Usage-oriented semantic enrichment of open data

Nathalie Aussenac-Gilles, Amina Annane, Pascal Gaillard, Christophe Baehr

April 8, 2021

UT3/IRIT, UT2J/MSHS-T, Météo France/DESR
`https://www.irit.fr/semantics4fair/`

Table of contents

Contexte

Les données en science météorologique

- Les données et le partage des données font parties de l'ADN constitutif des sciences météorologiques.
- Pour se faire, dès le départ, les codes, les normes, les travaux de l'OMM, le système de transmission mondial, les publications, les échanges scientifiques ont permis ce partage.
- Mais tout ceci a été fait pour l'unique communauté météorologique qui était à la fois productrice et consommatrice de ses données.
- L'Internet, et depuis peu, la Science Ouverte, changent la donne et nos données doivent être démocratisées.
- De plus, les services météorologiques ne sont plus les seuls à opérer des réseaux de mesures et à fournir de la donnée météorologique.

Les données en science météorologique

Pour notre établissement, nous pouvons recenser 3 situations :

- Des données sur nos machines ou sur des serveurs d'équipe fonctionnant en silos non interconnectés.
- Les données publiques de Météo France, partiellement ouvertes
- Les données de recherche déposées sur le pôle de données pour l'atmosphère AERIS, élément de l'IR Data-Terra.

On retrouve ainsi tous les types possibles de situations : données non-ouvertes, en silos, open mais difficiles d'accès ou un début de FAIRisation, mais uniquement avec un point de vue producteur.

Les principes FAIR [3]

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.

Manifeste pour la FAIRisation des données

Agir pour la FAIRisation des données, c'est

- faciliter l'accès à nos recherches
- permettre à nos données d'être interopérables
- améliorer l'analyse des données météorologiques par d'autres communautés
- travailler pour la durabilité de nos sciences
- être un accélérateur des progrès en sciences (météorologiques)

Données vs. Metadonnées

Données et Métadonnées sont à penser et à traiter de manière différentielle.

- Les Métadonnées sont dans notre culture, avec des headers qui ont (presque) toujours été dans nos pratiques.
- Avec le format GRIB, dès 1985, l'OMM a standardisé le header et les Métadonnées.

Mais le Web (Sémantique) a fait évoluer l'usage des Métadonnées et leur format.

- Les Métadonnées (ou annotations) permettent d'associer aux données des informations structurées descriptives permettant leur recherche par des moteurs de recherche
- Plusieurs dimensions sont alors à considérer.

Données vs. Metadonnées

Les données et Métadonnées peuvent alors être 2 fichiers séparés avec des traitements différents.

Avec les Métadonnées, il s'agit de faciliter et améliorer l'accès aux données recherchées par un utilisateur.

Les Métadonnées doivent alors être opens et correctement formatées pour être trouvable par les moteurs de recherche, accessible de manière pérenne, opérable facilement et réutilisable dans le temps.

Ce sont, elles, les clés d'une science des données ouvertes. Mais également pour d'éventuelles données avec accès restreint.

En revanche, il y a un impératif pour les données : avoir un identifiant unique et pérenne, comme l'est le DOI.

Présentation du projet Semantics4FAIR

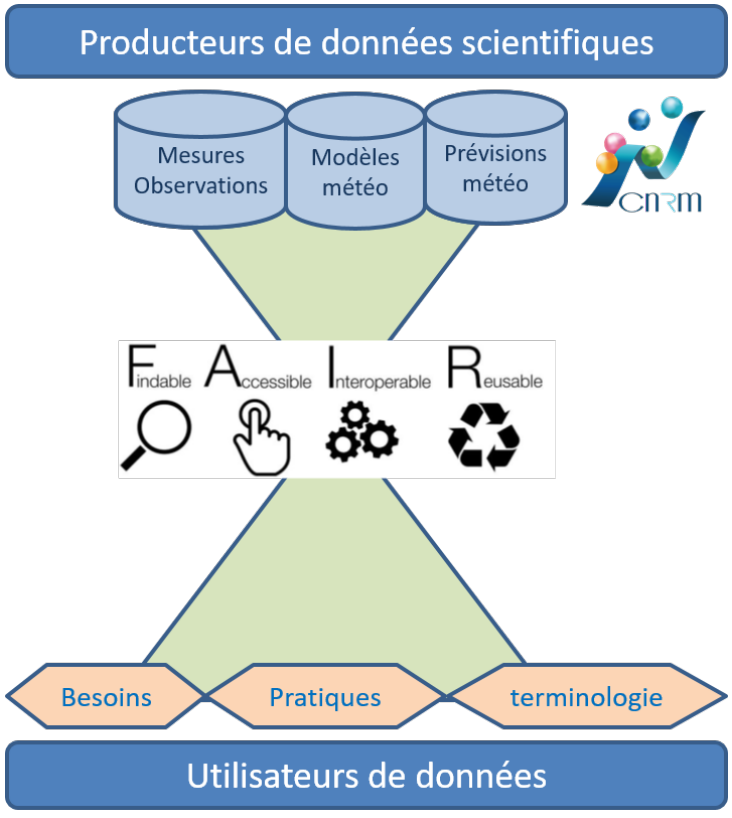
Semantics4FAIR : motivations

Objectifs pour Météo France

- Passer de données Open à des données FAIR
- favoriser leur réutilisation par de multiples communautés scientifiques hors météorologie

Deux approches pour un diagnostic :

- critères FAIR
- étude ergonomique des besoins, pratiques et difficultés des utilisateurs



Une hypothèse : répondre à l'aide des technologies et vocabulaires sémantiques aux besoins des utilisateurs

Semantics4FAIR : approche

Ontologies

Ontologies

Producteur Données

Sources de Données / Fouille de Données

2

RDF

Formalisation

Findable Accessible Interoperable Reusable

(Représentation sémantique)

Semantic Web

1

Formalisation des besoins

Usages / Utilisateurs

Analyse ergonomique des usages et besoins

Le travail en ergonomie de la recherche dans Semantics4FAIR

Ergonomie cognitive et analyse du travail

- L'ergonomie cognitive étudie les interactions avec un dispositif (des données), en prenant en compte la situation (ses objectifs).
- Dans Semantics4FAIR : interactions d'un chercheur en sciences de l'environnement avec des données Météo.
- L'étude inclut l'utilisation et la production.
- Objectif final : rapprocher les deux représentations de la forme des données à produire et utiliser.

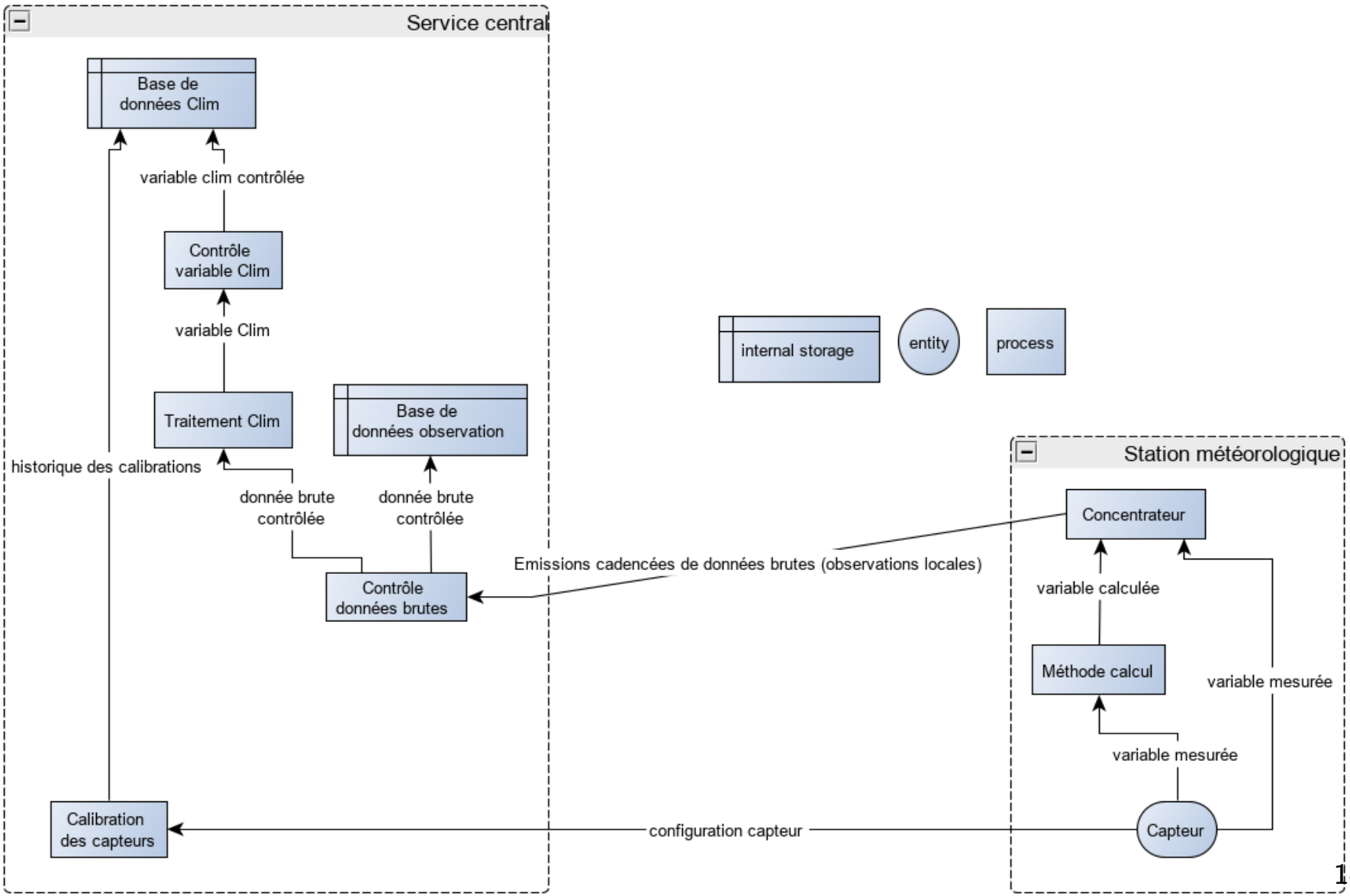
Côté utilisation, entretien et analyse de l'activité réelle (observation de terrain) :

- comprendre les objectifs du chercheur (déclarée et non déclarés) ;
- comprendre quelles données sont utilisées pour y arriver ;
- la manière dont elles sont utilisées en pratique ;
- celles dont il a besoin et celles dont il aurait besoin, celles qu'il n'arrive pas à obtenir ;
- faire un bilan des données disponibles et leur forme.

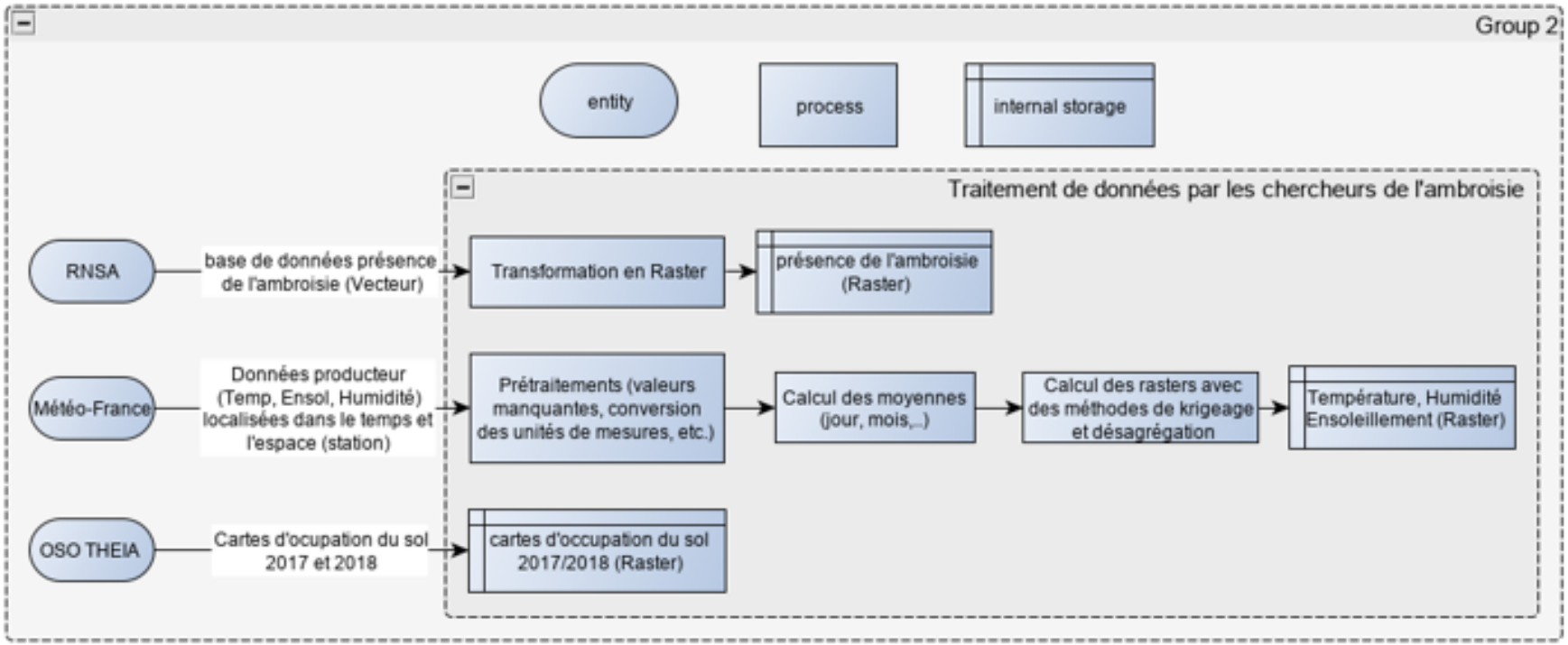
Coté production, entretien et observation (normalement...)

- comprendre l'objectif (déclaré ou non) de la création de données par Météo France (en interne et surtout dans cas en externe) ;
- comprendre ce que les créateurs de données pensent de l'usage final de leurs données par des utilisateurs ;
- comprendre comment Météo France produit les données ;
- comment Météo France formule les données produites ;
- comment elles sont présentées et pourquoi.

Flux de données expliqué à l'utilisateur



Flux de traitement des données par l'utilisateur



Evaluation de la FAIRness des données publiques de Météo France

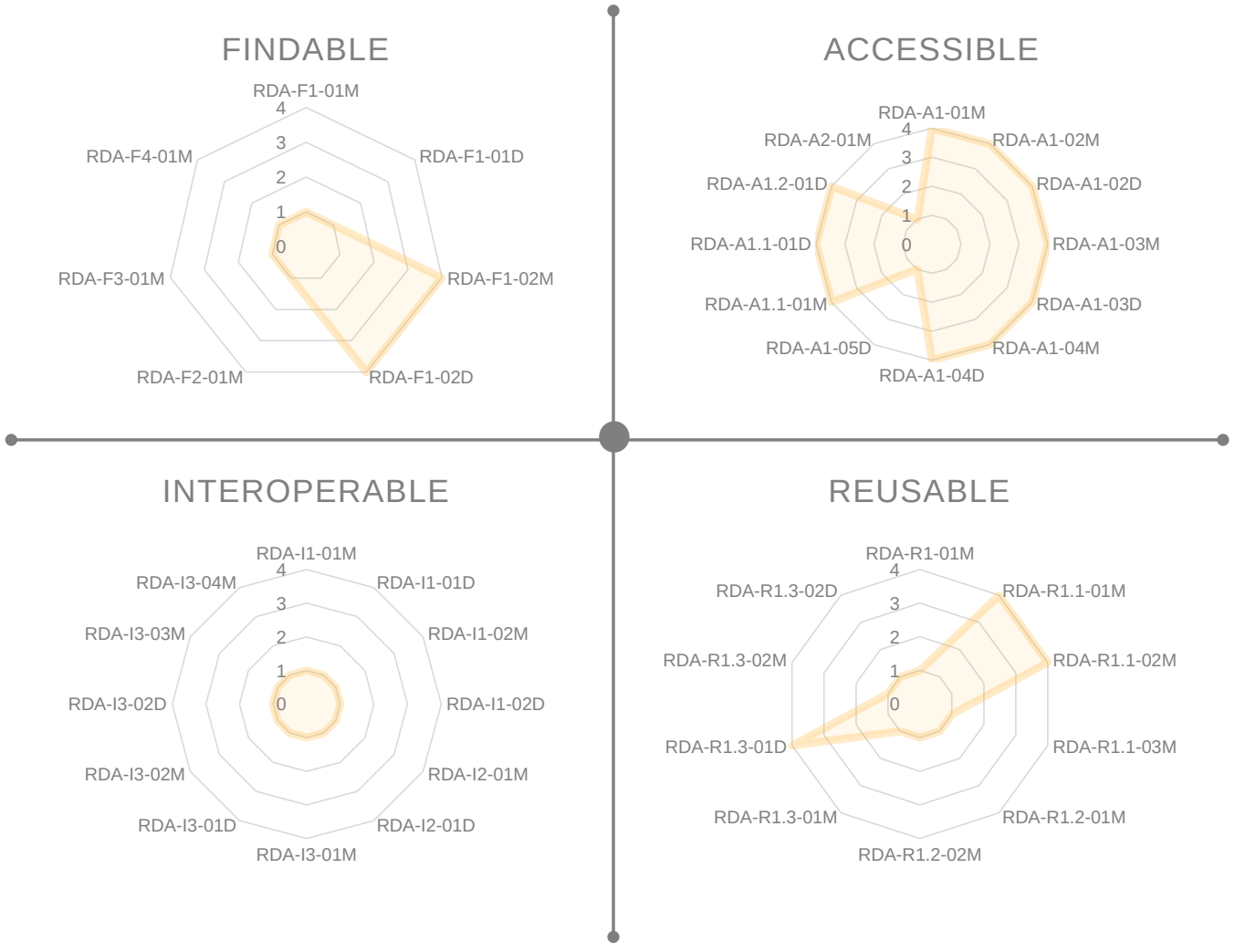
Modèle d'évaluation:FAIR Data Maturity Model [2]

Three components:

- Indicators: the individual aspects of FAIRness that are evaluated
  - 41 indicators extracted from FAIR principles
  - e.g., "RDA-F1-01M Metadata is identified by a persistent identifier"
- Priorities: the relative importance of the indicators (essential, important, useful)
- Evaluation methods: the way that the results of the evaluation of the indicators can be given a value
  1. Measuring progress
  2. Pass/fail

Résultats d'évaluation selon le modèle de la RDA (1/2)

Jeu de données évalué: DONNÉES SYNOP ESSENTIELLES OMM

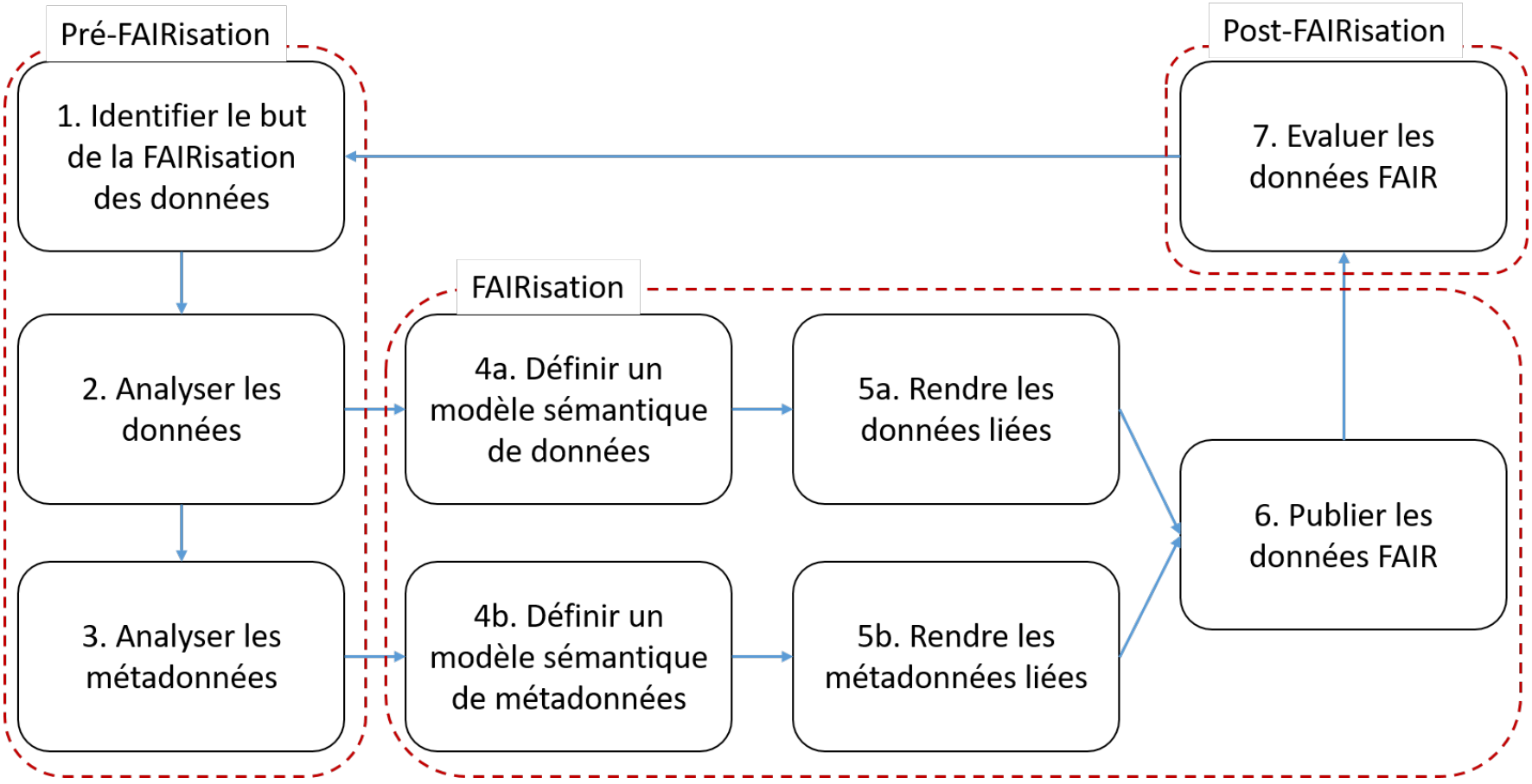


Résultats d'évaluation (2/2)

- Findable
  - Pas d'identifiants persistants et pérennes (e.g., DOI)
  - Pas de métadonnées sémantiques qui facilitent l'indexation des données.
- Accessible
  - Pas d'API pour accéder aux (méta)données automatiquement, mais passage par le formulaire
- Interoperable
  - Pas de modèle sémantique de (méta)données
  - Pas de métadonnées sémantiques
- Réutilisable
  - Pas de métadonnées sémantiques sur la provenance des données
  - Pas de métadonnées sémantiques pour interpréter les données, e.g., une définition très brève des acronymes du fichier des données Synoptiques

Vers la FAIRification des données météorologiques

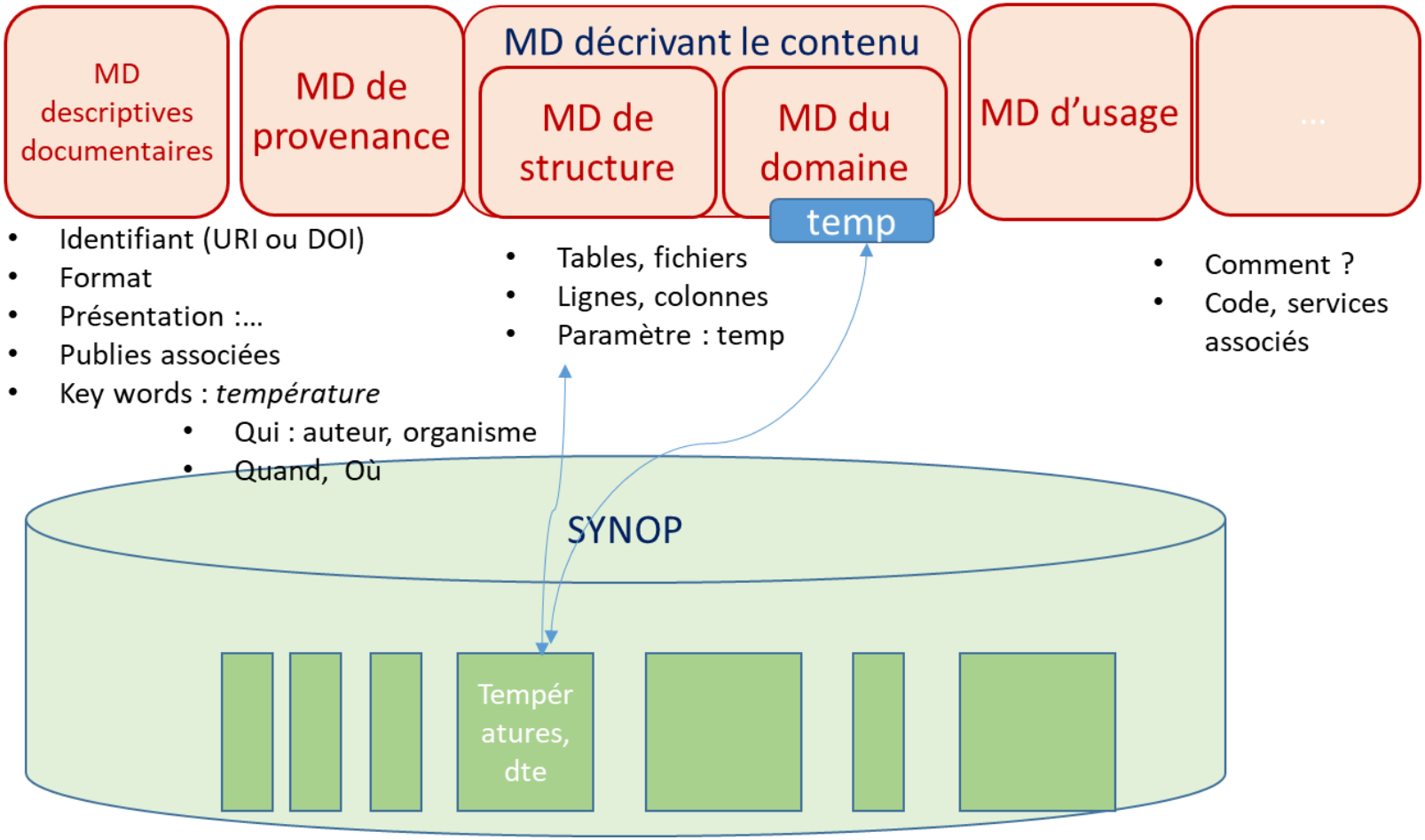
Processus de FAIRisation [1]



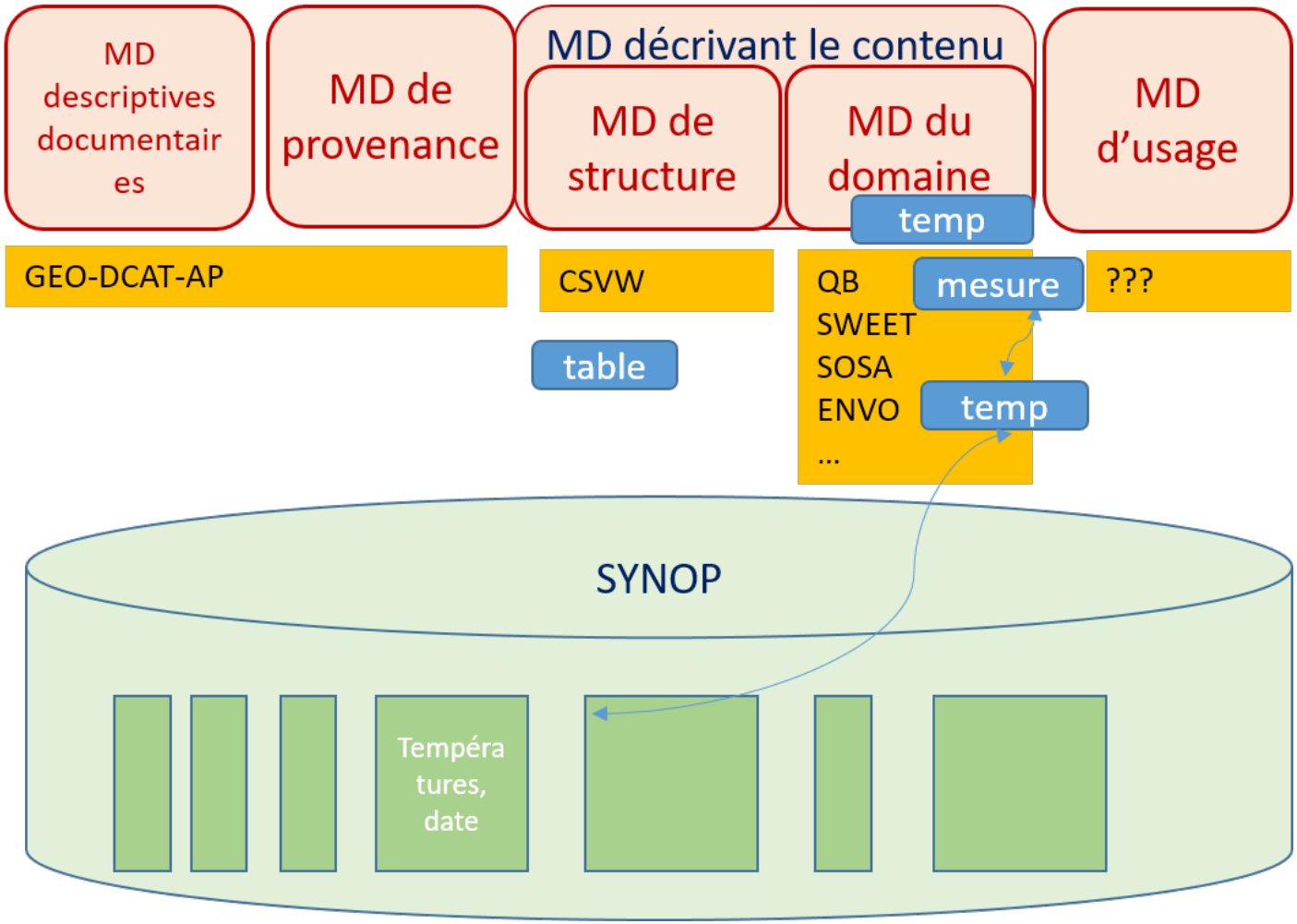
une stratégie FAIR en plusieurs étapes

1. Déterminer les MD adaptées (Diversification des méta-données)
2. Choisir des vocabulaires adaptés
   - point de vue Météo France
   - points de vue utilisateurs
3. Décrire chaque jeu de données
   - identifiants unique pour chaque jeu de données et ses métadonnées (DOI ou URI)
   - représentation sémantique des méta-données données
4. Construire un portail / référentiel de jeux de données météo
5. Diffuser les meta-données sur d'autres portails (ex : celui de UFT-MIP)

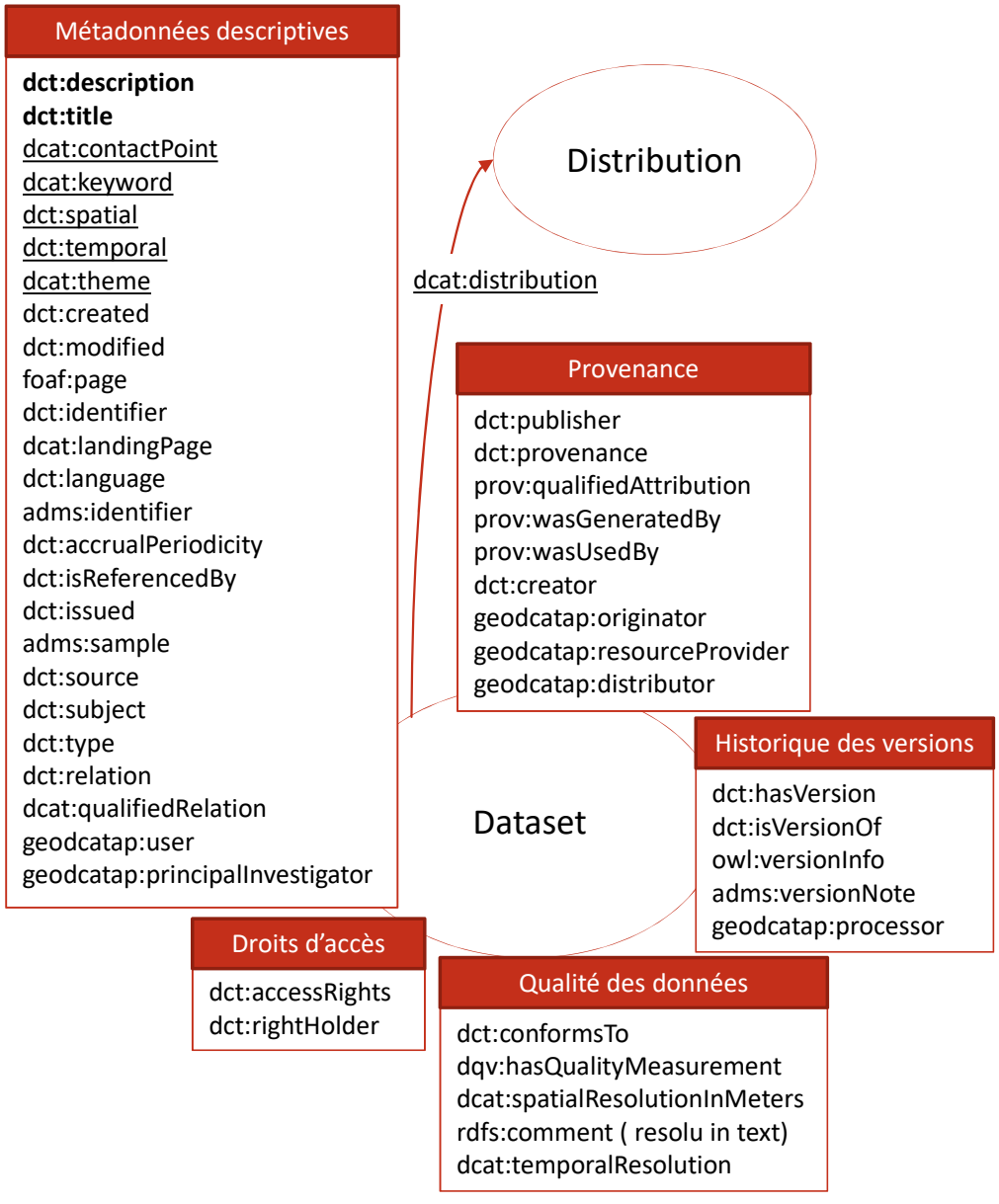
Diversité des métadonnées



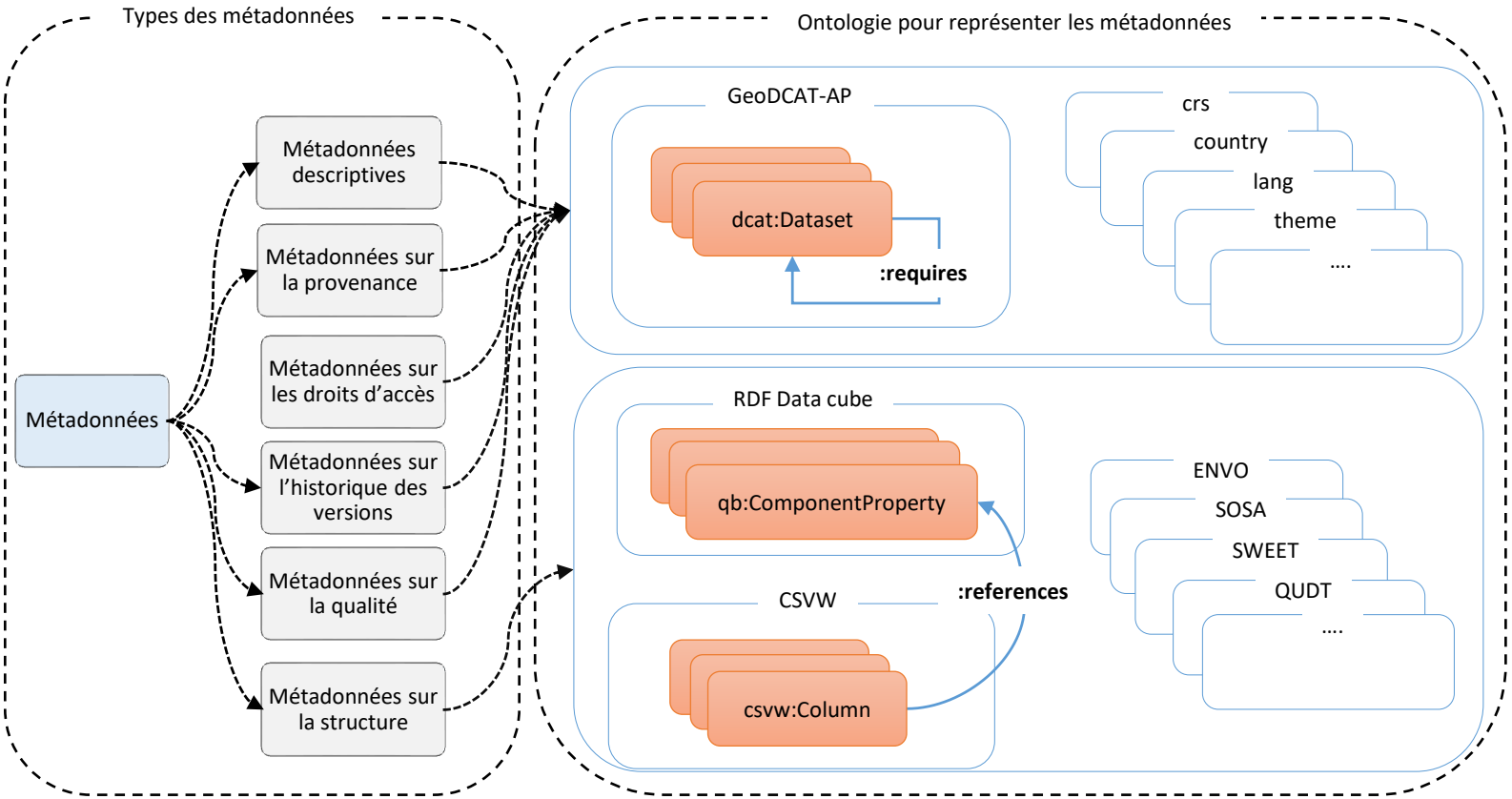
Vocabulaires de métadonnées



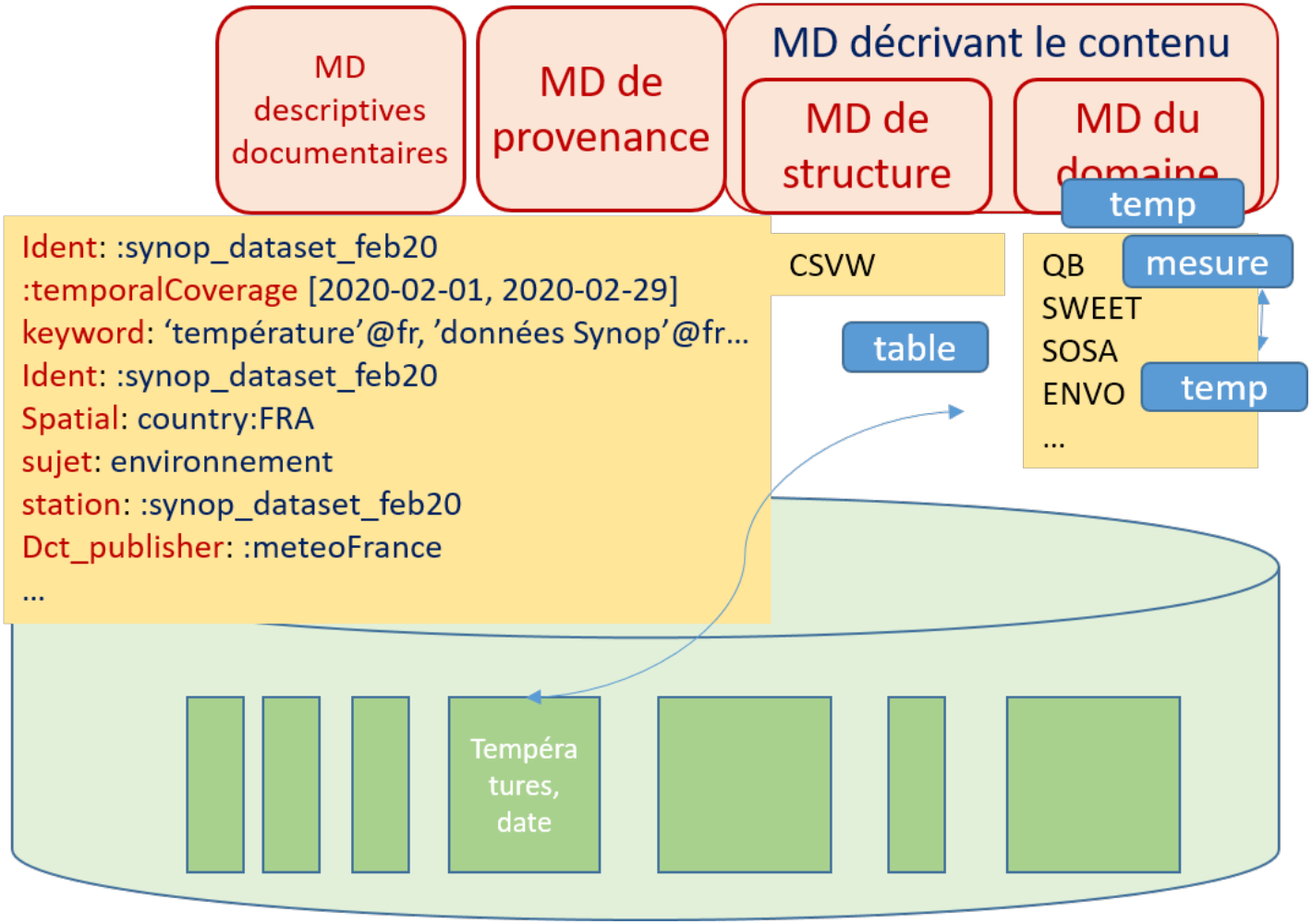
Les différentes facettes de Geo-DCAT-AP



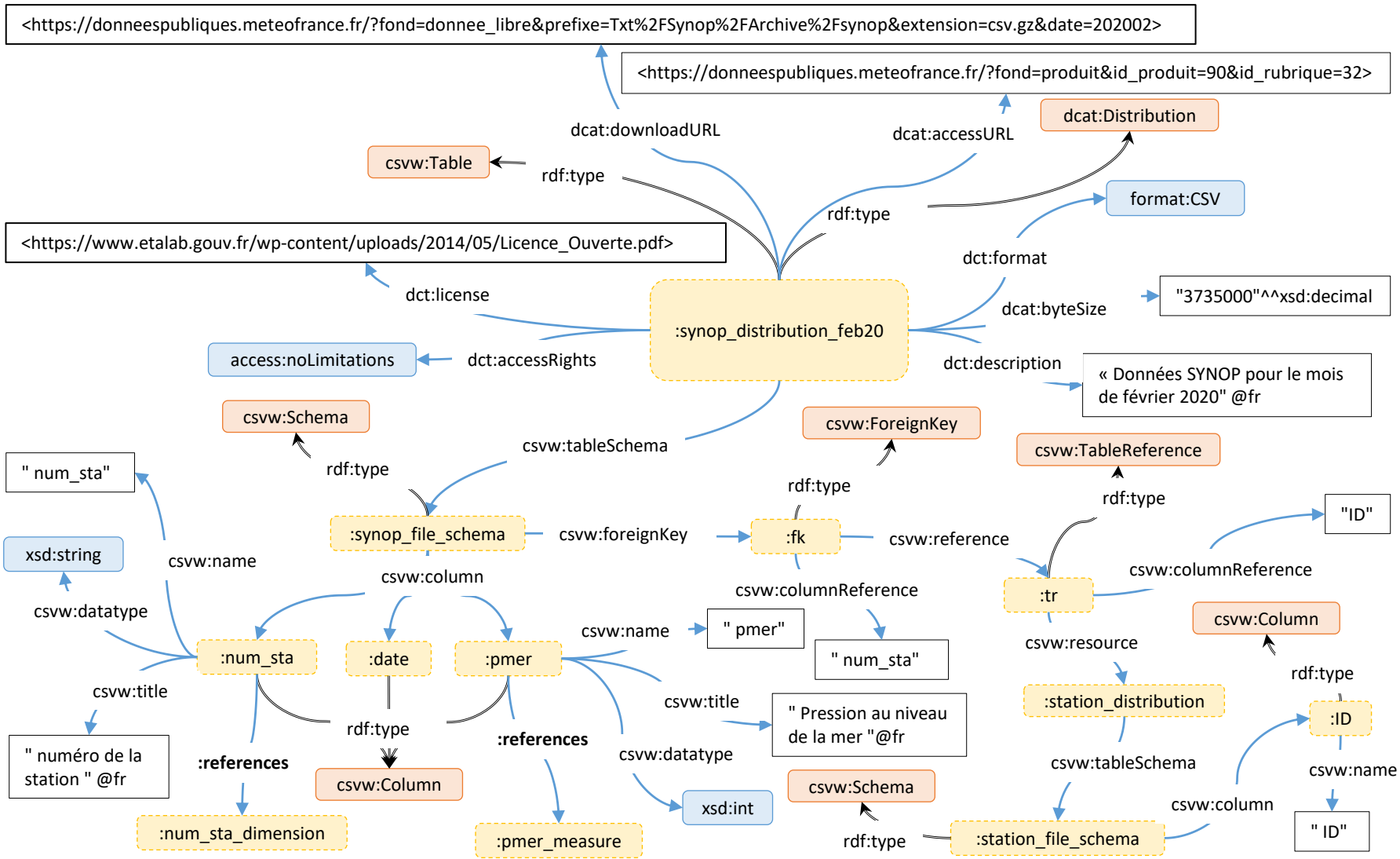
Mise en relation des vocabulaires pour décrire SYNOP



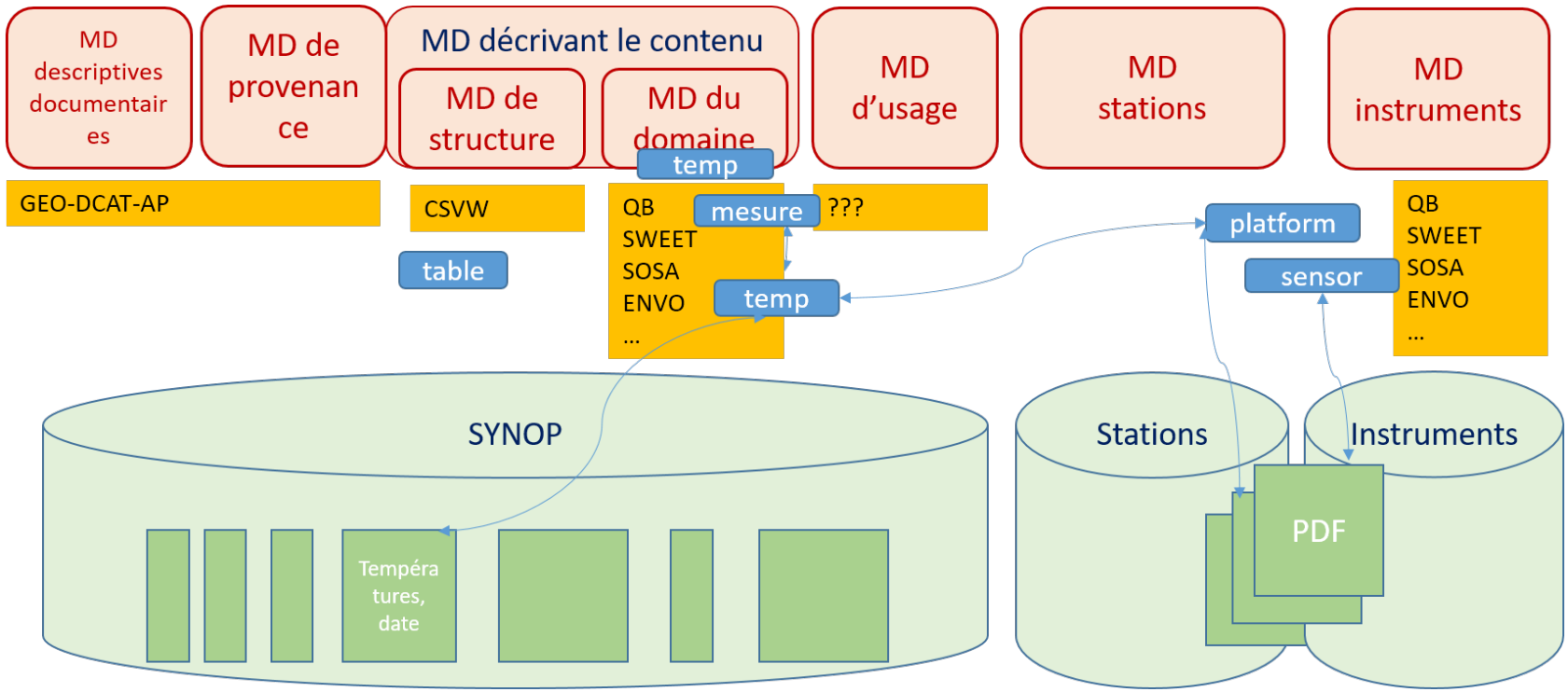
Décrire SYNOP avec des métadonnées sémantiques



Graphe RDF de métadonnées de SYNOP



Vers un meilleur niveau de FAIRisation des données SYNOP



Limites

- Les fiches de postes (des stations météorologiques) sont très intéressantes pour documenter les données SYNOP (les différentes localisations, les instruments utilisés, etc.) mais en format PDF et contiennent beaucoup de données hétérogènes => pas facile à exploiter. `https://donneespubliques.meteofrance.fr/?fond=contenu&id_contenu=37`
- Ces données sont moins volumineuses, on peut les instancier et les publier sur le LOD?

Du jeu de données au portail de données sémantisées

- Construction d'un portail / référentiel de jeux de données météo
  - Décrire ainsi plusieurs jeux de données (ouvertes ou non)
  - Définir un protocole d'accès aux données (contact, requêtes, ...)
  - Logique d'usage à recueillir et modéliser
- Diffuser les meta-données sur d'autres portails
  - Portails de domaine (i.e. Terra Data), de site (i.e. UFT-MIP)
  - Portails des données de la recherche (i.e. R3DATA), Portails généraux (Dataset Search)
  - Démarche dans 2 directions : push (se faire connaître) / pull (être repéré par d'autres)

Conclusion et Intérêt pour le CNRM

bilan du travail réalisé

- Travail réalisé et résultats réutilisables
  - Modèles (ontologies) pour décrire les MD
  - Exemple de description d'un jeu de données opérationnelles (SYNOP)
  - Template (patron) pour la saisie de MD pour des jeux de données météo
  - Modèle et vocabulaire des besoins exprimés par le chercheur en biologie
- En cours :
  - Prototype permettant de charger des ontologies / vocabulaires de domaines pour décrire des MD
  - Possibilité de définir de nouveaux templates de domaines
  - Description de MD d'autres jeux de données
  - Mise en relation de la vision utilisateur avec la description des données des producteurs

Des défis pour une recherche collaborative (1)

- Aller plus loin dans l'aide à établir la communication entre vocabulaire utilisateurs / producteurs de données
  - Expliciter les logiques d'usage
  - Vers des méta-données d'usage (qui a déjà utilisé ces données ? pour en faire quoi ? à l'aide de quels logiciels ?)
  - Lien avec IA
  - Ajouter d'autres profils d'utilisateur (processus de recherche et centres d'intérêt propres), définis en collaboration avec le CNRM pour les usages CRNM , et d'autres utilisateurs

Des défis pour une recherche collaborative (2)

- Vers des métadonnées de qualité (cf Geo-DCAT-AP)
  - Qualité a priori estimée par le producteur : nettoyage des données, précisions sur données manquantes etc)
  - Qualité a porteriori, évaluée par les utilisateurs
- Reste à faciliter le "croisement" de données venant de disciplines différentes
  - Opportunité : use-case de DataNoos
  - Gestion de workflows et de processus
  - Ambition à mettre en place avec tous les partenaires

Enjeux pour les données de la recherche

- Données de la recherche vs données opérationnelles
  - Données de campagnes : fenêtre temporelle plus réduite, objectif spécifique, données adaptées à l'objectif
  - Besoin de les documenter ; peu réutilisées hors communauté
  - Méthode, format, principe de collecte réutilisables
  - et les données aussi !
- Ouvrir les données :Nombreuses perspectives pour le CNRM
  - Valoriser les recherches
  - Valoriser les données
  - Initier de nouvelles manières de faire de la recherche à partir des données
  - Permettre de nouvelles collaborations

Merci de votre attention - Des questions ?

Infos sur DataNoos :

https://datanoos.univ-toulouse.fr/

RDA

La RDA est une organisation communautaire liée à la recherche, créée en 2013 par la Commission Européenne, la National Science Foundation et le National Institute of Standards and Technology aux Etats-Unis, et le Ministère Australien de l'Innovation. Sa mission est de construire les ponts sociaux et techniques pour que les chercheurs et les innovateurs de différents pays et de toutes les disciplines puissent partager ouvertement leurs données pour relever les grands défis de la société.

References i


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.


F. D. M. M. W. G. RDA.
FAIR Data Maturity Model. Specification and Guidelines, June 2020.


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.