Séminaire In-OVIVE, 21 septembre 2021

https://www6.inrae.fr/reseau-in-ovive/Actions-du-reseau/Seminaires/Seminaire-du-21-septembre-2021

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

Amina ANNANE, Nathalie Aussenac-Gilles, Mouna Kamel, Cassia Trojahn, Catherine Comparot (IRIT)

et Christophe Baehr (CNRM)

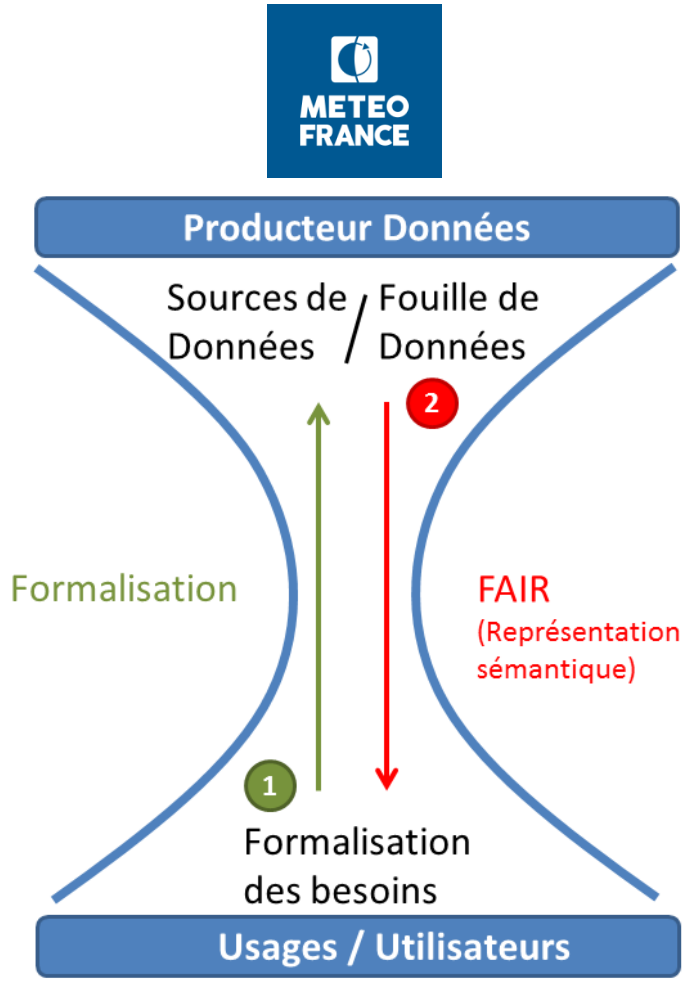
Réseau IN-OVIVE

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
  - GET-OMP: Géosciences Environnement Toulouse ; Observatoire Midi-Pyrénées
  - 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

GET

CNRM

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

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)

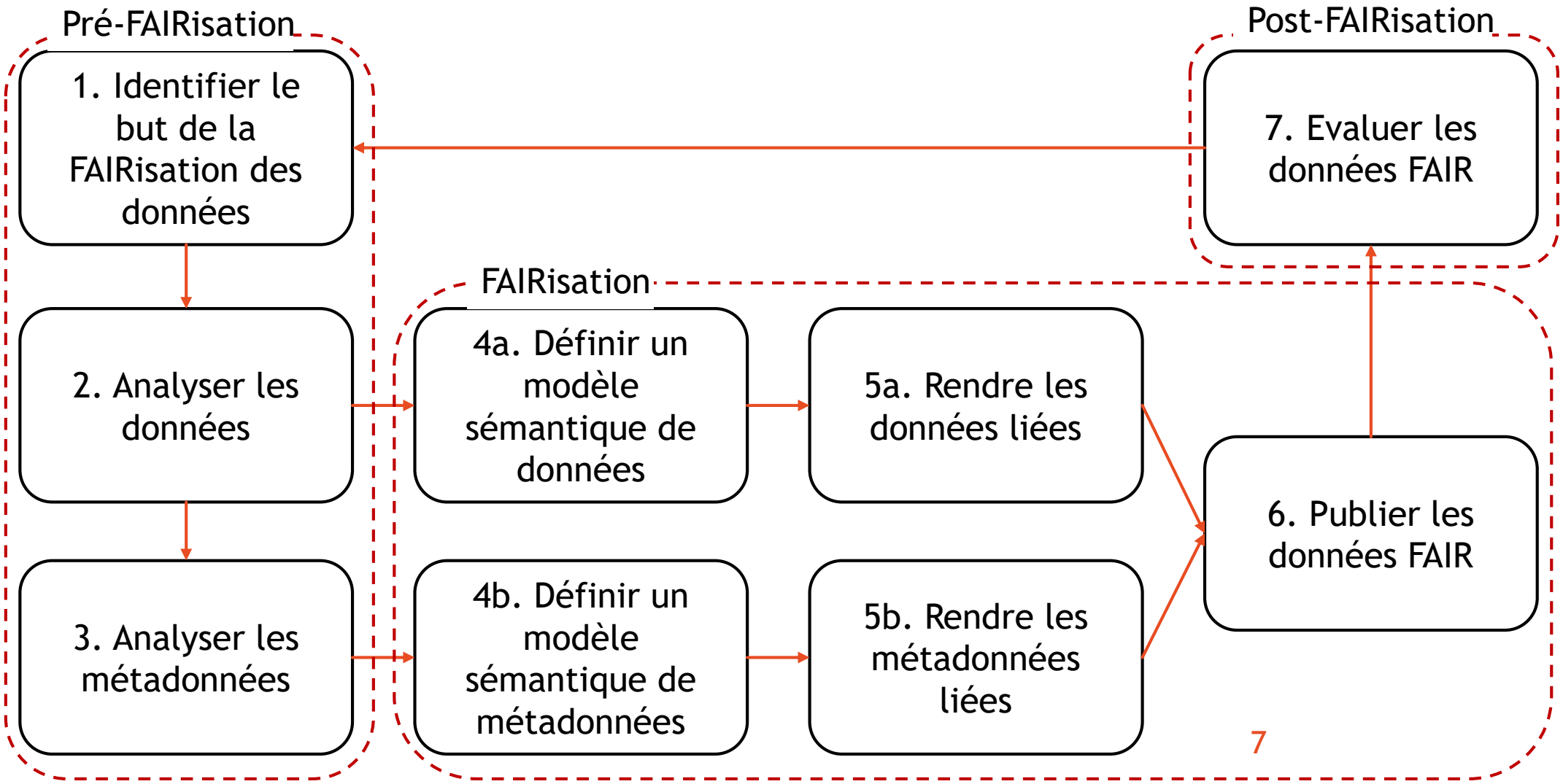
- I1. Les (méta)données utilisent un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances.
- I2. Les (méta)données utilisent des vocabulaires qui adhèrent aux principes FAIR.
- I3. Les (méta)données ont des liens documentés vers d'autres (méta)données.

Reusable (Réutilisable)

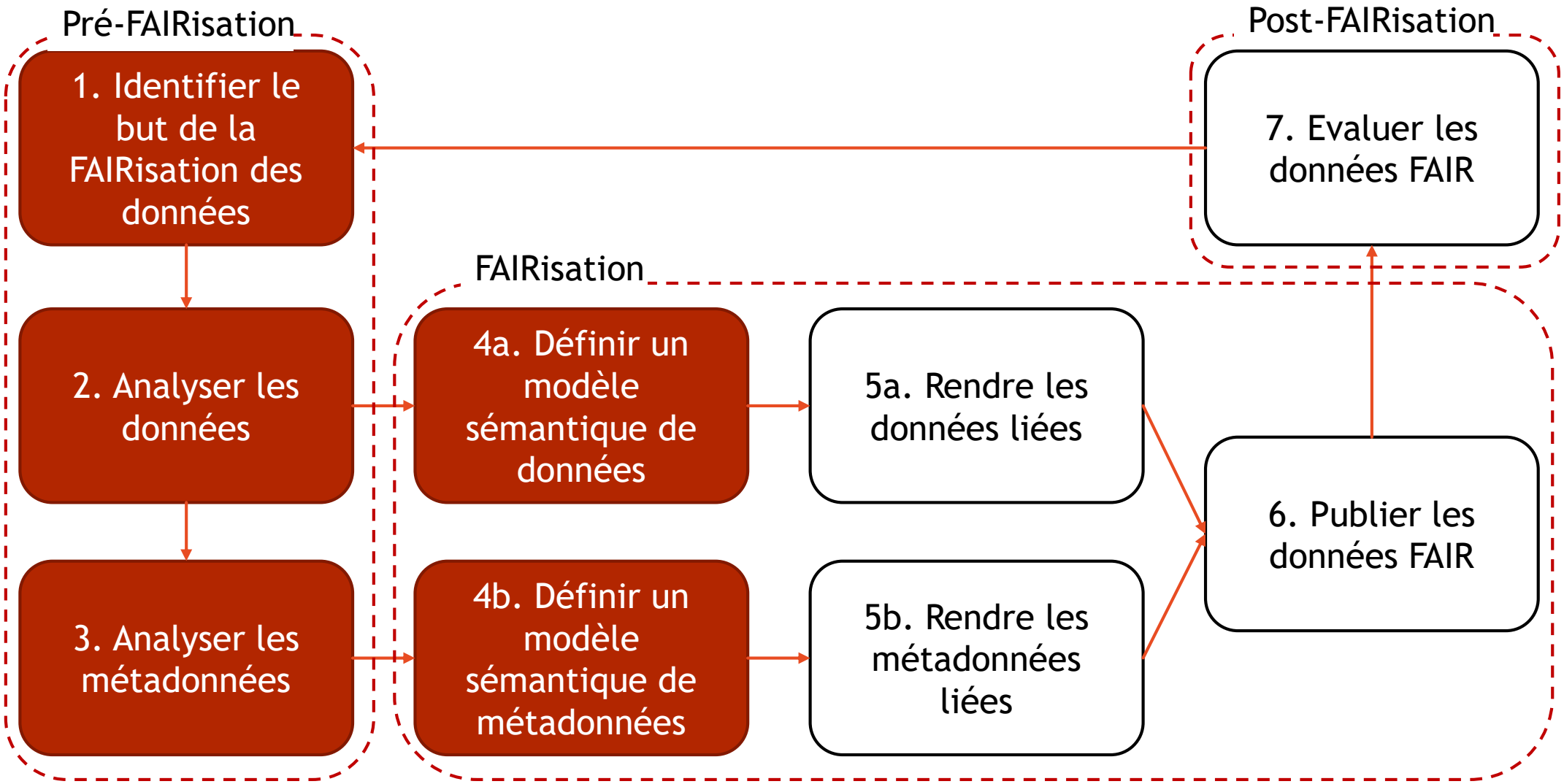
- R1. Les (méta)données ont des attributs multiples et pertinents.
  - R1.1. Les (méta)données sont mises à disposition selon une licence explicite et accessible.
  - R1.2. Les (méta)données sont associées à leur provenance.
  - R1.3 Les (méta)données sont conformes aux standards des communautés indiquées.

(Wilkinson et al., 2016)

Processus de FAIRisation (Jacobsen et al., 2020)



Processus de FAIRisation (Jacobsen et al., 2020)



Plan

- Pré-FAIRisation: Analyse des données et des métadonnées
- FAIRisation: développement du modèle sémantique
  - Méthodologie de développement
  - Spécification
  - Sélection d'ontologies
  - Intégration d'ontologies
- Evaluation
  - Instanciation du modèle développé avec le jeu de données SYNOP
  - Evaluation de l'impact de l'utilisation du modèle sur le degré de FAIRisation
- Conclusion et perspectives

Pré-FAIRisation: Analyse des données et des métadonnées

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)

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

Extrait des données

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

Documentation: Liste des stations

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

Documentation: Description des paramètres

SYNOP un jeu de données météorologiques (suite)

numer_sta	date	pmer	tend	cod_tend	dd	ff	t	…
7005	20200201000000	100710	-200	8			85.450000	…
7015	20200201000000	100710	-170	7			84.950000	…
7020	20200201000000	100630	-40	5			284.150000	…
7027	20200201000000	100770	-130	6	200	5.500000	285.650000	…
7	20200201000000	100830	-230	6	200	7.000000	285.150000	…
	200201000000	101140	-190	8	210	4.900000	285.4	
	200201000000	100780	-60	8	230	4.500000	284.7	
	…	…	…	…	…	…		

Valeurs codées…

API d'accès aux données non disponible?

Extrait des données

Des patrons, des unités de mesure et des liens vers des codes…

Fichier pdf..

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

Documentation: Li

Pas de métadonnées sémantiques..

Descriptif	Mnémonique	type	unité
Indicatif OMM station	numer_sta	car	
Date (UTC)		car	AAAAMMDDHHMISS
Pression au niveau m		int	Pa
Variation de pression en 3 he	tend	int	Pa
Type de tendance baromé	cod_tend	int	code
Direction du vent moy 10 mn	dd	int	degré
Vitesse du vent m yen 10 mn	ff	réel	m/s
Température	t	réel	K
Point de rosée		réel	K

Quelle température?

Lien cassé

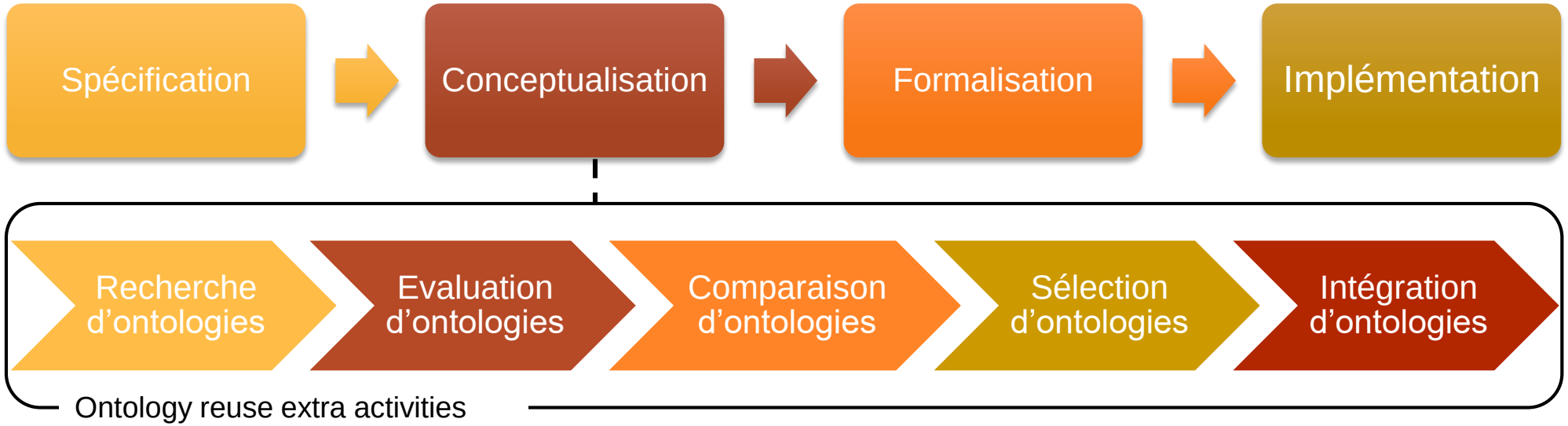
C'est quoi le point de rosée?

Documentatio paramètres

FAIRisation: Développement du modèle sémantique pour la représentation des méta(données)

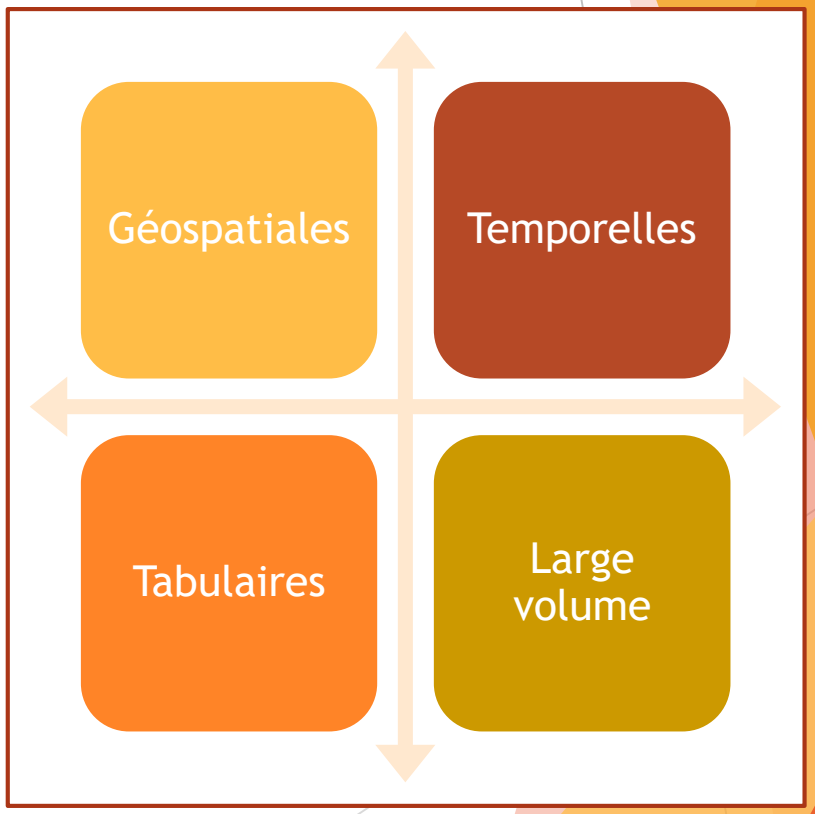
Méthodologie de développement du modèle

- Le développement du modèle a été basé sur la réutilisation des ontologies existantes afin d'adhérer au principe « I »
- Nous avons adopté la méthodologie NeOn-Scénario 03 « Réutilisation d'ontologies» (Baonza, Pérez, & Villazón, 2008)



Spécification (1/3): Caractéristiques des données météorologiques d'observation

- 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).



Spécification (2/3)

- 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 »?
  - ……
- Pas de transformation des données météorologiques en RDF:
  - Un coût élevé: nécessitent un investissement important (des ressources humaines et matérielles)
  - 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 (3/3)

I. Représenter les métadonnées des jeux de données météorologiques

II. Représenter le contenu des données sans les transformer en RDF

- Représenter le schéma des données
- Représenter les entités sémantiques incluses dans les données et les définir
- Représenter les valeurs codées

III. Représenter la structure des distributions de données

Recherche et évaluation d'ontologies

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 (1/4)

I. Représenter les métadonnées des jeux de données météorologiques

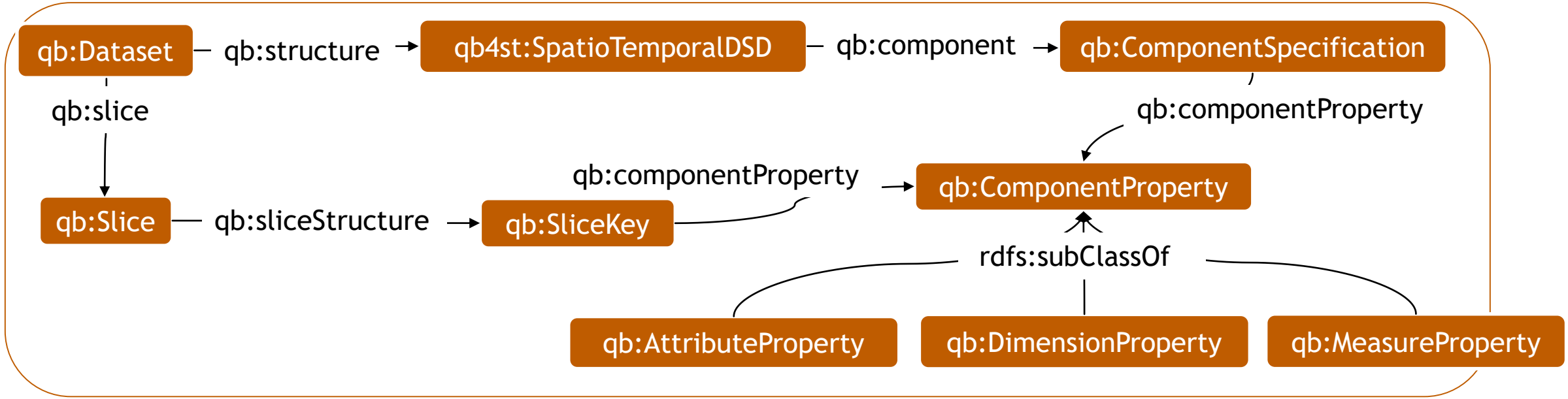
- GeoDCAT-AP: le vocabulaire choisi pour représenter les métadonnées
  - Une spécification de DCAT, dédiée aux données géospatiales (DCAT est FAIR selon https://fairsharing.org/)
  - Permet la représentation des différentes catégories de métadonnées pour adhérer aux principes « F » et « R ».

Métadonnées descriptives	Métadonnées sur les droits d'accès	Métadonnées sur la qualité	Métadonnées sur la provenance	Métadonnées sur l'historique des versions
dct:description **dct:title** dcat:contactPoint dcat:keyword dct:spatial dct:temporal dcat:theme dct:created	dct:accessRights dct:rightHolder	dct:conformsTo dqv:hasQualityMeasurement dcat:spatialResolutionInMeters rdfs:comment (resolu in text) dcat:temporalResolution	dct:publisher dct:provenance prov:qualifiedAttribution prov:wasGeneratedBy prov:wasUsedBy dct:creator geodcatap:originator geodcatap:resourceProvider	dct:hasVersion dct:isVersionOf owl:versionInfo adms:versionNote geodcatap:processor

Sélection d'ontologies (2/4)

II. Représenter le contenu des données sans les transformer en RDF

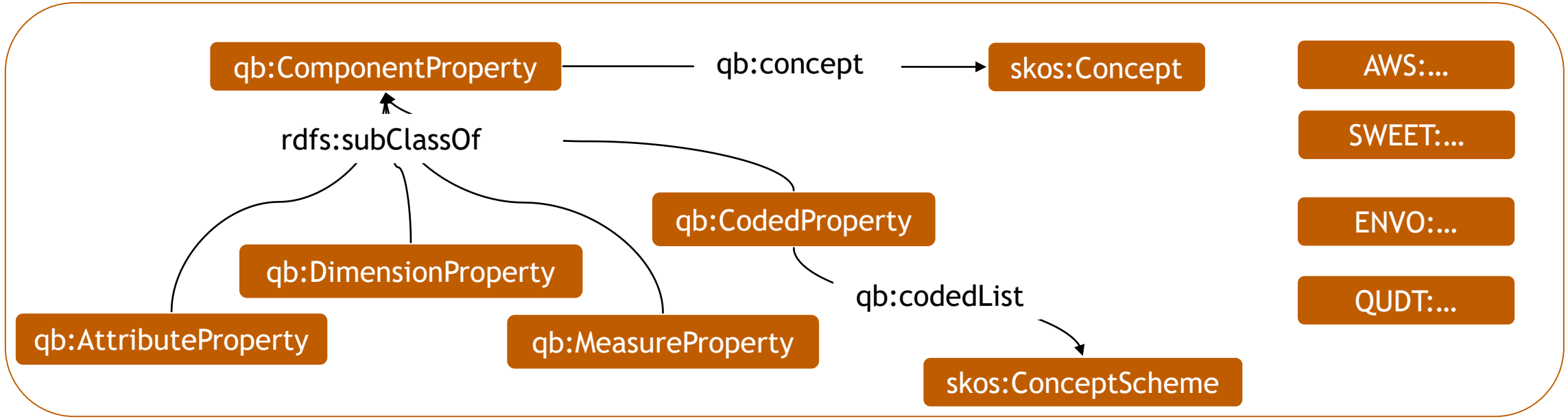
- Représenter le schéma des données: **RDF Data Cube** (qb) et QB4ST
  - Représenter le schéma des données multidimensionnelles indépendamment du format des distributions
  - W3C recommandation, un vocabulaire FAIR



Sélection d'ontologies (3/4)

II. Représenter le contenu des données sans les transformer en RDF

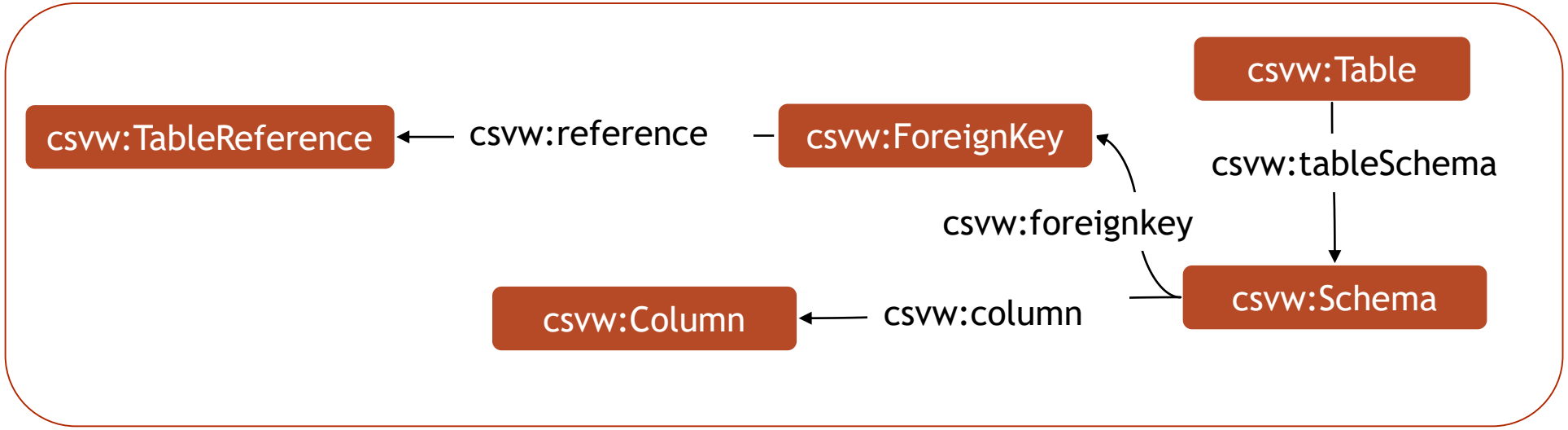
- Représenter les entités sémantiques incluses dans les données et les définir:
  - utiliser la propriété « qb:concept » pour lier les mesures, dimensions et attributs aux concepts de domaine des ontologies : AWS, SWEET, ENVO, SOSA, QUDT.
- Représenter les valeurs codées: utiliser la propriété « qb: CodedProperty » et « qb:codedList »



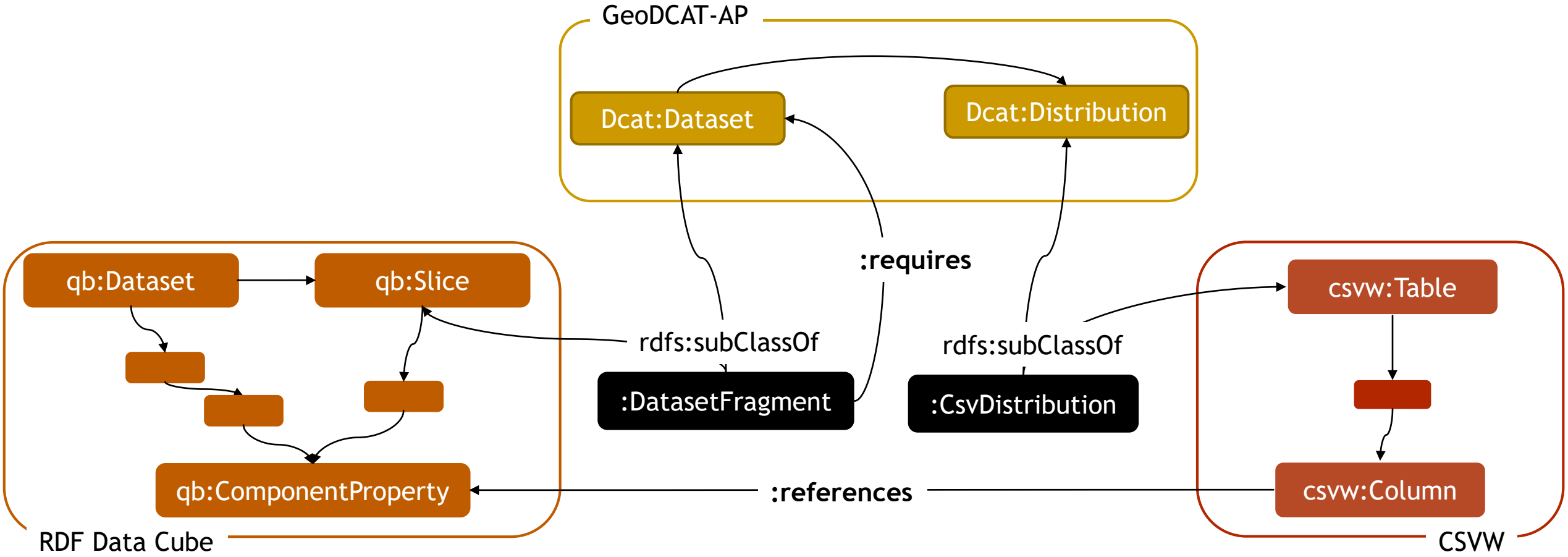
Sélection d'ontologies (4/4)

III. Représenter la structure des distributions de données

- CSVW
  - Représenter la structure des distributions csv



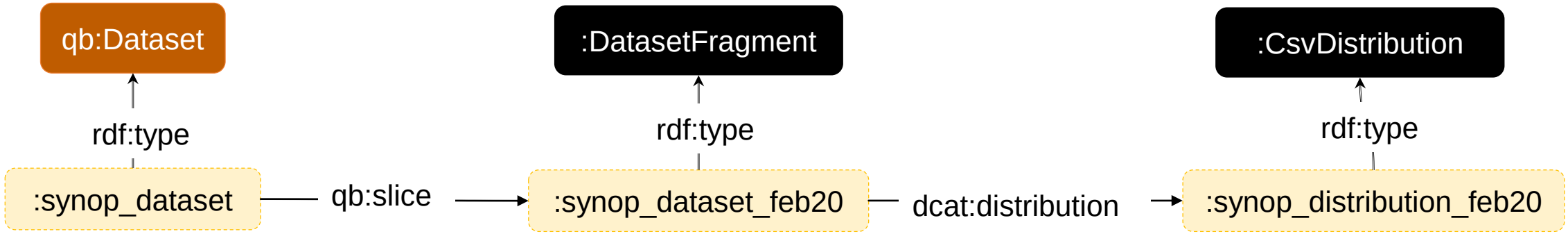
Intégration des ontologies sélectionnées



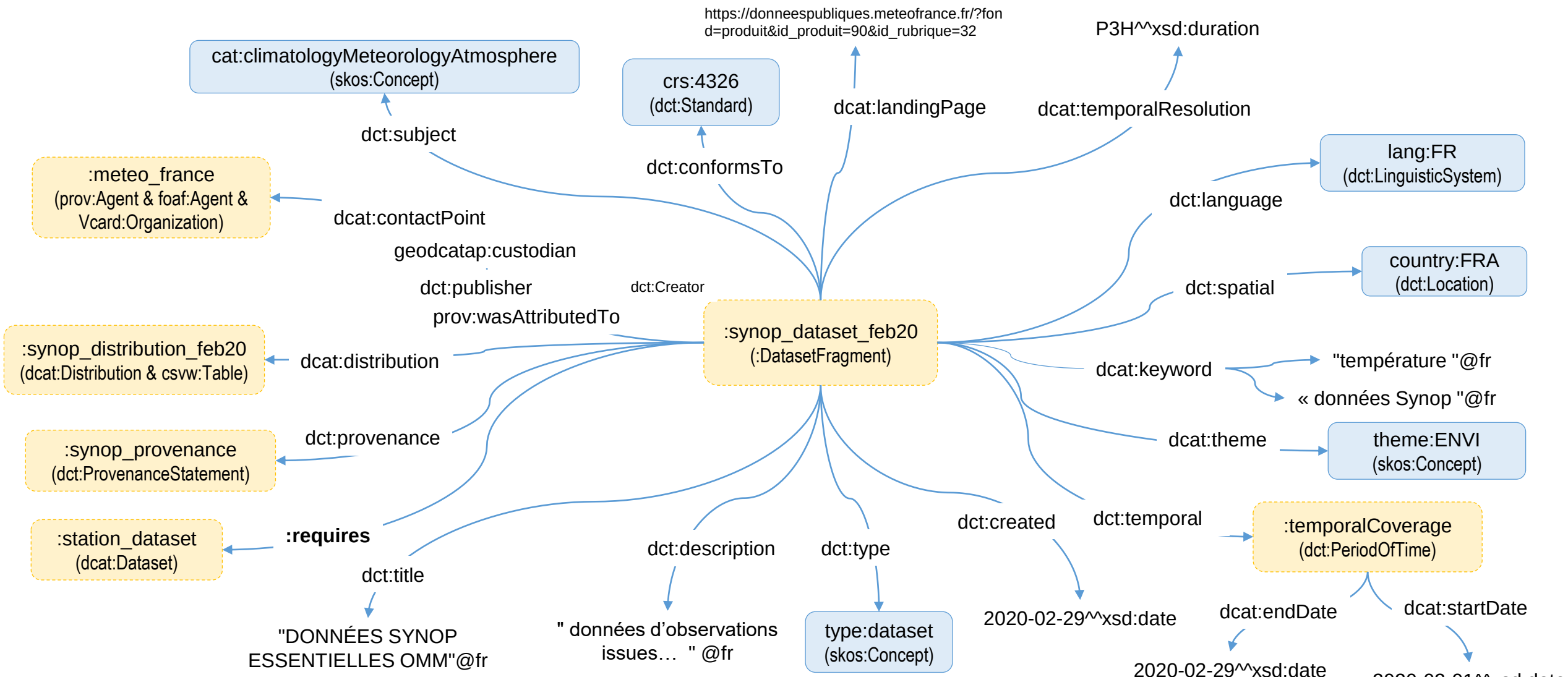
Evaluation

Instanciation des données SYNOP du mois de février 2020

- 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

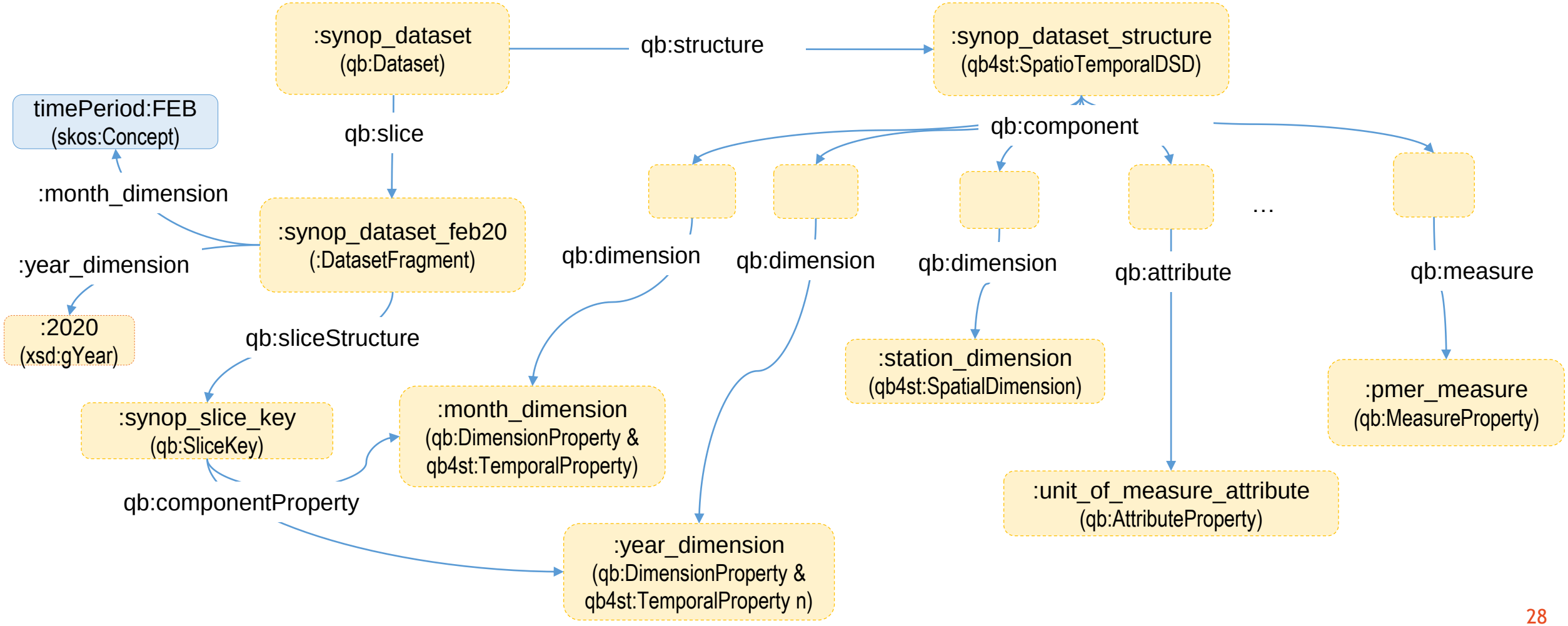


Instanciation des données SYNOP du mois de février 2020



concept — terme appartenant à un vocabulaire contrôlé — instance

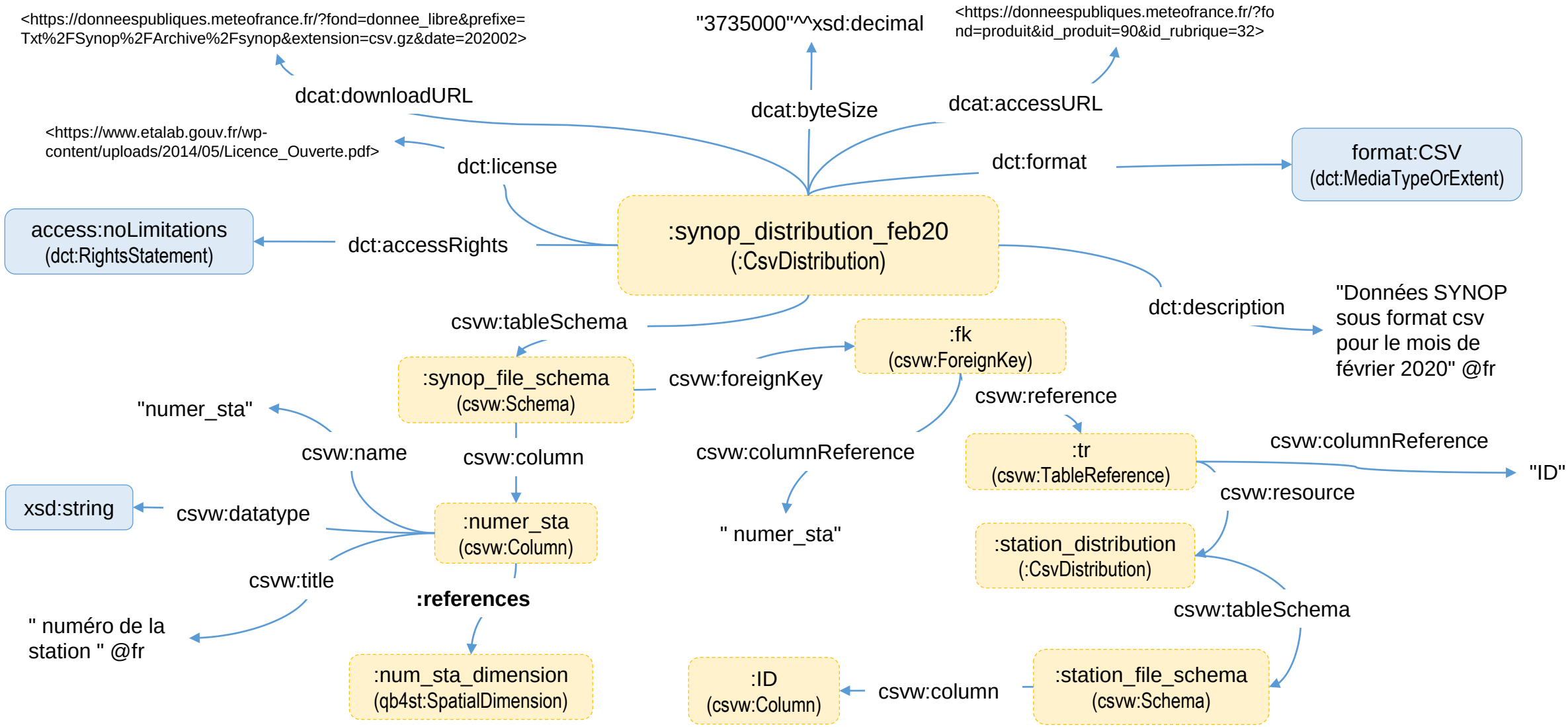
Instanciation des données SYNOP du mois de février 2020



Instanciation des données SYNOP du mois de février 2020

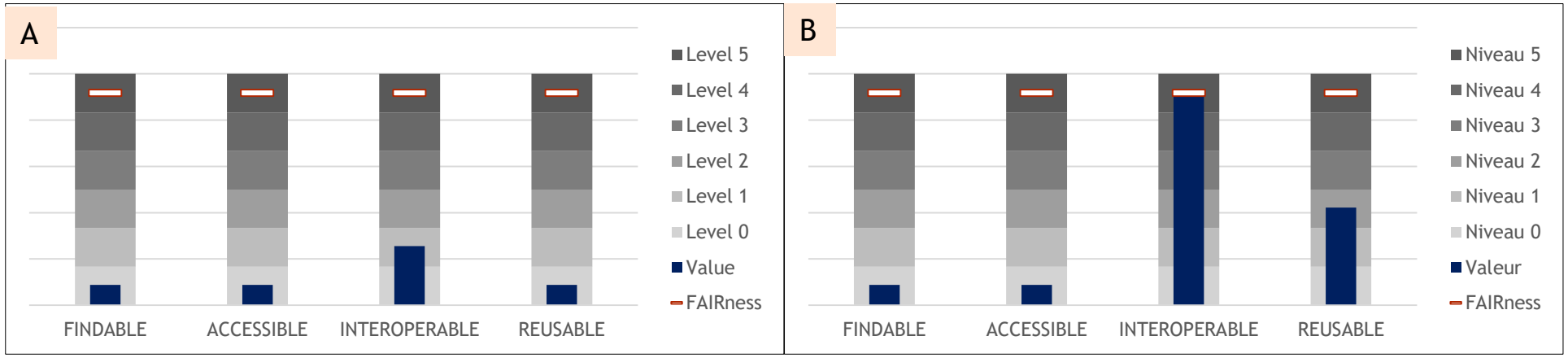
```
:sea-level_pressure a qb:MeasureProperty;
                    rdfs:label "pmer"^^xsd:string;
                    skos:altLabel "pression au niveau mer"@fr;
                    skos:altLabel "sea-level pressure "@en;
                    rdfs:range xsd:int;
                    :unitOfMeasure qudt:Pascal;
                    qb:concept [ a <http://sweetontology.net/propPressure/SeaLevelPressure>, skos:Concept];
                    skos:definition " Air pressure at sea level is the quantity often abbreviated as MSLP or
                    PMSL. Air pressure is the force per unit area which would be exerted when the moving gas
                    molecules of which the air is composed strike a theoretical surface of any orientation. "Mean
                    sea level" means the time mean of sea surface elevation at a given location over an
                    arbitrary period sufficient to eliminate the tidal signals.”@en.
```

Instanciation des données SYNOP



Evaluation du degré de la FAIRisation

- Evaluation du degré de FAIRisation avant et après la génération des métadonnées selon le « FAIR maturity model » de la RDA (FAIR Data Maturity Model WG, 2020)



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

Rapport détaillé sur l'évaluation se trouve sur le lien :
https://hal.archives-ouvertes.fr/hal-03197115/document

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 « 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 des algorithmes de recherche sémantique de datasets qui exploitent la représentation fine du schéma de données.

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.
- Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L. O. B., Mons, B., Schultes, E., Roos, M., & Thompson, M. (2020). A Generic Workflow for the Data FAIRification Process. *Data Intelligence*, *2*(1–2), 56–65.
- Baonza, M. D. F., Pérez, A., & Villazón, B. (2008). *NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse*.
- 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/)