Mise en oeuvre de ressources sémantiques dans dans l'IR Data Terra :

Une approche sémantique pour la découverte des données pluri-disciplinaires du système Terre

Jean-Christophe Desconnets (IRD, ESPACE-DEV)
Direction Technique Data Terra

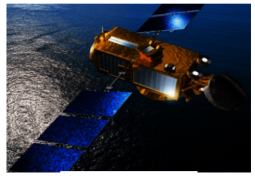
Plan

- L’infrastructure de recherche Data Terra ?
- Les enjeux d’interopérabilité sémantique
- Approche proposée : un modèle de métadonnées centré utilisateur
- Déclinaison opérationnelle
  - Privilégier les pratiques actuelles
  - Travaux en cours

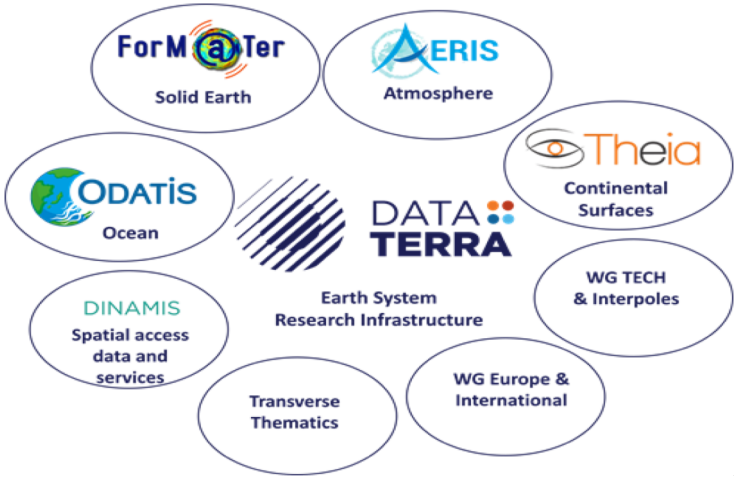
Infrastructure de recherche Data Terra

Plate-forme intégrée (en devenir) de données et de services distribuées pour l'observation et la compréhension du système terre et de l'environnement sur l'ensemble du cycle de la donnée, de son acquisition (spatiale, sols, in-situ) jusqu'à ses multi-usages (qualification/validation, stockage, traitements/extraction de connaissances, produits, services, …)

- **Faciliter l'accès et l'utilisation** des **données et produits** de qualité sur l'ensemble des **compartiments du système Terre**
  => Données spatiales, aéroportées, sols, in-situ
- Développer des services de visualisation et de traitements adaptés aux besoins, à l'accroissement de la volumétrie et aux avancées technologiques
- Favoriser la mutualisation, interopérabilité, émergence d'approches multi- et inter-disciplinaires
- Servir les communautés scientifiques, les acteurs de l'action publique et de l'innovation
- Mettre en œuvre une stratégie nationale, européenne et internationale




Organisation de Data Terra



Data Terra est fondée sur quatre pôles correspondant à chacun des grands compartiments du Système Terre : surfaces continentales, atmosphère, océans et terre solide, complétés par des services transverses.

- 26 organismes et universités
- 4 pôles de données
- 6 services (DINAMIS) et groupes de travail transversaux
- 30 Centres de Données et de Services (CDS) et Infrastructures de données spatiales (IDS)
- 25 Consortium d'Expertise Scientifique
- 187 ETPT / 400 scientifiques, ingénieurs et techniciens
- 33 M€ (2016), 39 M€ (2017), > 40 M€ (2019)
- Plus de 500 produits et services, plus de 15000 utilisateurs
- 50 000 To (2018) ; 100 000 To (2022/2023) ; 150 Peta (2025)

Les enjeux d'interopérabilité sémantique

Extraits du document du GT Science Data Terra, octobre 2020 :

"Les **verrous des défis prioritaires aux interfaces des pôles** thématiques ont permis d'identifier des besoins communs tant pour la mise à disposition de données qu'à la **capacité de combiner des données de différentes sources** (satellites, in situ, modélisation, simulation) et de différentes échelles. "

Défi : une climatologie de référence du milieu côtier Français

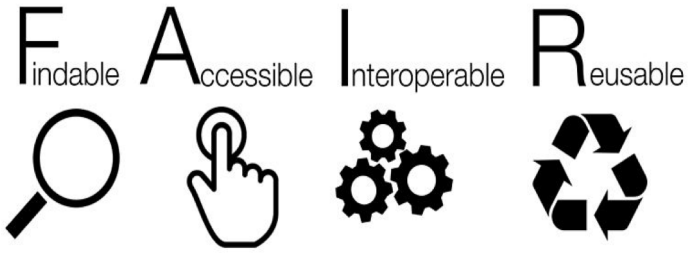
Verrous : Besoin d'accéder à une combinaison de l'ensemble des données acquises à l'interface continent-océan (in situ, satellite) ;

Défi : Observation et modélisation des impacts des changements globaux

Verrous: Pouvoir accéder à l'ensemble des données (non seulement celles produites par les labos et les SNOs, mais aussi par les gestionnaires) sur un continuum donné pour favoriser l'approche intégrée ; besoin d'outils pour compiler, mettre à disposition, combiner observation in situ, satellites, modèles ;

Les besoins : une découverte des données transversale aux compartiments du système

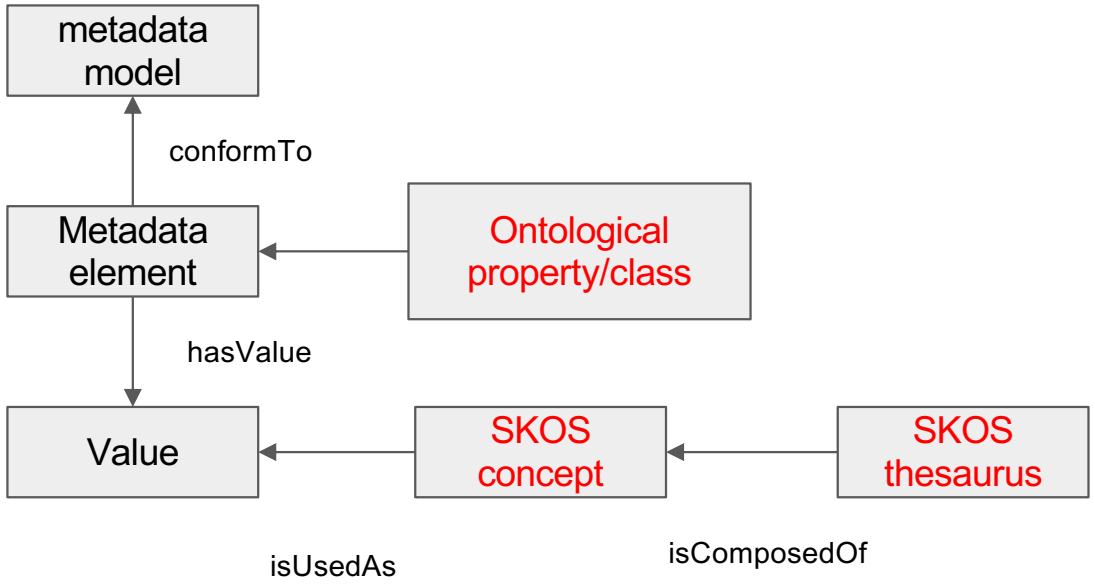
- **Découverte** des **données**, des **services** et des **traitements** qui **traversent les compartiments** du système Terre
- **Vue de l'ensemble** des données et services pour qu'ils puissent être interrogés et exploités de manière **interopérable et tranversale**



Approche proposée

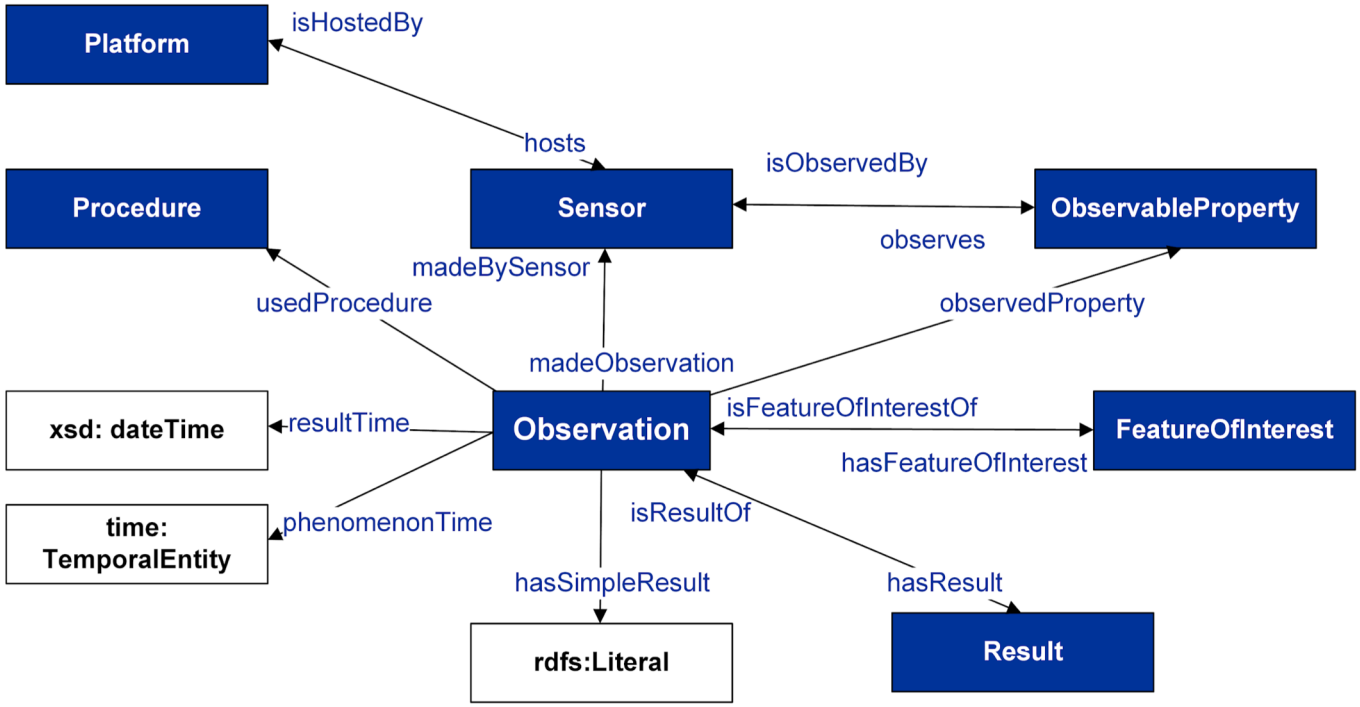
une ontologie d’observation comme modèle de métadonnées pour lier les données aux interfaces des pôles

Positionnement de l’approche



notion de modèle de métadonnées et liens avec les ontologies

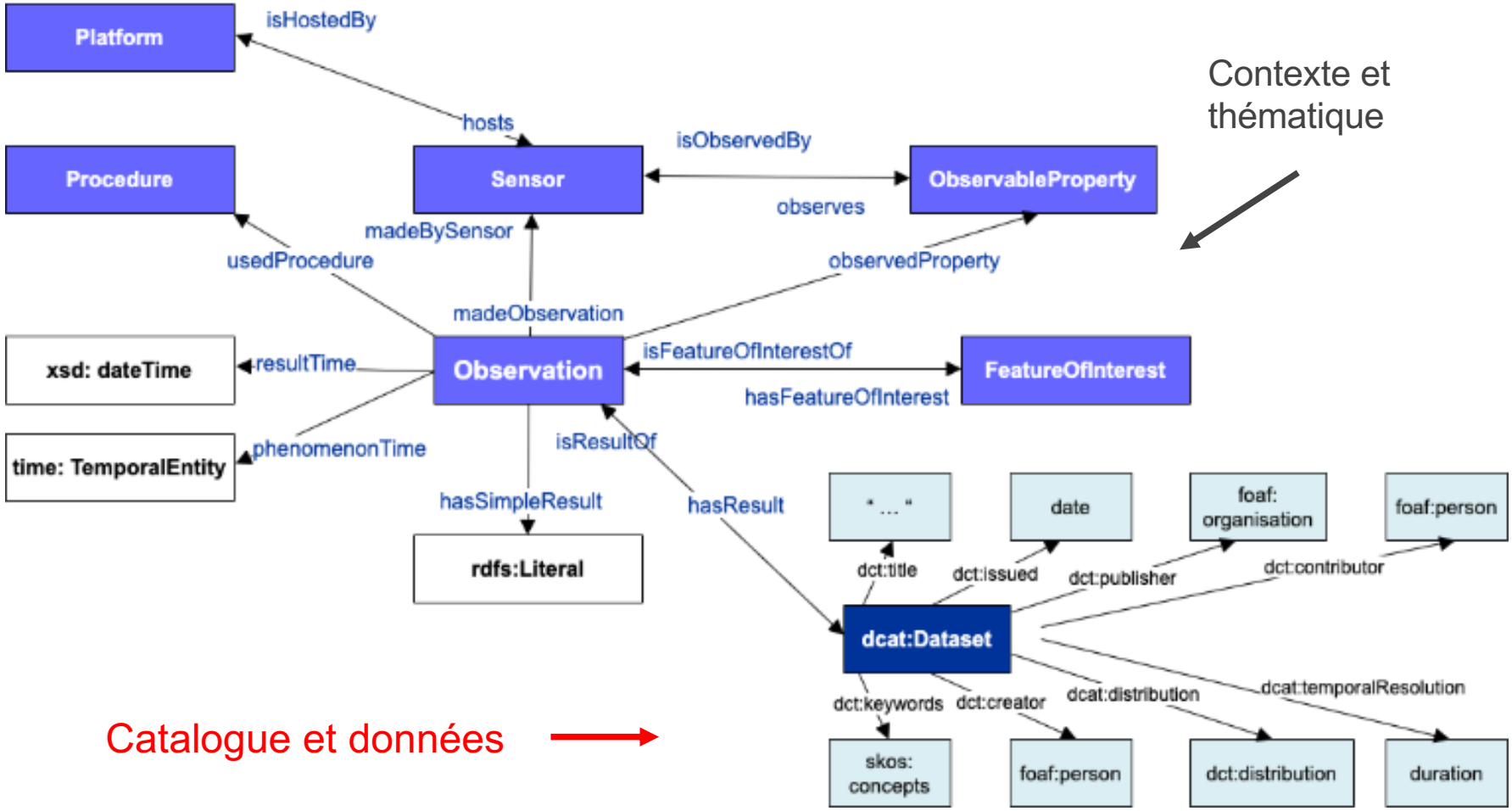
Le paradigme d'observation pour modéliser la connaissance du Système Terre et des dispositifs d'acquisition et diffusion



SOSA : Sensor, Observation, Sample and Actuator : rattache les données au contexte d'acquisition, les objets d'étude et leurs caractéristiques propriétés

Complétée par l'ontologie de catalogue DCAT

Data Catalog Vocabulary pour décrire les **caractéristiques** des jeux de **données**



A user-centric metadata model to foster sharing and reuse of multidisciplinary datasets in environmental and life sciences

Valentina Beretta [a], Jean-Christophe Desconnets [a], Isabelle Mougenot [b], Muhammad Arslan [a], Julien Barde [c], Véronique Chaffard [d]

https://doi.org/10.1016/j.cageo.2021.104807

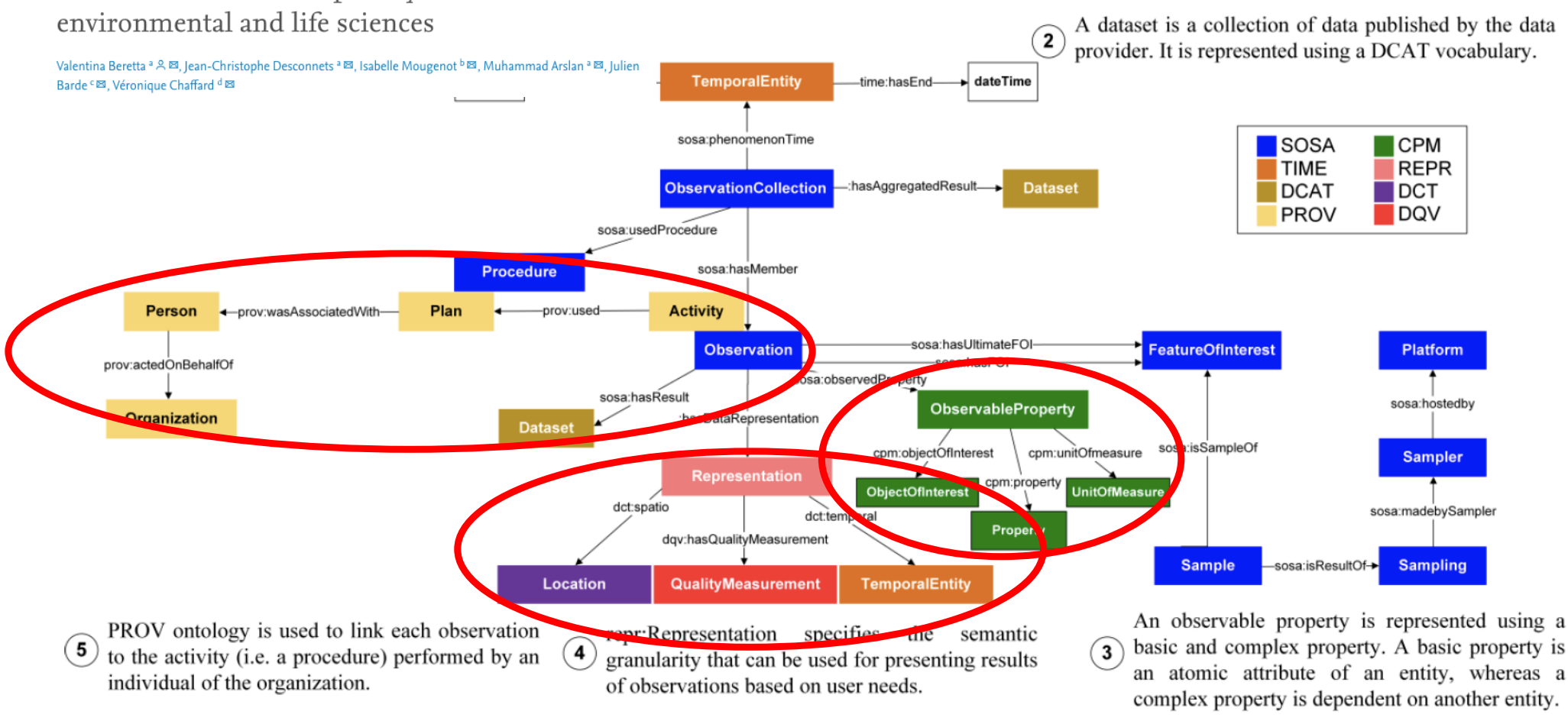
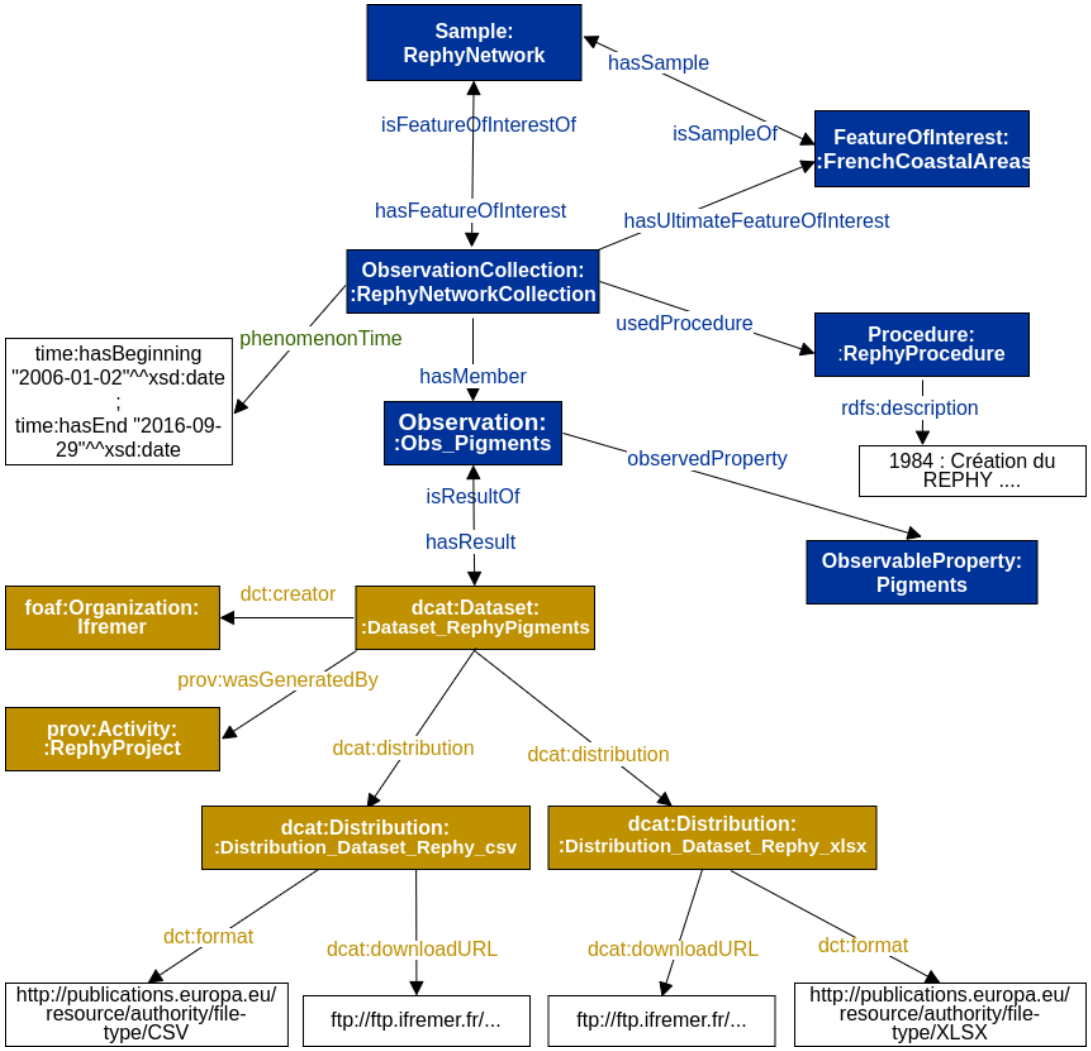


Fig. 1. Domain-neutral dataset discovery model [38]

Annoter les données avec les concepts clés de l'ontologie

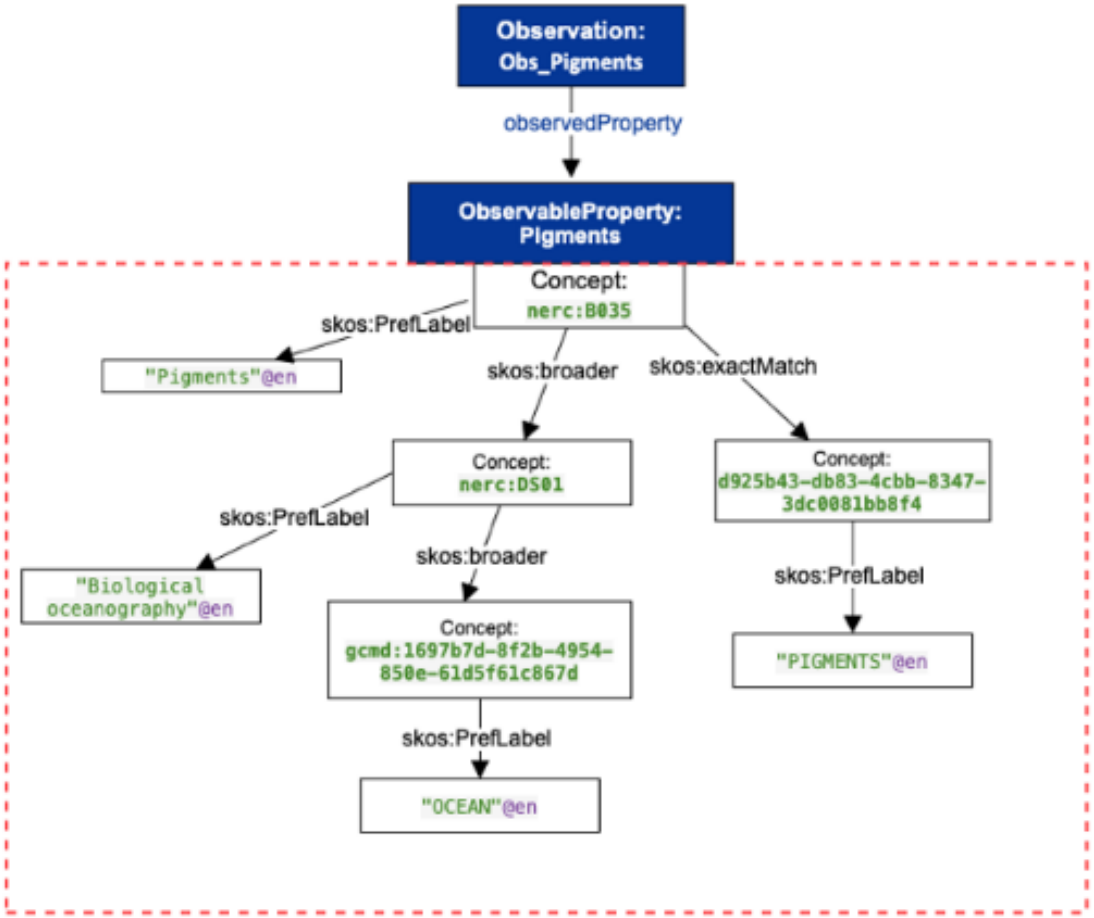
Réseau d'Observation et de Surveillance du Phytoplancton et de l'Hydrologie dans les eaux littorales



REPHY Network
https://www.odatis-ocean.fr/donnees-et-services/acces-aux-donnees/catalogue-complet#/metadata/c5dd9e6f-b45f-4cd6-984d-95d13c8d1f1f

Enrichir ces concepts avec les vocabulaires disciplinaires

Et exploiter les alignements entre **vocabulaires disciplinaires** pour naviguer sur les concepts aux interfac des compartiments



Déclinaison opérationnelle

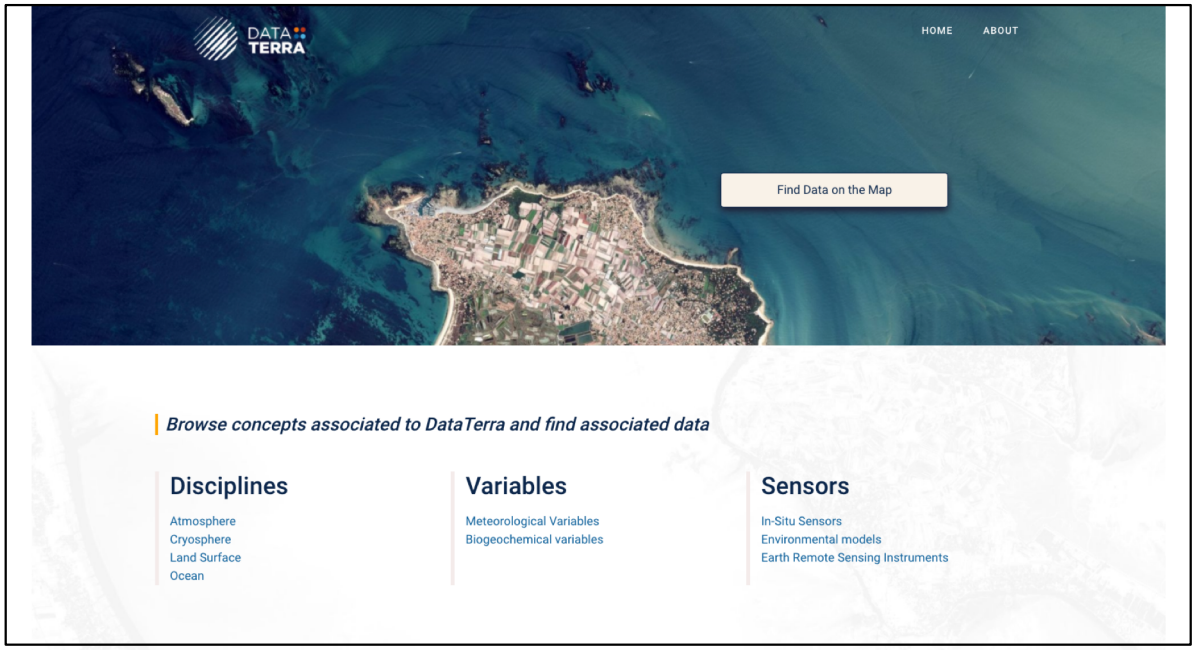
découvrir les données en naviguant dans les compartiments de la Terre, les capteurs et les propriétés observées

Le portail de découverte et d'accès des données Data Terra (PoC)

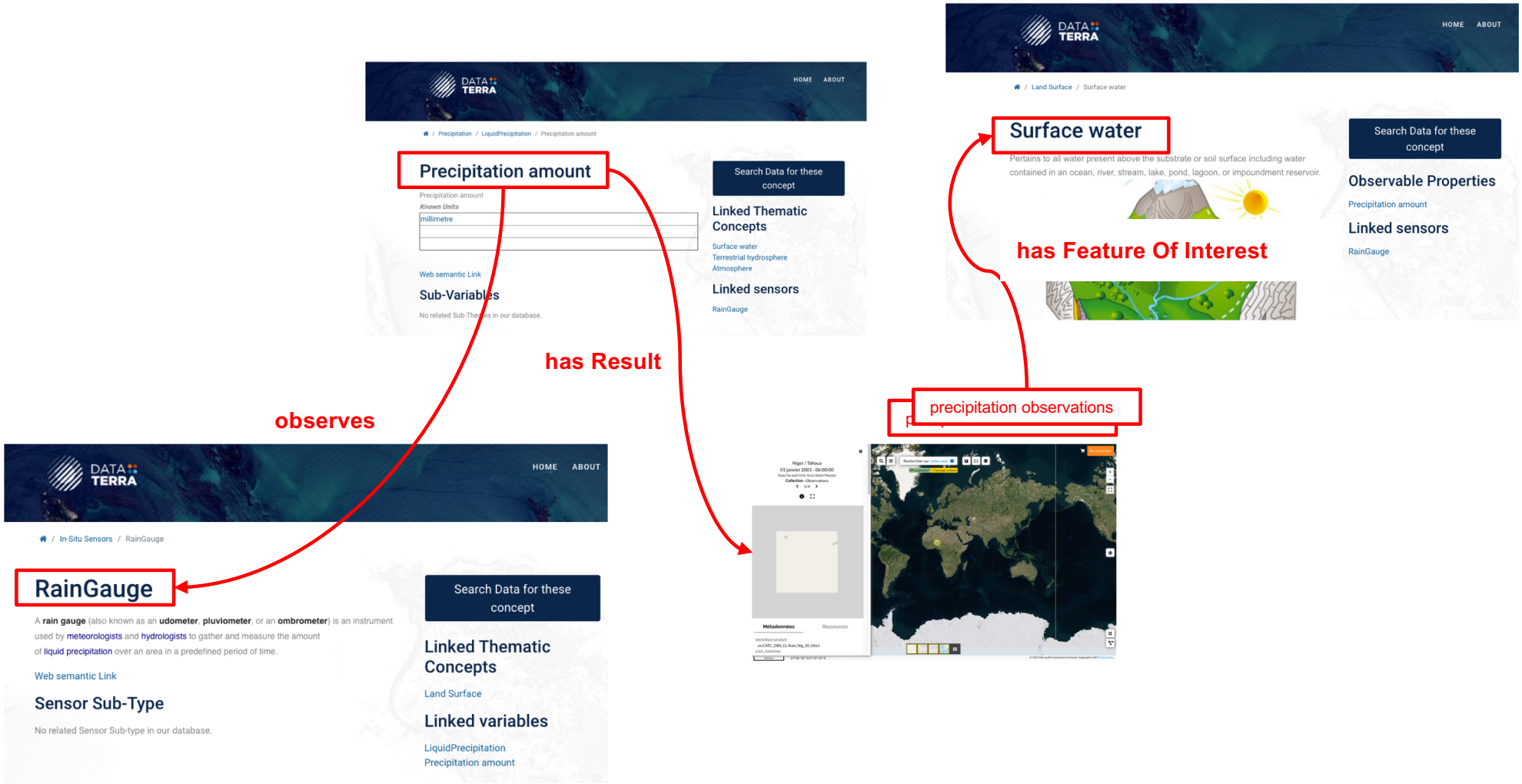
Un climatologue veut réaliser des réanalyses des données climatiques. Il cherche des données de précipitations *in-situ* en afrique subsaharienne

1 - Il interroge le catalogue ou
2 - il part à la découverte des données

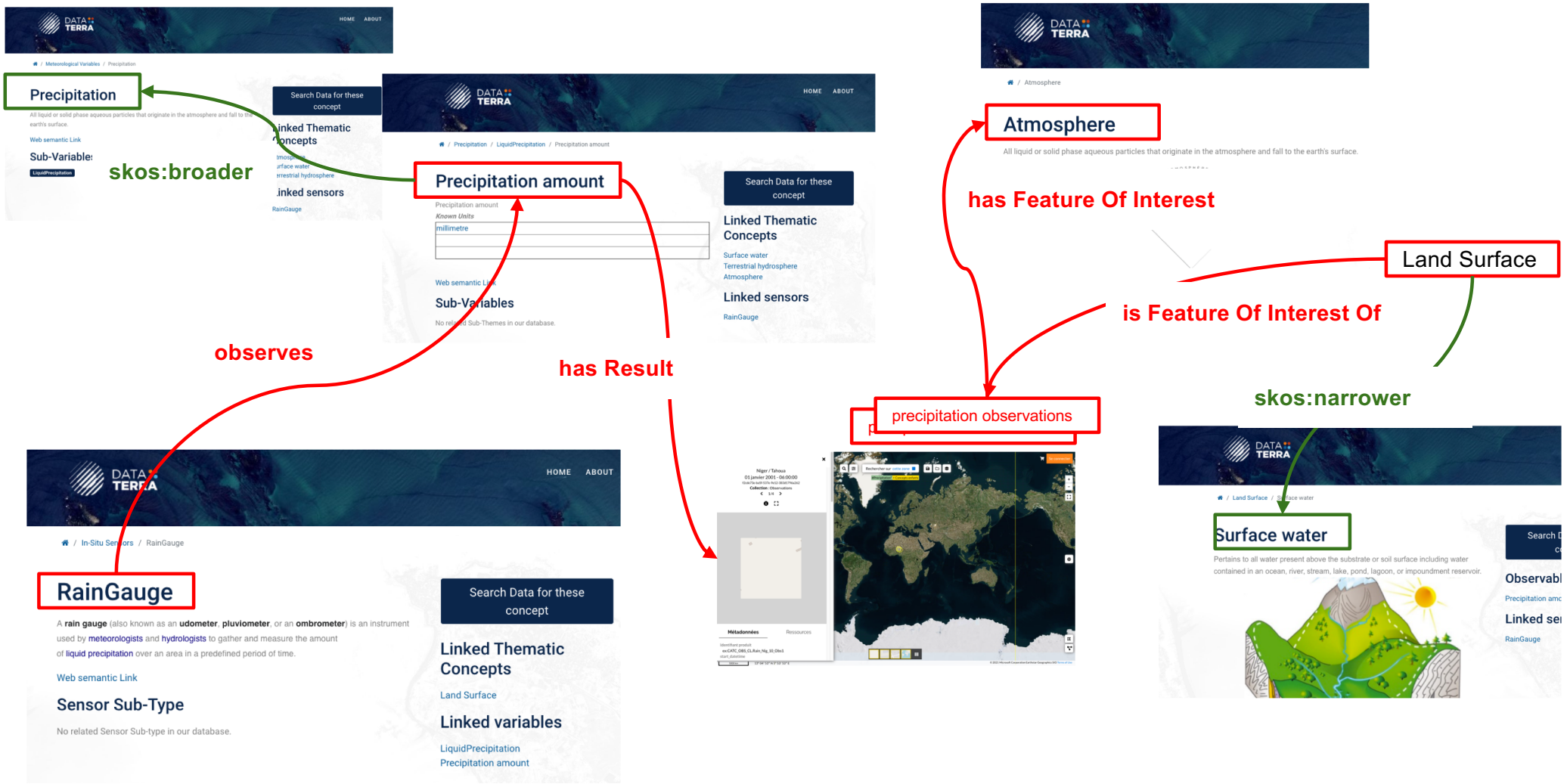
https://dataterra.geomatys.com/



Sous le capot



… enrichi par les vocabulaires disciplinaires



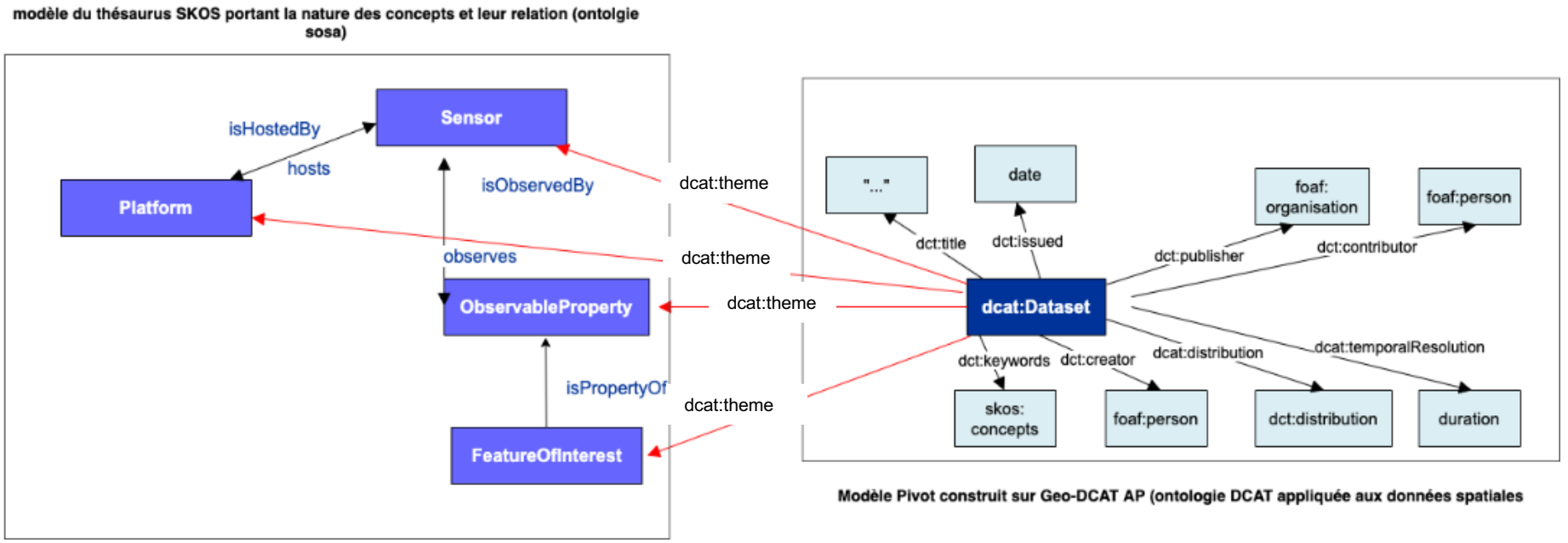
Limites et contraintes du changement de paradigme dans les catalogues

- Modèle spécifique Data Terra, EOSC préconise DCAT et ses déclinaisons GEO-DCAT AP
- Difficultés de transformation des catalogues existants (annotation avec SOSA)
- Mise en œuvre longue : demande en même temps une maturité sur les catalogues et la formalisation/standardisation des terminologies disciplinaires

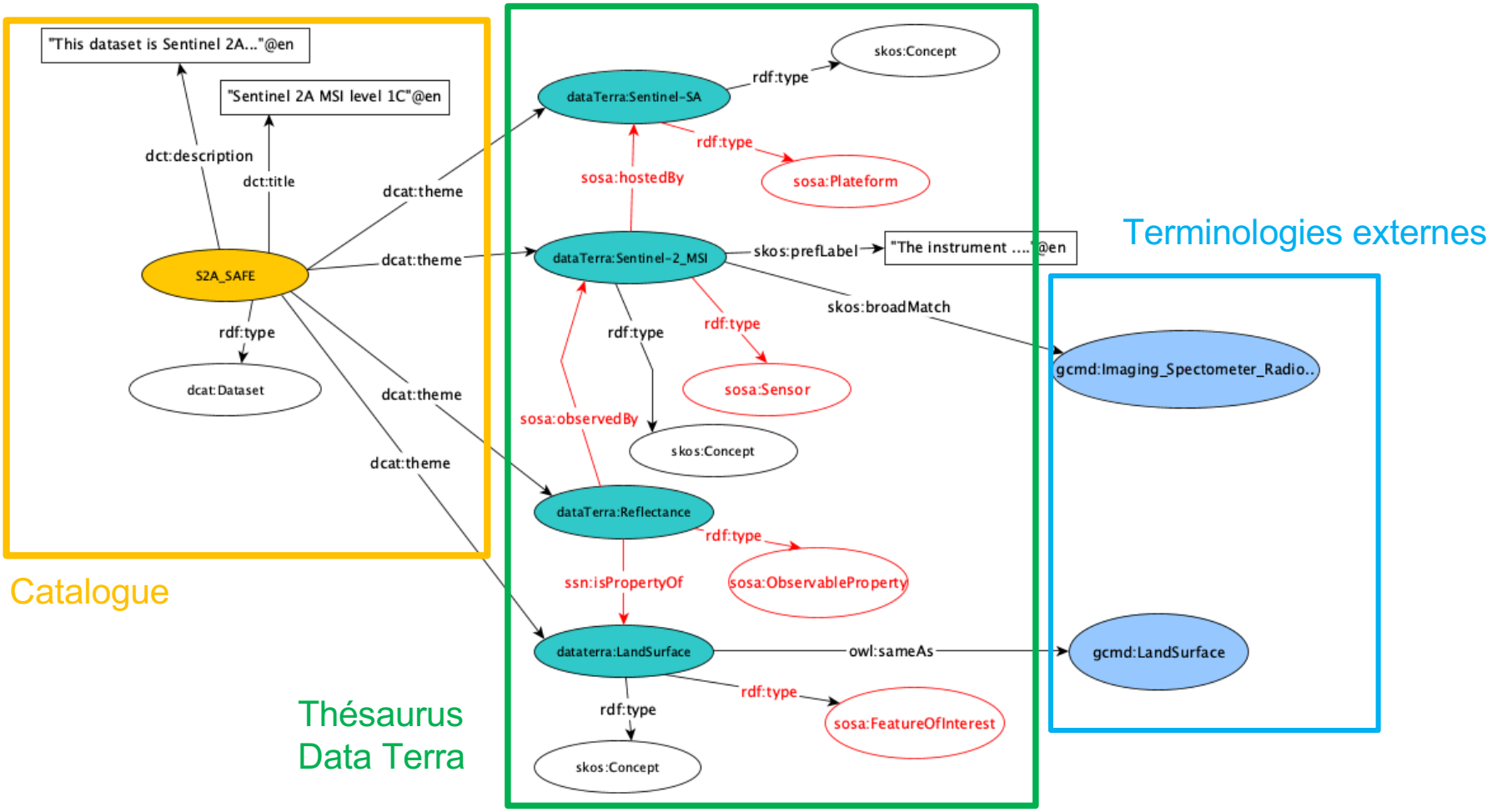
Et dans la vraie vie, comment on s’en sort ?

Garder une approche de catalogage classique ….

- Mais en y ajoutant le contexte d'observation (SOSA) dans les mots clés



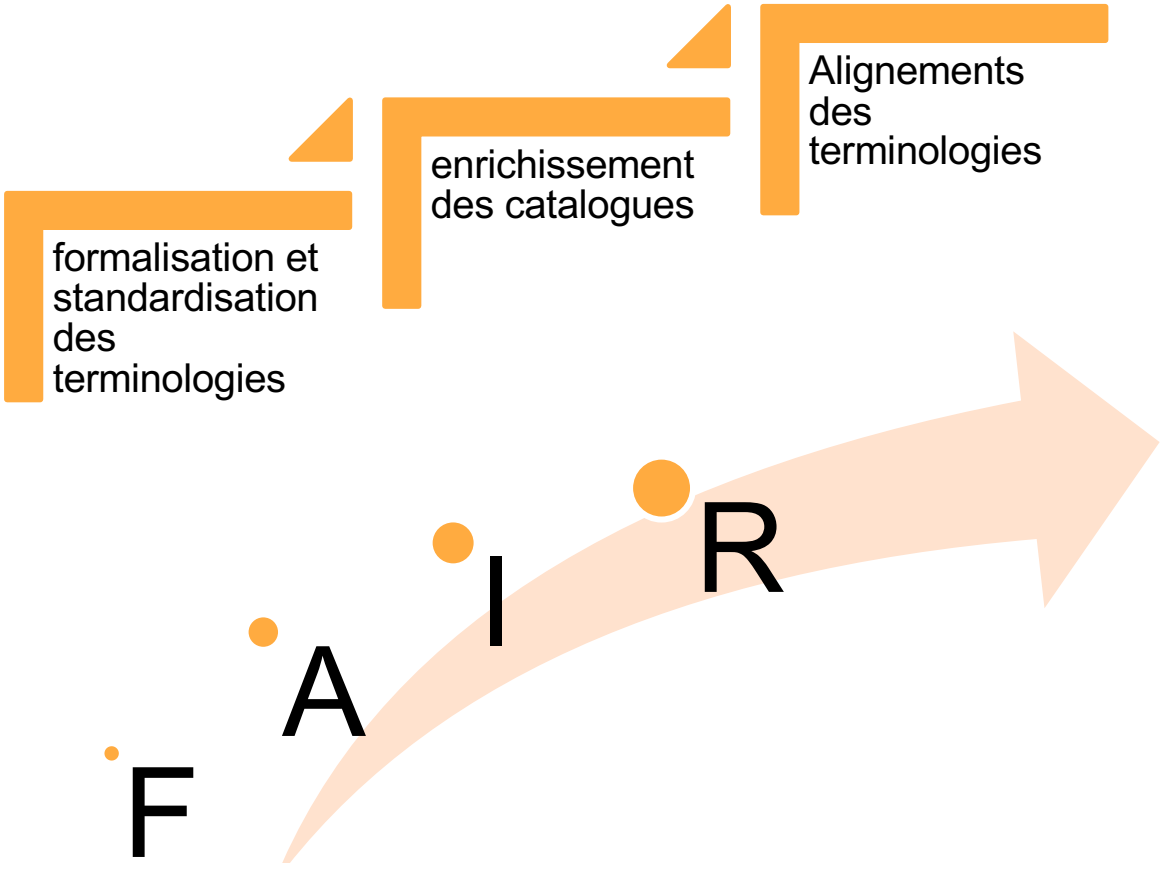
… Et typer les concepts SKOS avec les classes SOSA



Extrait du graphe de la fiche métadonnées décrivant un fichier SAFE et les concepts Skos associés

Bénéfices de cette approche

- On dissocie la standardisation (syntaxique et sémantique) des catalogues de leur enrichissement sémantique

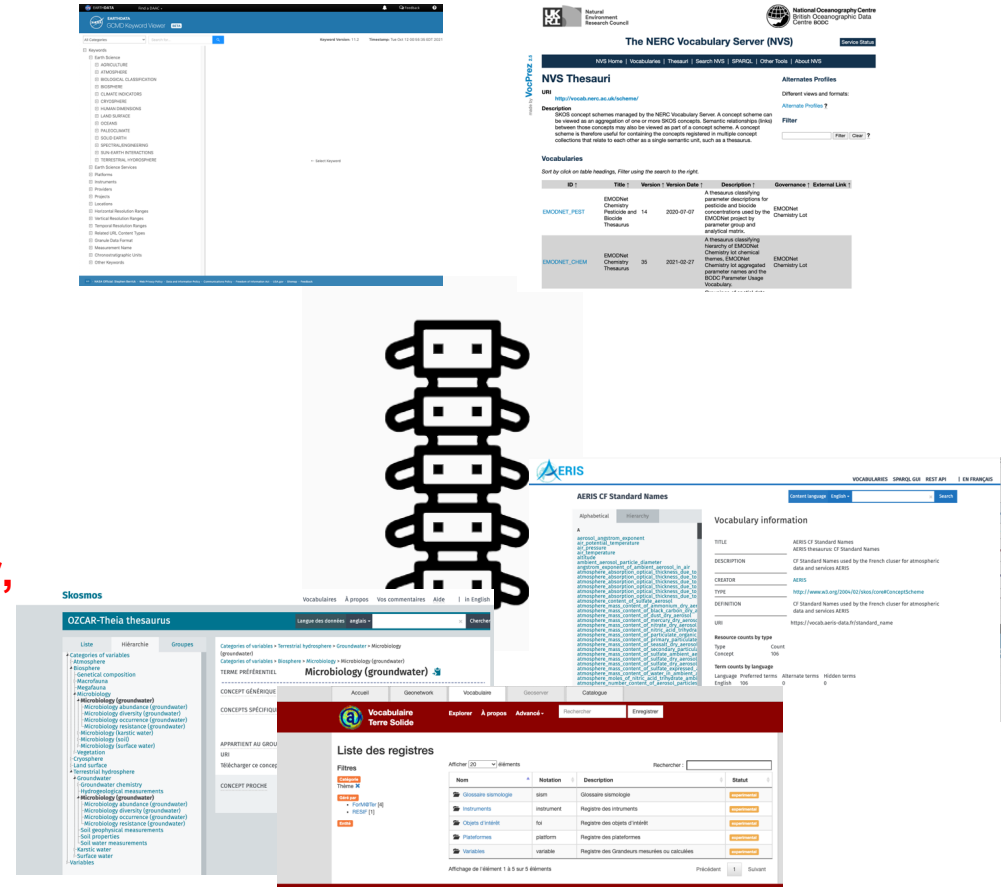


Trajectoire de construction des ressources terminologiques

1 – **Fairisation des terminologies** dans les pôles de données (compartiments) sur la base des ressources terminologiques existantes

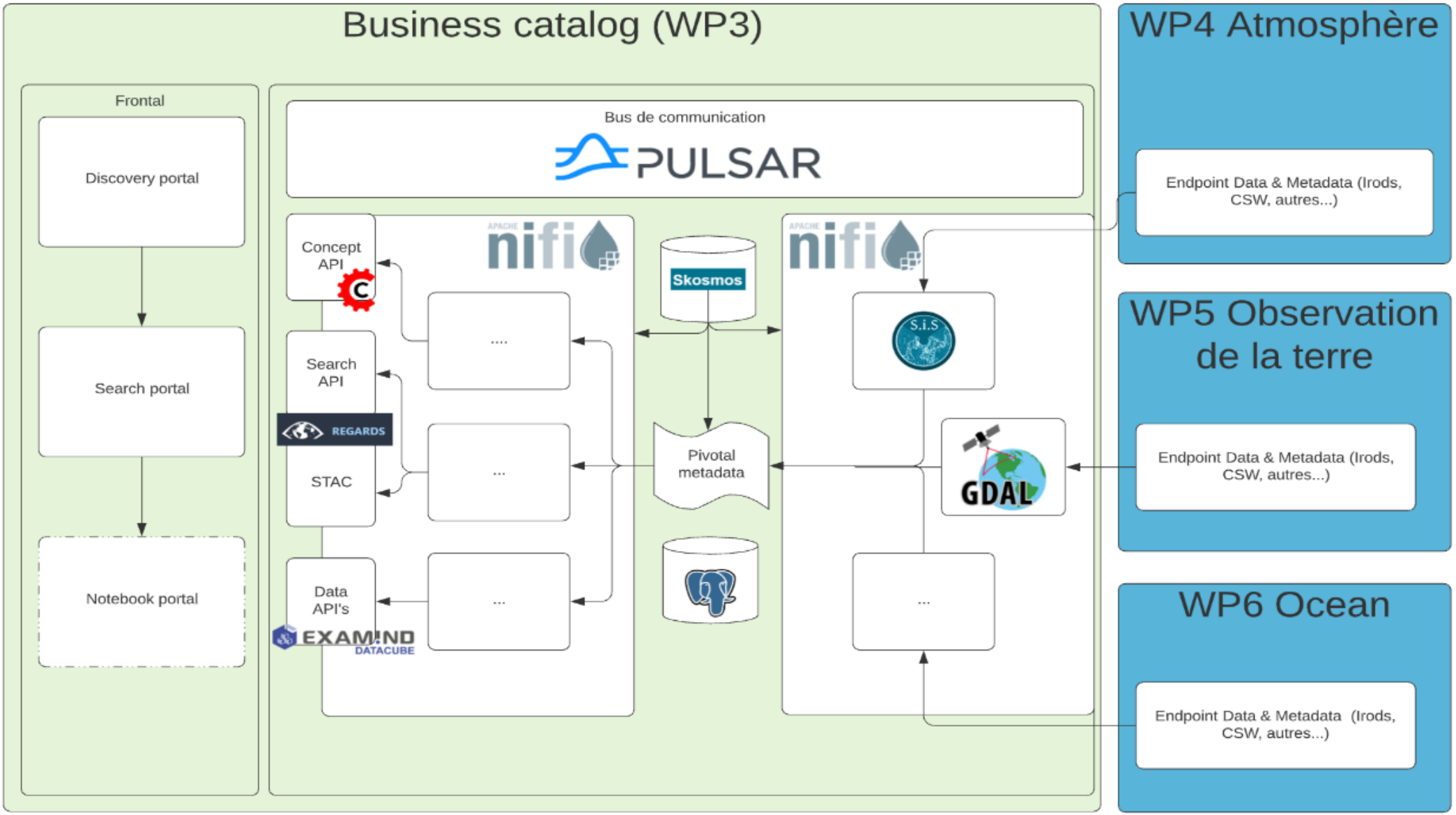
2 – Elaboration d'une « **colonne vertébrale** » **terminologique** pour le Système Terre autour de la vision d'observation : **Variable, Plateforme-capteur, Objet d'étude**

3 – Enrichissement des ressources terminologiques par des **alignements**

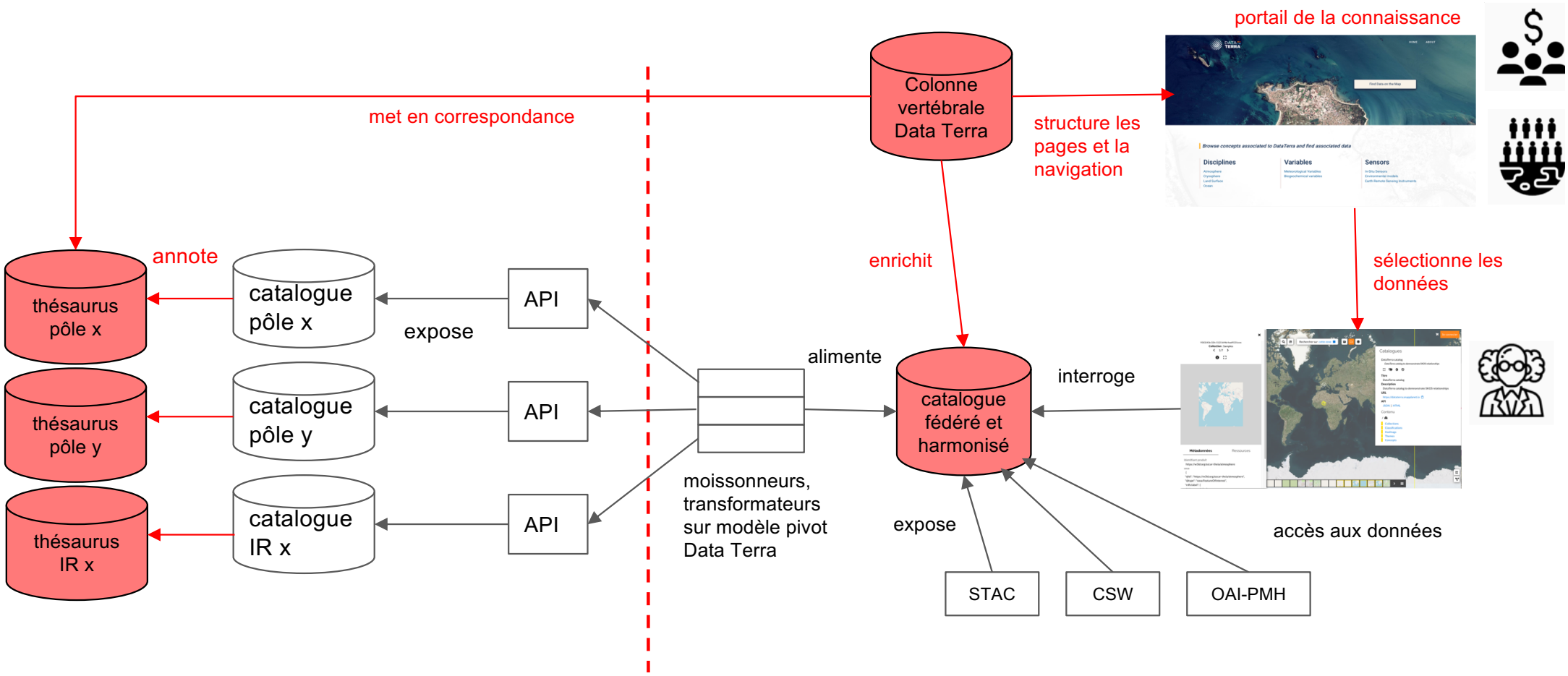


Des questions ?

Jean-Christophe Desconnets (IRD, ESPACE-DEV)
Direction Technique Data Terra
Jean-christophe.Desconnets@ird.fr


Business catalog (WP3)
Frontal
Discovery portal
Search portal
Notebook portal
Bus de communication
PULSAR
Concept API
Search API
REGARDS
STAC
Data API's
EXAMIND
DATACUBE
....
...
...
Skosmos
Pivotal metadata
S.I.S
GDAL
...
WP4 Atmosphère
Endpoint Data & Metadata (Irods, CSW, autres...)
WP5 Observation de la terre
Endpoint Data & Metadata (Irods, CSW, autres...)
WP6 Ocean
Endpoint Data & Metadata (Irods, CSW, autres...)

Architecture et composants entre pôles et Data Terra



Trois pistes pour avancer

1. Créer, structurer et enrichir les vocabulaires existants en lien avec les enjeux de découverte et d'accéder aux données Data Terra

1. Améliorer la Fairisation des vocabulaires

1. Améliorer l'utilisation des vocabulaires dans les métadonnées des catalogues de données

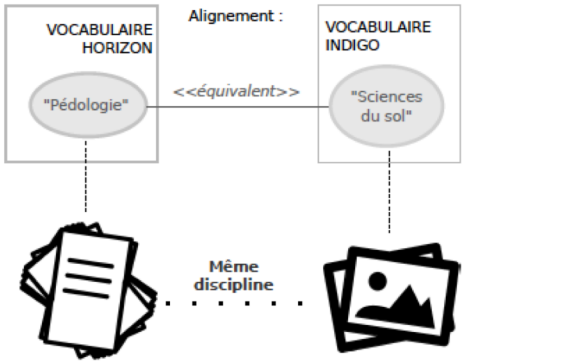
Rendre les vocabulaires standards et accessibles

Imposer un vocabulaire standard existant est difficilement envisageable

Chaque discipline a adapté son vocabulaire à ses besoins

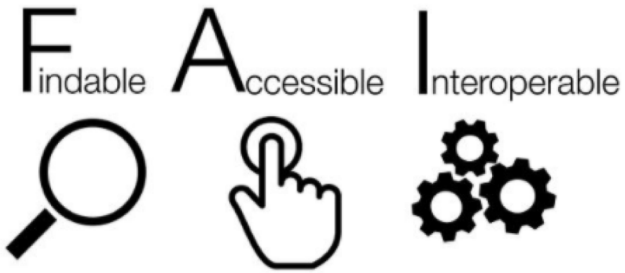
Ces vocabulaires doivent être **accessibles** dans des **formats standards** pour être **traités et interprétés automatiquement**

- ~~Solution 1 : imposer un vocabulaire commun~~
- Solution 2 : aligner les vocabulaires



Rendre les terminologies FAIR

Notamment sur



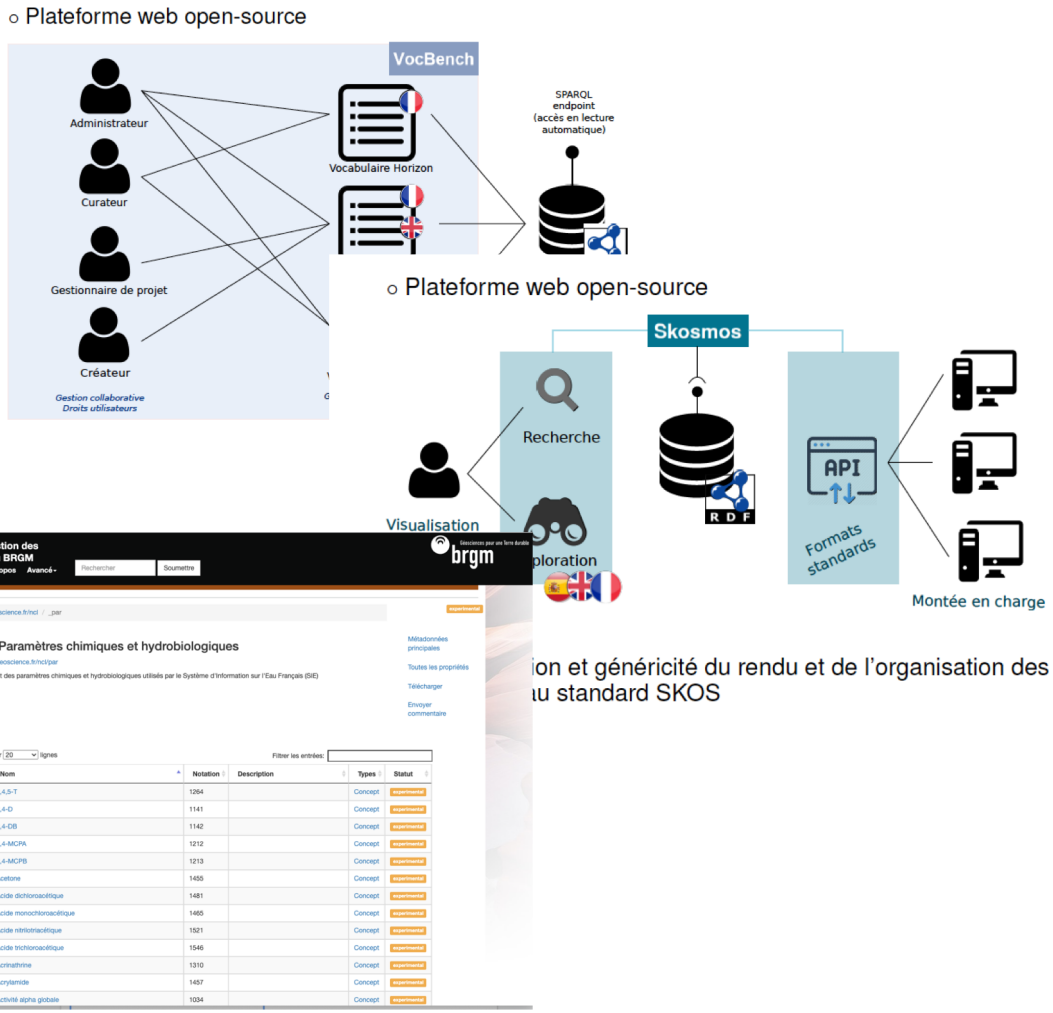
Les formaliser (volet scientifique)
Les maintenir et les faire évoluer (scientifique)
Les aligner (scientifique et technique)
Les préserver (technique)
Les partager et les exposer (technique)

Quels sont les outils qui peuvent nous aider dans ces activités ?

Quels outils pour les pôles de données ?

Des outils sur étagère

Voc Bench : Plateforme open source multilingue pour la gestion collaborative d'ontologies en OWL, de thésaurus en SKOS et plus généralement



Skosmos : Outil open source permettant de naviguer et de publier des ressources en SKOS sur le web. Il propose une API REST pour accéder aux référentiels, d'ensembles de données en RDF.

UKGovLD Registry : outil gestion et d'exposition de registres de données liées ont pour fonctionnalités prinicipal la création, la maintenance et l'évolution des listes de codes de leurs identifiants (URI).