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

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

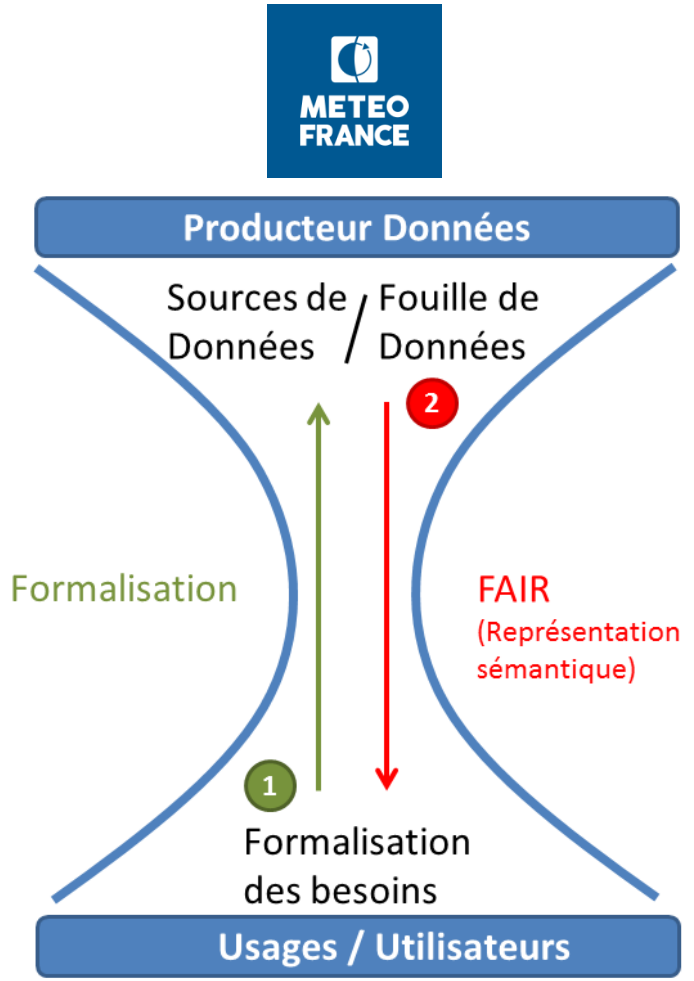
SEMANTICS 4 FAIR

Contexte et motivations

SEMANTICS 4 FAIR

- Un projet multidisciplinaire financé par l'ANR qui regroupe plusieurs laboratoires de recherche:
  - IRIT: Institut de Recherche en Informatique de Toulouse
  - CNRM (Météo-France): Centre National de recherches météorologiques
  - OMP: Observatoire Midi-Pyrénées
  - GET: Géosciences Environnement Toulouse
  - MSH-T: Maison des Sciences de l'Homme et de la Société de Toulouse

- Le but du projet est de faciliter la réutilisation des données météorologiques en améliorant leur degré de FAIRisation



IRIT

MSHS TOULOUSE MAISON DES SCIENCES DE L'HOMME ET DE LA SOCIÉTÉ USR 3414

Réutilisation des données météorologiques

- Les données météorologiques sont essentielles pour avancer la recherche dans plusieurs domaines:
  - Agriculture
  - Biologie
  - Transport maritime
  - Aviation
  - médecine
  - …

Réutilisation des données météorologiques

- Cas de l'ambroisie:
  - L'ambroisie est une plante très allergisante, responsable de divers symptômes allergiques (rhinite, conjonctivite, urticaire, toux, eczéma…) liées à la dissémination de son pollen, à partir du mois d'août jusqu'en octobre.
  - Il est considéré aujourd'hui comme un polluant biologique par les autorités sanitaires.
  - Des chercheurs à l'Observatoire Midi-Pyrénées (OMP) veulent étudier la corrélation entre les conditions climatologiques et la propagation de la plante

  - Besoin de réutiliser des données météorologiques

SYNOP un jeu de données météorologiques

DONNÉES SYNOP ESSENTIELLES OMM

Description

Données d'observations issues des messages internationaux d'observation en surface (SYNOP) circulant sur le système mondial de télécommunication (SMT) de l'Organisation Météorologique Mondiale (OMM). Paramètres atmosphériques mesurés (température, humidité, direction et force du vent, pression atmosphérique, hauteur de précipitations) ou observés (temps sensible, description des nuages, visibilité) depuis la surface terrestre. Selon instrumentation et spécificités locales, d'autres paramètres peuvent être disponibles (hauteur de neige, état du sol, etc.)

Métropole et outre-mer - Fréquence : 3 h - Format : ASCII

Conditions d'accès ⊖

- Sans redevance sous Licence Ouverte d'Etalab. La source à indiquer est "Météo-France". Quelques suggestions : "Source : Météo-France" ou "Informations créées à partir de données de Météo-France".

Moyens d'accès ⊖

- Téléchargement direct via le formulaire ci-dessous.

Documentation ⊖

- Descriptif des paramètres de données SYNOP essentielles OMM
- Liste des stations essentielles (format csv)
- Liste des stations essentielles (format GeoJSON)

Téléchargement ⊕

Téléchargement de données archivées ⊕

SYNOP un jeu de données météorologiques (suite)

Extrait du fichier qui documente les données

Descriptif	Mnémonique	type	unité
Indicatif OMM station	numer_sta	car	
Date (UTC)	date	car	AAAAMMDDHHMISS
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

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	…
…	…	…	…	…	…

Extrait des données

SYNOP un jeu de données météorologiques (suite)

Fichier pdf, pas de définition des paramètres, labels non précis, Lien cassé, valeurs codifiées, etc.

Descriptif	Mnémonique	type	unité
Indicatif OMM station	numer_sta	car	
Date (UTC)	date	car	AAAAMMDDHHMISS
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

numer_sta	date	pmer	ff	t	…
7005	20200201000000	100710	3.200000	285.4500	
7015	20200201000000	100710	7.700000	284.950000	
7020	20200201000000	100630	8.400000	284.150000	
7027	20200201000000	100770	5.500000	285.650000	…
…	…	…	…	…	…

Acronymes
Définition des mesures
Mauvaise indexation
Valeurs codées
API d'accès non disponible
…

Les principes FAIR

Findable (re-trouvable)

- F1. Les (méta)données sont associées à un identifiant unique et pérenne.
- F2. Les (méta)données sont décrites avec des métadonnées **riches**.
- F3.Les métadonnées incluent clairement et explicitement l'identifiant des données qu'elles décrivent
- F4. Les (meta)données sont enregistrées ou indexées dans un dispositif permettant de les rechercher.

Accessible (Accessible)

- A1. Les (méta)données sont accessibles par leur identifiant, via un protocole standardisé.
  - A1.1 Le protocole utilisé est ouvert, libre et peut être implémenté de manière universelle.
  - A1.2 Le protocole utilisé permet l'accès par autorisation et authentification si besoin.
- A2. Les métadonnées restent accessibles même si les données ne le sont pas ou plus.

Interoperable (Interopérable)

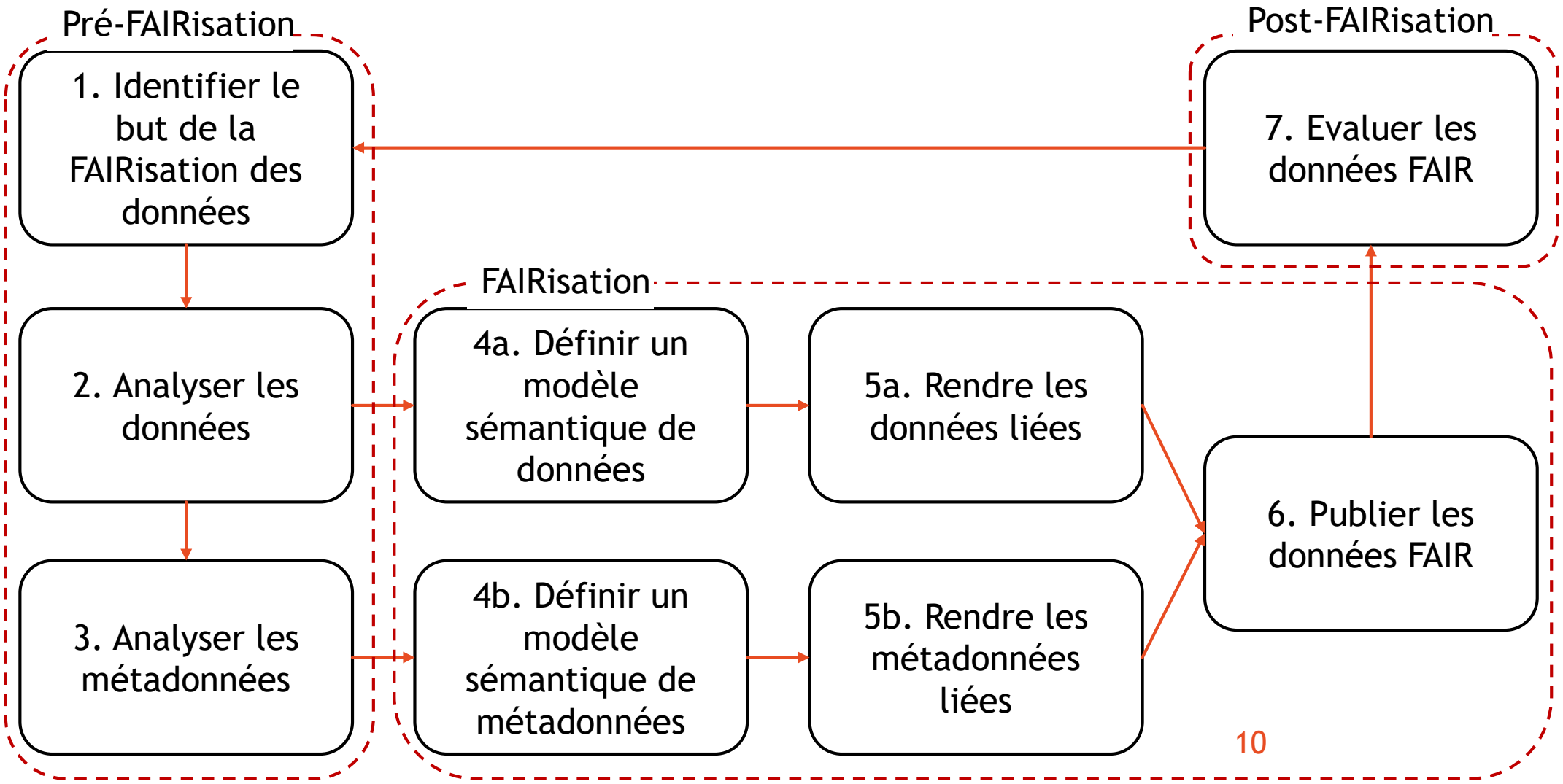
- I1. Les (méta)données utilisent un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances.
- I2. Les (méta)données utilisent des vocabulaires qui adhèrent aux principes FAIR.
- I3. Les (méta)données ont des liens documentés vers d'autres (méta)données.

Reusable (Réutilisable)

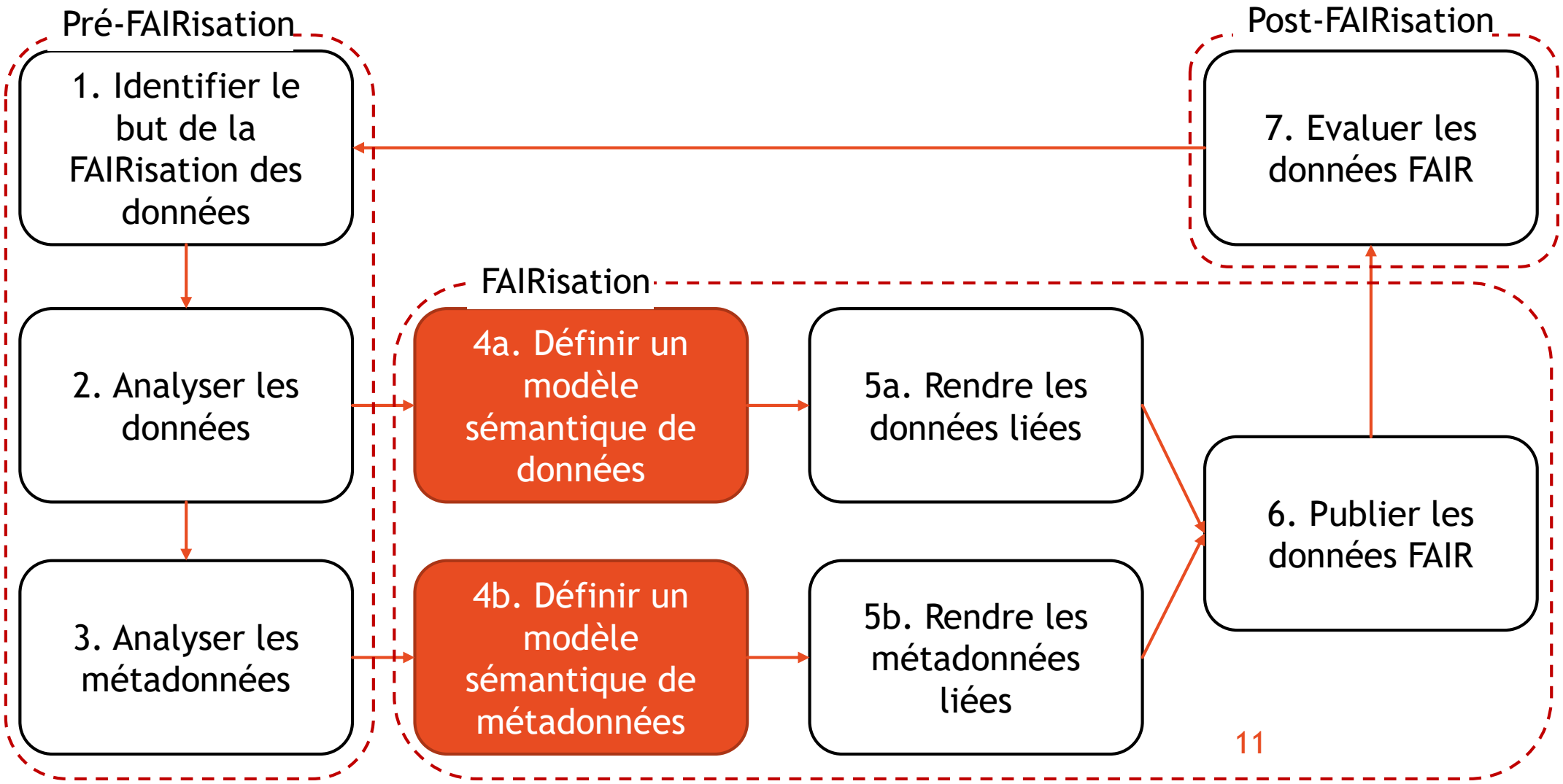
- R1. Les (méta)données ont des attributs multiples et pertinents.
  - R1.1. Les (méta)données sont mises à disposition selon une licence explicite et accessible.
  - R1.2. Les (méta)données sont associées à leur provenance.
  - R1.3 Les (méta)données sont conformes aux standards des communautés indiquées.

(Wilkinson et al., 2016)

Processus de FAIRisation



Processus de FAIRisation



Plan

- Développement du modèle
  - Méthode de développement
  - Spécification
  - Sélection d'ontologies
  - Intégration d'ontologies
- Evaluation
  - Instanciation des données SYNOP
  - Evaluation de l'impact de l'utilisation du modèle sur le degré de FAIRisation
- Conclusion et perspectives

Un modèle sémantique pour la représentation des (méta)données des datasets météorologiques

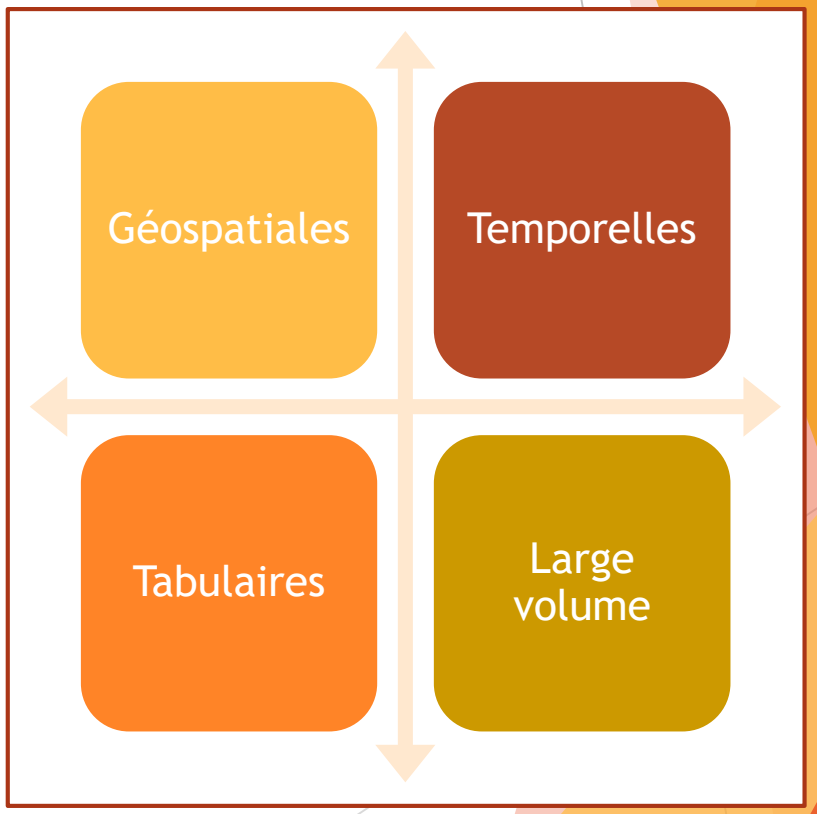
Méthode de développement du modèle

- Le développement du modèle a été basé sur la réutilisation des modèles existants afin d'adhérer au principe « I »

Spécification

- Données météorologiques d'observation dites « in situ » : ce sont des 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).

- Caractéristiques des données météorologiques d'observation



Spécification

- Pas de transformation des données météorologiques en RDF:
  - Un coût élevé: nécessitent un investissement important (des ressources humains et matériels)
  - Pas efficace pour l'interrogation des données: génère un graphe RDF immense
  - Les logiciels existants traitant les données spatio-temporelles ne traitent pas forcément des données RDF

Spécification

- Competency questions:
  - Quelle est la signification du paramètre « point de rosé »?
  - Dans quel format peut on télécharger les données?
  - Quel est le type de température/humidité fourni dans les données SYNOP?
  - Quelle est la méthode de mesure/ comment est calculé tel paramètre?
  - Quelle est la signification des valeurs du paramètre « Temps présent »?
  - ……

- Deux besoins principaux:
  - Représentation des métadonnées des jeux de données météorologiques
  - Représentation des données, notamment via:
    - Représentation des schémas de données des jeux de données météorologiques (i.e., expliciter la sémantique des entités incluses)
    - Représentation des valeurs codées
    - Représentation de la structure des distributions de données

Sélection d'ontologies: vocabulaires existants

Metadata

- INSPIRE schema
- DCAT
- DCAT-AP
- GeoDCAT-AP
- ADMS
- VoiD
- (Frosterus et al. 2011)
- (Parekh et al. 2004)

Data and data schema

- SOSA
- GeoSPARQL
- Time
- PROV-O
- RDF data cube
- QB4ST

ontologies générales

- SWEET
- ENVO
- AWS
- Irstea ontologie
- CANDELA ontologie

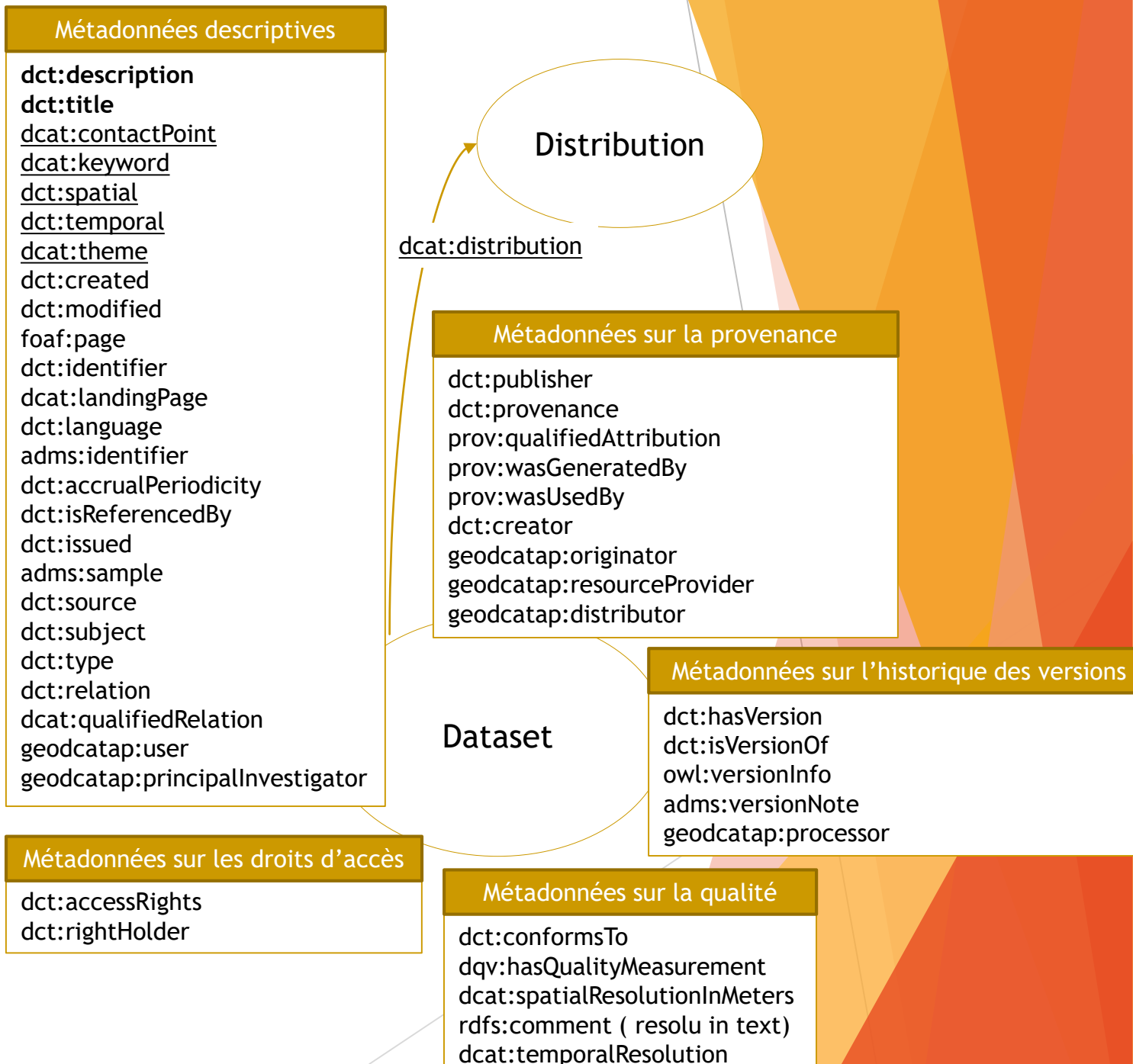
ontologies de domaine

Data structure

- CSVW
- JSON-LD

Sélection d'ontologies

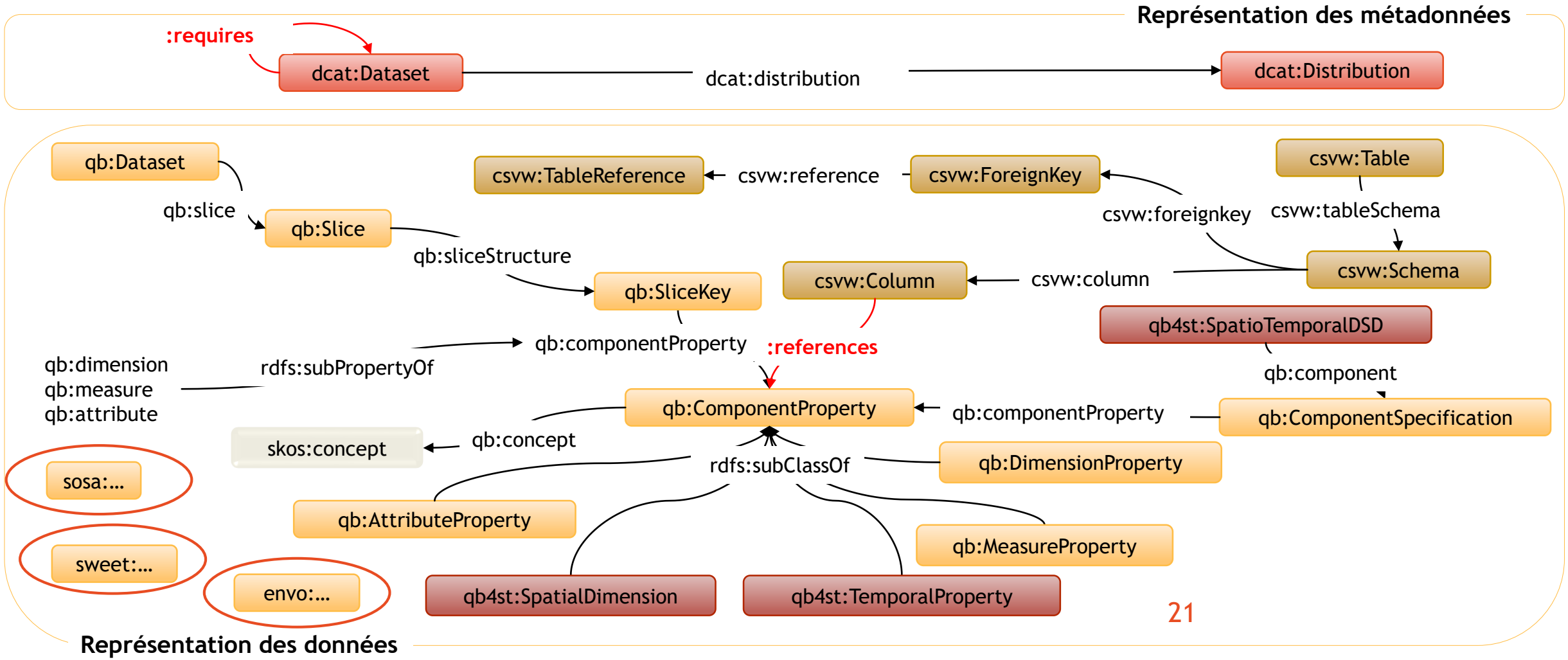
- GeoDCAT-AP: le vocabulaire choisi pour représenter les métadonnées
  - Une spécification de DCAT qui est FAIR (selon https://fairsharing.org/)
  - Permet la représentation des différentes catégories de métadonnées pour adhérer aux principes « F » et « R ».



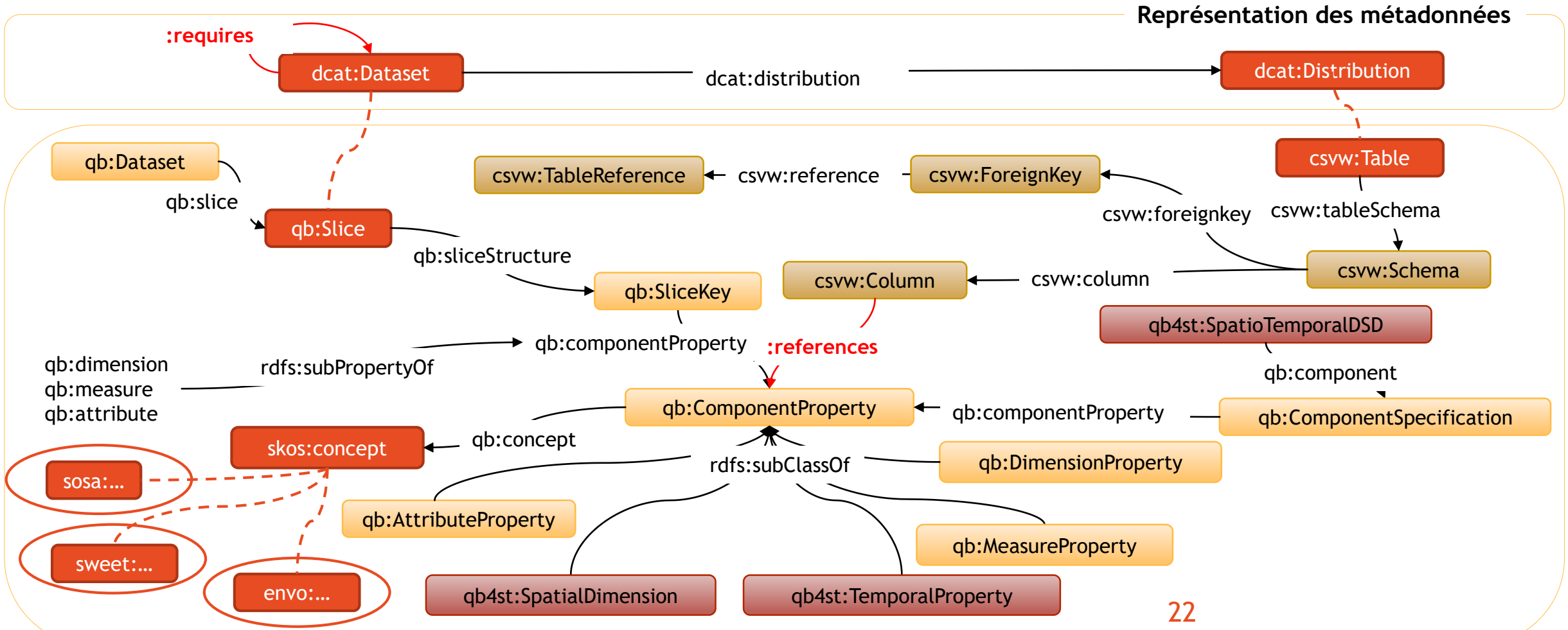
Sélection d'ontologies

- RDF Data Cube et QB4ST
  - Représenter le schéma de données multidimensionnelles indépendamment du format des distributions
  - W3C recommandation, un vocabulaire FAIR
- CSVW
  - Représenter la structure des distributions tabulaires
- Ontologies de domaines pour expliciter les entités sémantiques incluses dans les jeux de données
  - SWEET, ENVO
  - QUDT
  - SOSA
- Les ontologies réutilisées sont FAIR (selon https://fairsharing.org/)

Intégration d'ontologies: Notre modèle



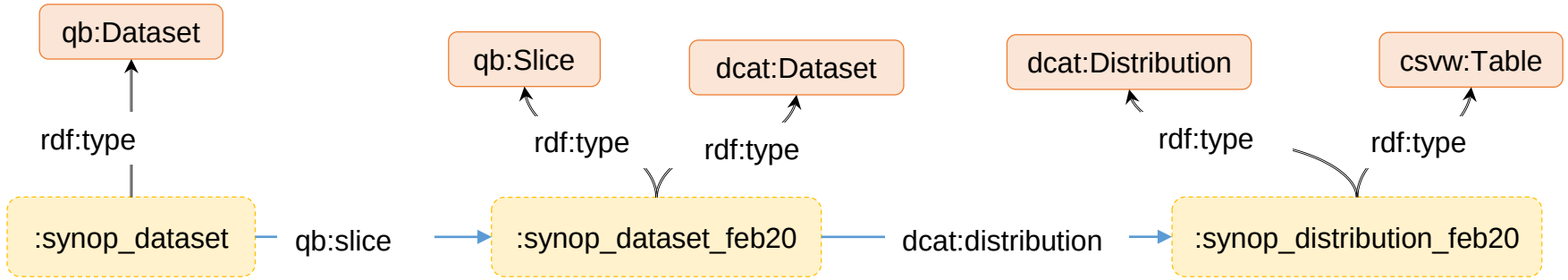
Intégration d'ontologies: Notre modèle



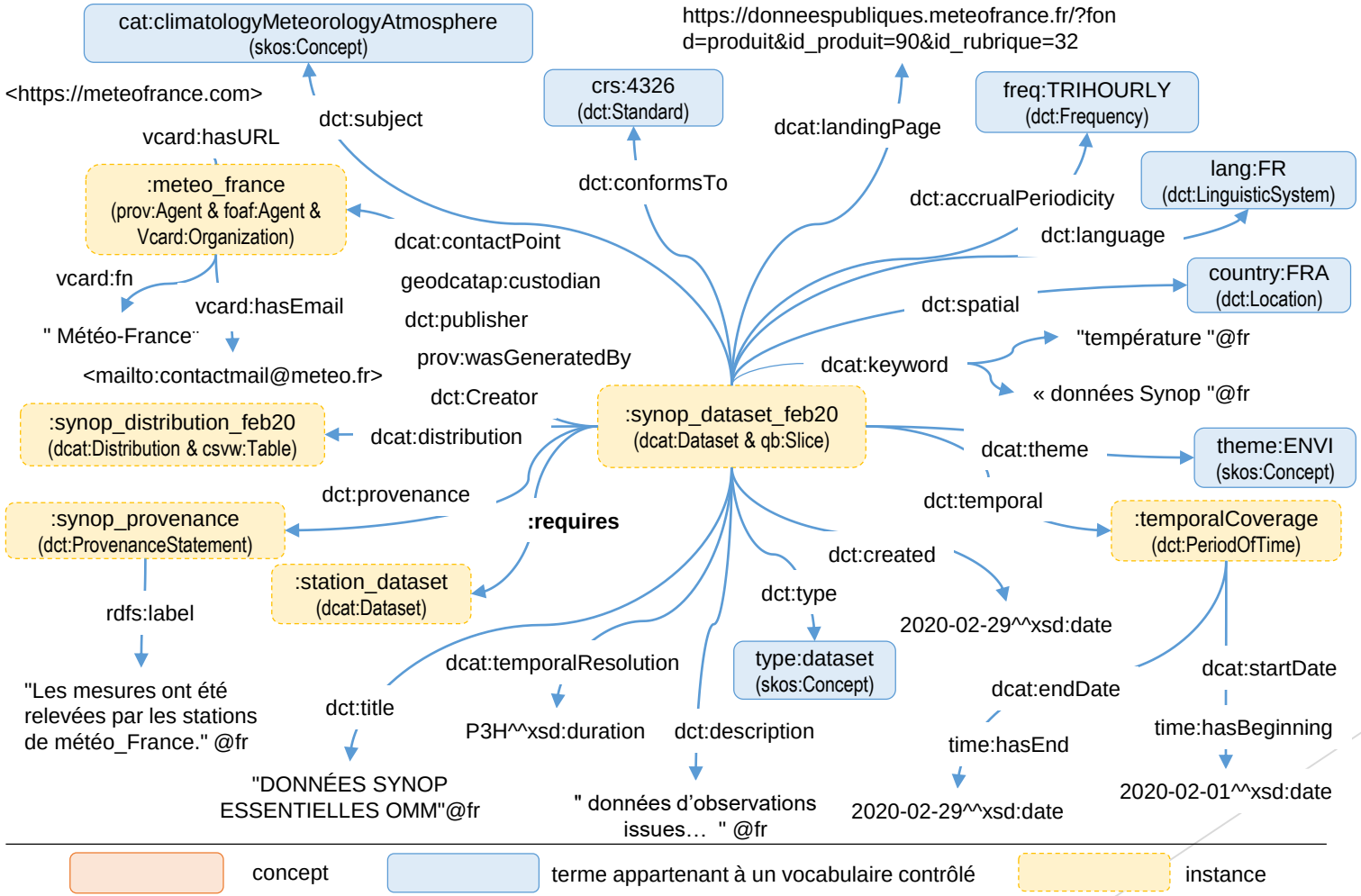
Evaluation

Instanciation des données SYNOP

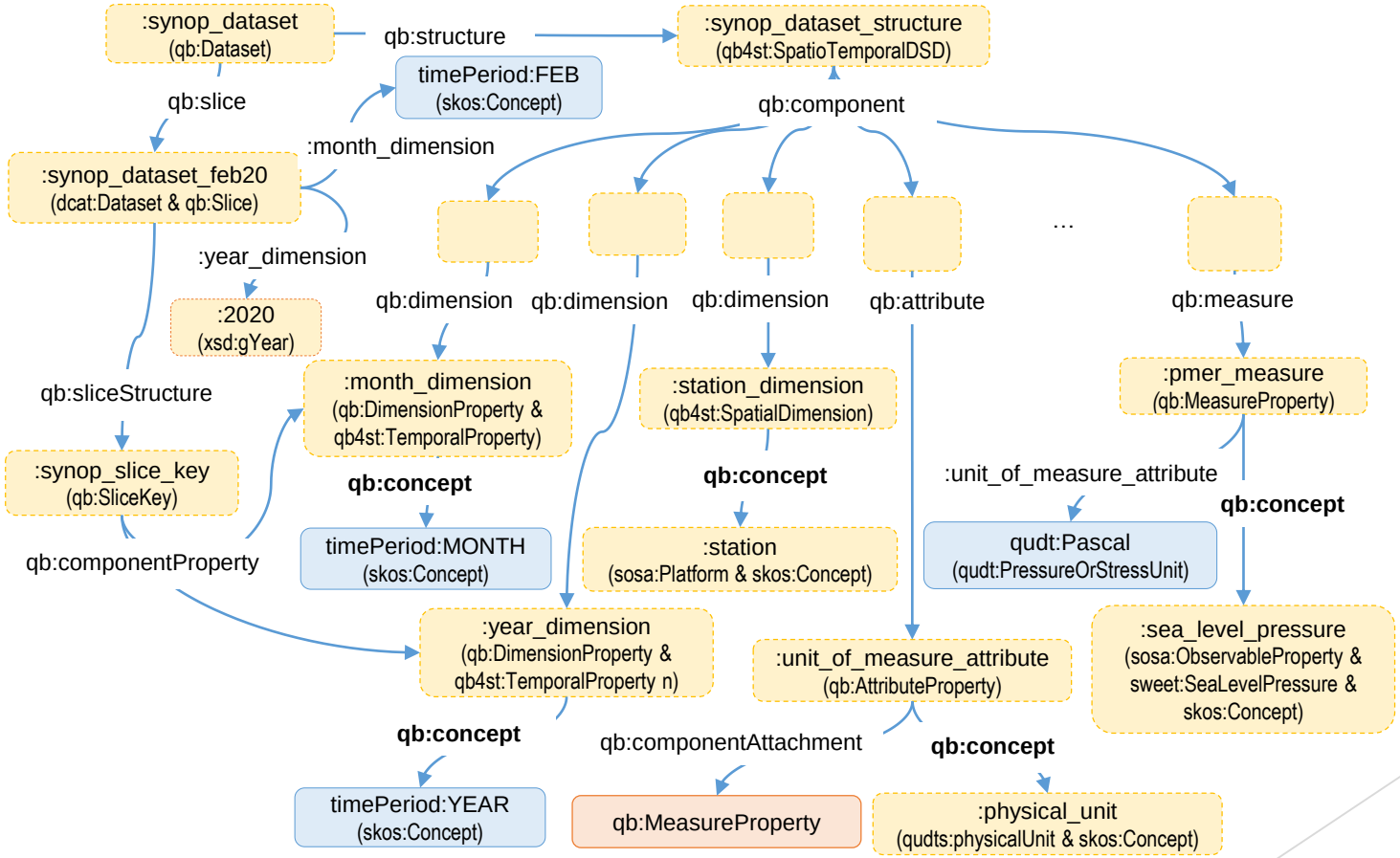
- Organisation des données:
  - Archive existante depuis Janvier 1996
  - L'archive se compose d'un ensemble de fichier csv
  - Chaque fichier inclut les données d'un seul mois



Instanciation des données SYNOP



Instanciation des données SYNOP



Instanciation des données SYNOP

```
MeteOnto:station_pressure a qb:MeasureProperty;
        rdfs:label "pres"^^xsd:string;
        skos:altLabel "pression à  la station"@fr;
        skos:definition "Pression déduite de la lecture d'un baromètre à  la station après
        corrections et, si nécessaire, réuction de sa valeur à  l'altitude de la station."@fr;
        skos:altLabel "station pressure"@en;
        skos:definition "Pressure deduced from the reading of a barometer at the station
        after applying instrument corrections and, if necessary, reducing its value to the
        height of the station."@en;
        rdfs:range xsd:int;
        MeteOnto:unitOfMeasure qudt:Pascal.
```

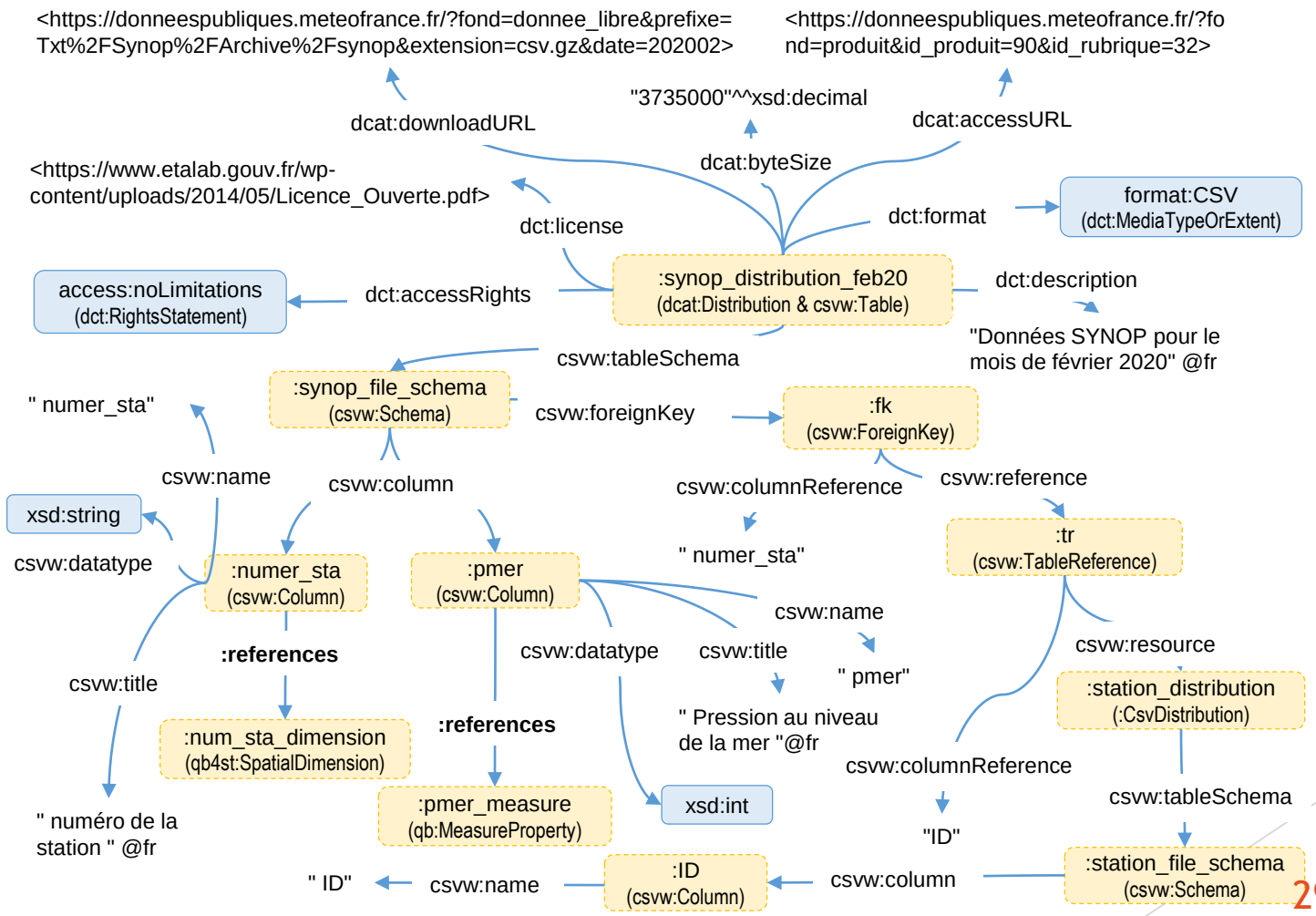
Représentation d’une mesure en RDF

Instanciation des données SYNOP

```
MeteOnto:sea-level_pressure_ a cube:MeasureProperty;
                rdfs:label "pmer"^^xsd:string;
                skos:altLabel "pression au niveau mer"@fr;
                skos:altLabel "sea-level pressure "@en;
                rdfs:range xsd:int;
                MeteOnto:unitOfMeasure qudt:Pascal;
                cube:concept [ a <http://sweetontology.net/propPressure/SeaLevelPressure>,
skos:Concept].
```

Représentation d’une mesure en RDF

Instanciation des données SYNOP



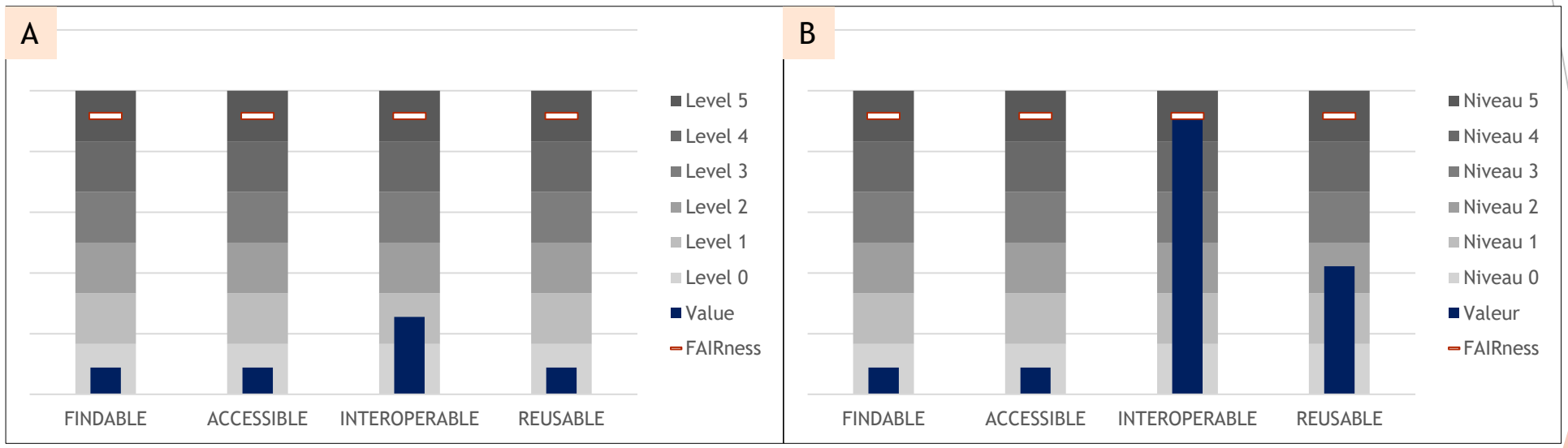
Evaluation du degré de la FAIRisation

- FAIR data maturity model (DOI: 10.15497/rda00050)
  - Proposé par la RDA (Research Data Alliance), publié en juin 2020
  - Trois composants:
    1. Indicateurs: les aspects individuels de la FAIRisation qui sont évalués (41)
       i. E.g., from the principle “F1. (Meta)data are assigned a globally unique and persistent identifier”, four indicators are defined
    2. Priorité: l’importance relative à chaque indicateur (essentiel, important et utile)
    3. Méthodes d’évaluation: la manière avec laquelle les valeurs sont attribués aux résultats de l’évaluation des indicateurs
       i. Mesurer le progrès (Measuring progress)
       ii. Mesurer la conformité ou l'échec (Measuring pass-or-fail)

	Essential	Important	Useful
Level 0	○		
Level 1	●		
Level 2	●	◐	
Level 3	●	●	
Level 4	●	●	◐
Level 5	●	●	●

○ None of the indicators are satisfied
◐ Half of the indicators are satisfied
● All indicators are satisfied

Evaluation du degré de la FAIRisation



Level 0	Pas FAIR
Level 1	FAIR critères essentiels uniquement
Level 2	FAIR critères essentiels + 50 % critères importants
Level 3	FAIR critères essentiels + 100% critères importants
Level 4	FAIR critères essentiels + 100% critères importants+ 50% critères utiles
Level 5	FAIR critères essentiels + 100% of important criteria + 100% critères utiles

Conclusion et perspectives

Conclusion

- Un modèle sémantique basé sur des vocabulaires de référence « FAIR » pour représenter les données météorologiques d'observation
- En plus des métadonnées générales, le modèle représente le schéma des données et explicite les entités sémantiques à l'aide des ontologies de domaine
- Une première évaluation a montré que:
  - Le modèle permet de bien représenter le jeu de données représentatif SYNOP
  - Le modèle est consistant avant et après l'instanciation
  - Les métadonnées générées de l'instanciation permettent d'améliorer le degré de FAIRisation, notamment les principes « F », « I », et « R »

Perspectives

- Une évaluation plus poussée du modèle
  - E.g., Instancier de nouveaux jeux de données pour évaluer la capacité du modèle à représenter les métadonnées des datasets, et l'enrichir si besoin
- Développer un outil interactif facilitant la saisie des métadonnées à base d'un modèle sémantique (travail en cours…)
- Publier et indexer les métadonnées générées sur des portails de données après leur avoir généré des identifiants persistent
- Développer un algorithme de recherche sémantique de datasets qui exploite la représentation fine du schéma de données (travail en cours…)

Merci de votre attention
Des questions ?

Références bibliographiques

- M. D. Wilkinson *et al.*, “Comment: The FAIR Guiding Principles for scientific data management and stewardship,” *Sci. Data*, vol. 3, 2016.
- C. Roussey, S. Bernard, G. Andre, and D. Boffety, “Weather Data Publication on the LOD using SOSA/SSN Ontology,” Semant. Web, 2019.
- L. Lefort, “Ontology for Meteorological sensors,” 2010. [Online]. Available: https://www.w3.org/2005/Incubator/ssn/ssnx/meteo/aws#.
- K. Hans Peter de, N. Rouquette, R. Burkhart, H. Espinoza, and L. Lefort, “Library for Quantity Kinds and Units: schema, based on QUDV model OMG SysML(TM), Version 1.2,” 2011. [Online]. Available: https://www.w3.org/2005/Incubator/ssn/ssnx/qu/qu.
- M. Perry and J. Herring, “OGC GeoSPARQL-A geographic query language for RDF data,” *OGC Candidate Implement. Stand.*, p. 57, 2012.
- R. G. Raskin and M. J. Pan, “Knowledge representation in the semantic web for Earth and environmental terminology (SWEET),” Comput. Geosci., vol. 31, no. 9, pp. 1119–1125, Nov. 2005.
- W3C, “Data Catalog Vocabulary (DCAT) - Version 2,” 2020. [Online]. Available: https://www.w3.org/TR/vocab-dcat-2/.
- G. Atemezing et al., “Transforming meteorological data into linked data,” Semant. Web, vol. 4, no. 3, pp. 285–290, 2013.
- M. Frosterus, E. Hyvönen, and J. Laitio, “DataFinland-A semantic portal for open and linked datasets,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, vol. 6643 LNCS, no. PART 2, pp. 243–254
- The RDF Data Cube Vocabulary (January 2014) – W3C recommendation (https://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/)