymes actuels et désuets, par les relations `skos:prefLabel`, `skos:altLabel` et `skos:hiddenLabel` respectivement. On lui associe également ses définitions (multiples et dans les différentes langues) par la relation `skos:definition`. Le lien sémantique existant entre un concept et les concepts intervenant dans ses propres définitions (termes en italique qui correspondent à des entrées lexicales) est représenté par la relation `skos:related`.

Quant aux entrées de type 2, elles ne seront pas formalisées. Comme dit précédemment, ces entrées utiles pour constituer un index alphabétique, ne font que référencer des termes descripteurs déjà présents en tant que synonymes dans des entrées de type 1 (référencées entre parenthèses), qui possèdent les définitions. La Figure 5 présente le modèle IMV.

La section suivante mentionne les étapes nécessaires à l'extraction des connaissances du lexique de l'OMM au format PDF, et à la génération de la ressource lexicale INMEVO selon le modèle IMV.

5.3 Extraction et représentation des connaissances

La mise en forme typographique et dispositionnelle du document (notamment le parallélisme des paragraphes présents dans les colonnes) est primordiale pour caractériser les différents éléments du lexique à mettre en correspondance. Pour l'exploiter, différents outils de conversion de document PDF au format texte ont été testés, comme les bibliothèques python *Tesseract* et *pdfminer*, ou encore des logiciels en ligne comme *Adobe*, *OnlineOCR*, *PDFConvert*. Chacun de ces outils présente des inconvénients liés à la perte d'éléments typographiques et dispositionnels, comme les caractères accentués pour le français, le gras ou l'italique, ou encore en fusionnant les 2 colonnes en une seule (le plus souvent l'une au-dessous de l'autre). Nous avons finalement opté pour le convertisseur en ligne *PDFconverter* qui nous a semblé le plus fiable au regard du taux d'erreurs observé. Cet outil fournit un document au format Word dans lequel texte, typographie (caractères accentués, gras et italiques) et indices dispositionnels (tabulations, retours à la ligne, parallélisme des deux colonnes, etc.) sont globalement conservés. Néanmoins, des pertes de mise en forme ont été observées par endroit, et aucune bibliothèque exploitant des documents Word de notre connaissance n'ont permis d'exploiter les deux colonnes en parallèle.

De fait, deux pages Word, une correspondant à la colonne EN et l'autre à la colonne FR, ont été produites à partir de chaque page Word issue du convertisseur. Au final, nous avons obtenu 694 pages EN et 694 pages FR à exploiter, les ièmes pages FR et EN contenant les mêmes entrées lexicales. Cette réorganisation du corpus présente l'avantage de minimiser la propagation

français		russe		espagnol	
L0530	éclair	P0180	прошедшая погода	A0280	acdar
C0820	éclaircie	C0820	прояснение	C0820	aclaramiento
H0340	éclair de chaleur	P0940	прямая связь	A0220	aclimatación
S0940	éclair diffus	P0930	прямая связь одной точки с несколькими	C1210	acondicionamiento de aire

FIGURE 4 – Synonymes de l'entrée C0820 via les index.

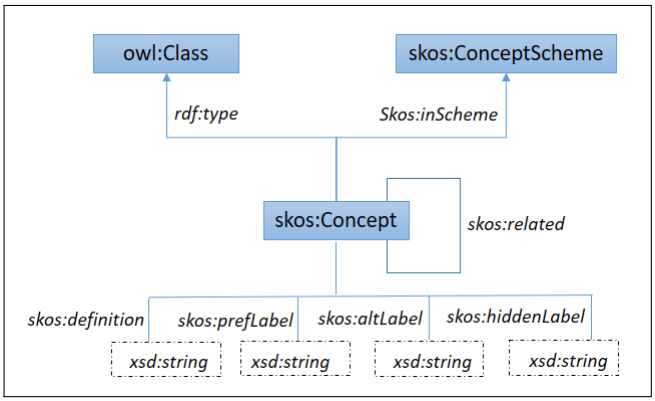


FIGURE 5 – IMV : Modèle sémantique pour la représentation du lexique de l'OMM.

des erreurs, lorsqu'il y en a, les erreurs se limitant à la seule page en cours de traitement.

Nous avons tout d'abord exploité les pages EN, les identifiants étant directement associés aux termes anglais. Nous donnons ci-dessous l'exemple de l'entrée lexicale C0780 du lexique de l'OMM, extraite et représentée selon le modèle IMV décrit Figure 5.

```
inmevo:C0780 rdf:type skos:Concept ;
  skos:prefLabel "clear air" @en ;
  skos:definition "Air which is devoid of clouds
    or fog." @en ;
  skos:definition "In some contexts, air which is
      devoid of any solid or liquid particles
      which would reduce oisibility." @en ;
  skos:related inmevo:V0390 .
```

Les pages en français ont été exploitées dans un deuxième temps. Les termes descripteurs n'étant plus associés à leurs identifiants dans les pages en FR, et l'OCRisation n'ayant pas permis de conserver un strict parallélisme (notamment en introduisant des sauts de lignes additionnels) entre les pages en EN et en FR, les principes d'extraction de connaissance mis en œuvre pour les pages en EN n'ont pas pu être appliqués.
Un moyen de retrouver la correspondance entre paragraphes EN et FR est de calculer une similarité sémantique entre eux. La méthode éprouvée dans cette version bêta de la ressource INMEVO repose sur une mesure de similarité calculée entre un paragraphe EN et la traduction en anglais du paragraphe FR. Pour cela, nous avons eu recours à l'API DeepL [26] pour la traduction, et à la distance de Levenshtein pour évaluer la similarité. Cette méthode a fourni de bons résultats du fait que la correspondance était recherchée entre les entrées lexicales figurant sur une seule page du corpus, et non pas sur la totalité du corpus.
Nous donnons ci-dessous l'exemple de l'entrée C0780 enrichie par les termes et définitions en français :

```
inmevo:C0780 rdf:type skos:Concept ;
  skos:prefLabel "clear air" @en ;
  skos:prefLabel "air clair" @fr ;
  skos:definition "Air which is devoid of clouds
    or fog." @en ;
  skos:definition "In some contexts, air which is
      devoid of any solid or liquid particles
      which would reduce oisibility." @en ;
  skos:definition "Air sans nuage ni
      brouillard." @fr ;
  skos:definition "Dans certains contextes, air ne
      contenant aucune particule solide ou liquide
      susceptible de réduire la visibilité." @fr ;
  skos:altLabel ""air limpide " @fr ;
  skos:related inmevo:V0390 .
```

Au final, la ressource INMEVO [27] est représentée par une instance `inmevo:INMEVO` de `skos:ConceptScheme`, par l'ensemble des instances `inmevo:A0010`, `inmevo:A0020`, ... de `skos:Concept` représentant les entrées lexicales, chacune d'elles étant reliée à `inmevo:INMEVO` par la relation `skos:inScheme`.

```
inmevo:INMEVO rdf:type skos:ConceptScheme .

inmevo:C0780 rdf:type skos:Concept ;
  skos:prefLabel "clear air" @en ;
  ...
  skos:related inmevo:V0390 ;
  skos:inScheme inmevo:INMEVO.

inmevo:C0790 rdf:type skos:Concept ;
  skos:prefLabel "clear air turbulence – CAT" @en ;
  ...
  skos:inScheme inmevo:INMEVO.

...
```

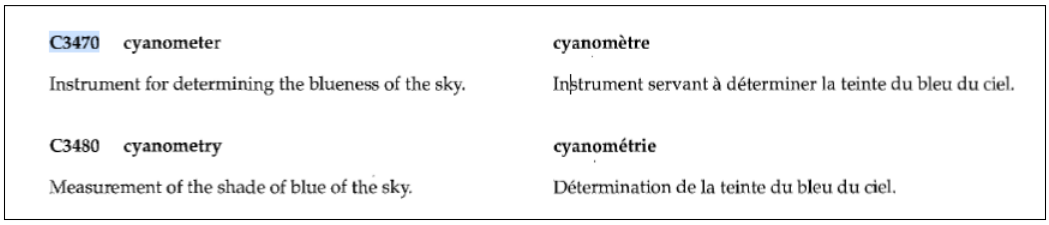
5.4 Evaluation de INMEVO

TABLE 2 – Evaluation de la ressource INMEVO. Les pourcentages indiquent les taux d'extractions correctes.

	Entrée de type 1
effectif	30
label@en	96%
label@fr	70%
def@en	96%
def@fr	86.6%
concepts reliés	84.2%
ocerisation @en	63%
ocerisation @fr	61%

26. https://www.deepl.com/

27. https://w3id.org/inmevo/



C3470 cyanometer	**cyanomètre**
Instrument for determining the blueness of the sky.	Instrument servant à déterminer la teinte du bleu du ciel.
C3480 cyanometry	**cyanométrie**
Measurement of the shade of blue of the sky.	Détermination de la teinte du bleu du ciel.

FIGURE 6 – Exemple d'entrées lexicales ayant conduit à des erreurs lors du processus d'extraction.

Nous avons sélectionné de façon aléatoire 30 entrées de type 1 de la ressource INMEVO, et les avons comparées manuellement aux entrées correspondantes du lexique de l'OMM. Cette comparaison a porté sur l'exactitude des labels extraits en français et en anglais, des définitions extraites en français et en anglais, des liens entre concepts (exprimés dans les définitions), ainsi que des erreurs d'OCRisation. Les résultats sont présentés Table 2.

Les erreurs générées lors du processus d'OCRisation sont pour la majorité dues à une mauvaise reconnaissance de caractères accentués pour le français, et aux caractères spéciaux utilisés dans les unités de mesure ou les formules mathématiques pour les deux langues. Par ailleurs, nous remarquons que les erreurs d'extraction sont plus fréquentes pour le français. Ce phénomène peut s'expliquer par deux facteurs interdépendants :

- la traduction de termes spécifiques au domaine n'est pas toujours appropriée. Par exemple, la traduction de *sonde de battage* (entrée lexicale R0970) selon le traducteur DeepL est *threshing probe* alors que le terme anglais associée à cette entrée est *ramsonde*.
- la correspondance entre un paragraphe anglais et celui issu de la traduction d'un paragraphe français est établie à l'aide d'une mesure de similarité entre chaînes de caractères, ce qui, lorsque la page comporte des entrées lexicales référençant des termes de la même famille, est source d'erreur.

L'exemple de la figure 6 illustre ces phénomènes, avec deux entrées lexicales C3470 et C3480 appartenant à la même famille (*cyanometer* et *cyanometry*), et une similarité plus élevée (selon la distance de Levenshtein) entre *Instrument for determining the shade of blue in the sky* (traduction obtenue pour *Instrument servant à déterminer la teinte du bleu du ciel*) et *Measurement of the shade of blue of the sky* qui est la définition de l'entrée lexicale C3480, qu'avec *Instrument for determining the blueness of the sky.*

6 FAIRisation de données météorologiques à l'aide de dmo-core enrichi

Le processus de FAIRisation d'un jeu de données météorologiques revient à instancier le modèle **dmo-core enrichi**, en suivant d'abord le processus de FAIRisation de **dmo-core** décrit en détail dans [13], et rappelé brièvement en section 3.3. Au terme de cette étape, toute instance de `qb:ComponentProperty` et reliée à une instance de `csvw:Column` peut également être reliée à l'instance de `skos:Concept` correspondant à l'entrée lexicale définissant la colonne si celle-ci existe dans la ressource INMEVO, par la relation `qb:concept`.

L'exemple représentant la colonne `pmer` à l'aide du modèle enrichi devient :

```
inmevo:INMEVO rdf:type skos:ConceptScheme .
inmevo:S0470 rdf:type skos:Concept ;
  skos:prefLabel "sea-level pressure" @en ;
  skos:prefLabel "scanneur" @fr ;
  skos:definition "Atmospheric pressure at mean
    sea level calculated from the observed station
      pressure." @en ;
  skos:definition "Pression atmosphérique au niveau
    moyen de la mer calculée d'après la pression
    mesurée à la station." @fr ;
    skos:inScheme inmevo:INMEVO .

:sea-level_pressure_col
  rdf:type owl:NamedIndividual, csvw:Column ;
  dmo-c:references  :pmer_measure ;
  csvw:datatype "xsd:int" ;
  csvw:name "pmer" ;
  csvw:title "pression au niveau mer" .

:sea_level_pressure rdf:type
  owl:NamedIndividual ,
  sweet:SeaLevelPressure ,
  skos:Concept ,
  <http://www.w3.org/ns/sosa/ObservableProperty> .

:pmer_measure rdf:type owl:NamedIndividual ,
                       qb:MeasureProperty ;
  qb:concept :sea_level_pressure ;
  qb:concept inmevo:S0470 ;
  :unit_of_measures_attr
     <http://qudt.org/vocab/unit#Pascal> .
```

7 Bilan et Perspectives

L'objectif d'améliorer le degré de FAIRisation (et plus précisément le degré de Réutilisabilité) des jeux de données météorologiques en documentant les colonnes des jeux de données tabulaires nous a conduit à enrichir le modèle **dmo-core** pour pouvoir intégrer une ressource lexicale exprimée en SKOS dans le schéma

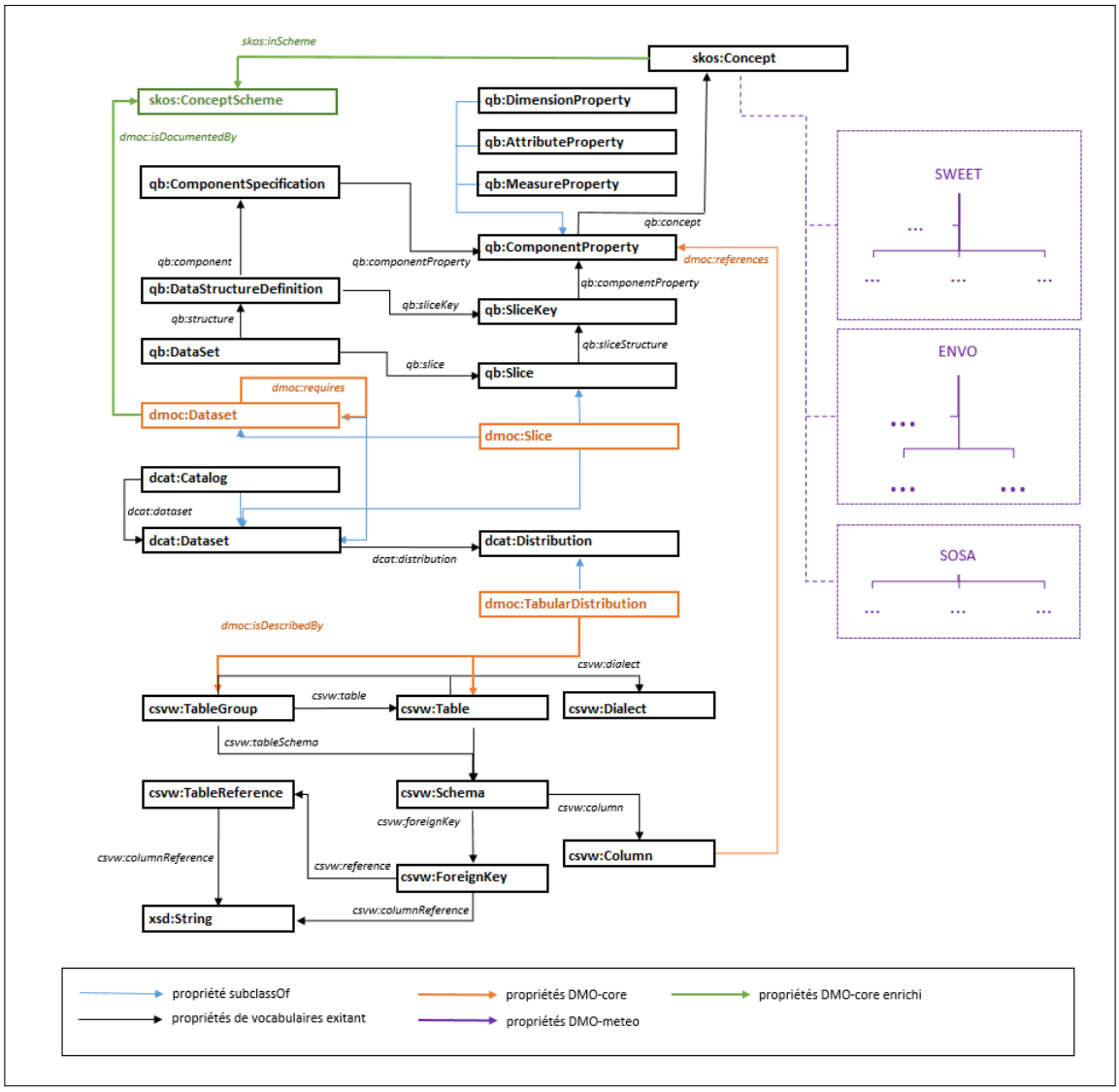


FIGURE 7 – Extension de dmo-meteo à l'aide d'une ressource lexicale au format SKOS.

d'annotation, et à construire la ressource lexicale INMEVO (actuellement dans une version bêta) spécifique au domaine de la météorologie. INMEVO est issue du Vocabulaire International de Météorologie produit par l'Organisation Mondiale de Météorologie, dont les entrées lexicales expriment des termes en quatre langues, des définitions riches et des termes synonymes. La ressource INMEVO est accessible et peut être utilisée dans diverses applications. Par ailleurs, l'approche proposée est générique, et peut être étendue à tout jeu de données, quel que soit le domaine, dès lors qu'une ressource lexicale du domaine formalisée en SKOS est disponible.

Plusieurs suites à cette étude sont envisagées. Un premier objectif est d'améliorer la qualité de la ressource INMEVO, (1) en améliorant le processus d'extraction notamment pour le français, afin de limiter les erreurs mentionnées dans la section 5.4, (2) en rajoutant des informations présentes dans le lexique et non encore prises en compte, comme le pays d'usage d'un terme (e.g. CA pour le Canada), et (3) en intégrant les vocabulaires espagnol et russe présents dans le lexique. Une vérification manuelle au final serait dans tous les cas nécessaire. Au delà du Vocabulaire International de Météorologie produit par l'OMM, cette ressource pourrait être mieux organisée en différenciant par exemple les phénomènes (e.g. dégagement, éclaircie) des propriétés de phénomènes (e.g. pression atmosphérique, niveau de la mer). Le deuxième objectif est de proposer des méthodes d'alignement automatique entre les concepts des ontologies de domaine (e.g. SWEET) et les concepts SKOS du schéma de concepts `inmevo:INMEVO`, toujours dans la perspective de documenter cette fois les ontologies du domaine de la météorologie.

Remerciements

Ce travail a bénéficié du soutien financier de l'ANR pour le projet Semantics4FAIR (2019-2022), contrat

ANR-19-DATA-0014-01.

Références

[1] S. Ahmadi, M. Constant, K. Fort, B. Guillaume, and J. P. McCrae. Convertir le trésor de la langue française en ontolex-lemon : un zeste de données liées. In *Journées LIFT 2021, Linguistique informatique, formelle et de terrain*, Grenoble, France, 2021.

[2] E. Amdouni and C. Jonquet. FAIR or FAIRer ? An integrated quantitative FAIRness assessment grid for semantic resources and ontologies. In *MTSR - 15th International Conference on Metadata and Semantics Research*. Springer, Nov. 2021.

[3] N. Calzolari. Approaches towards a "lexical web" : the role of interoperability. In J. Webster, N. Ide, and A. C. Fang, editors, *Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, pages 18—25. City University, 2008.

[4] M. Frosterus, E. Hyvönen, and J. Laitio. Datafinland - A semantic portal for open and linked datasets. In G. Antoniou, M. Grobelnik, and et al., editors, *8th Extended Semantic Web Conference, ESWC, Heraklion, Crete, Greece*, volume 6644 of *LNCS*, pages 243–254. Springer, 2011.

[5] D. Garijo, Ó. Corcho, and M. Poveda-Villalón. Foops ! : An ontology pitfall scanner for the FAIR principles. In O. Seneviratne, C. Pesquita, J. Sequeda, and L. Etcheverry, editors, *Proc. of the ISWC 2021 Posters, Demos and Industry Tracks : From Novel Ideas to Industrial Practice co-located with 20th Int. Semantic Web Conference (ISWC 2021)*, volume 2980 of *CEUR Workshop Proc.* CEUR-WS.org, 2021.

[6] R. D. Gratta, F. Frontini, F. Khan, and M. Monachini. Converting the parole simple clips lexicon into rdf with lemon. *Semantic Web*, 6 :387–392, 2015.

[7] 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 *Proc. of the 5th Int. Workshop on Semantic Sensor Networks*, volume 904, pages 1–16, 2012.

[8] J. McCrae, D. Spohr, and P. Cimiano. Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th Extended Semantic Web Con- ference (ESWC 2011) on The Semantic Web : research and ap- plications - Volume Part I*, pages 245—259, Berlin, Heidelberg, 2011. Springer-Verlag.

[9] R. Navigli, M. Bevilacqua, S. Conia, D. Montagnini, and F. Cecconi. Ten years of babelnet : A survey. In *Proceedings of IJCAI 2021*, pages 4559–4567, 2021.

[10] V. Parekh, J. Gwo, and T. W. Finin. Ontology based semantic metadata for geoscience data. In H. R. Arabnia, editor, *Conference on Information and Knowledge Engineering*, pages 485–490, 2004.

[11] A. Toral and M. Monachini. Simple-owl : a generative lexicon ontology for nlp and the semantic web. In *Workshop of Cooperative Construction of Linguistic Knowl- edge Bases, 10th Congress of Italian Association for Artificial Intelligence (IA*AI), Rome (Italy)*, 2007.

[12] C. Trojahn, M. Kamel, A. Annane, N. Aussenac-Gilles, and B. L. Nguyen. A FAIR Core Semantic Metadata Model for FAIR Multidimensional Tabular Datasets. In O. Corcho, L. Hollink, O. Kutz, N. Troquard, and F. J. Ekaputra, editors, *23rd International Conference on Knowledge Engineering and Knowledge Management (EKAW 2022)*, volume 13514 of *Lecture Notes in Computer Science book series (LNCS)*, pages 174 – 181, Bolzano, Italy, Sept. 2022. Springer.

[13] C. Trojahn, M. Kamel, A. Annane, N. Aussenac-Gilles, B.-L. Nguyen, and C. Baehr. FAIRification of Multidimensional and Tabular Data by Instantiating a Core Semantic Model with Domain Knowledge : Case of Meteorology. In E. Garoufallou, M.-A. Ovalle-Perandones, and A. Vlachidis, editors, *16th International Conference on Metadata and Semantics Research (MTSR 2022)*, volume TBA, page à paraître, London, United Kingdom, Nov. 2022. springer.

[14] M. Van Assem, A. Gangemi, and G. Schreiber. Conversion of wordnet to a standard rdf/owl representa- tion. In *Proceedings of LREC2006*, Genova, 2006. ELRA, Paris.

[15] M. Wilkinson, M. Dumontier, and *et al.* Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Sc. Data*, 6(1) :1–12, 2019.

[16] N. Yacoubi, C. Faron, F. Michel, F. Gandon, and O. Corby. A Model for Meteorological Knowledge Graphs : Application to Météo-France Observational Data. In *22nd Int. Conf. on Web Engineering, ICWE 2022*, Bari, Italy, July 2022.