L'interopérabilité sémantique dans l’infrastructure de recherche Data Terra et les pôles de données

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

Préambule

Travaux réalisés par des différents groupes et projets

- GT inter-pôles : Catalogue, vocabulaires
- Projet européen PHIDIAS
- Post doctorant de la Mission pour la Science Ouverte de l’IRD
- Prestation Geomatys et SnapPlanet (CNES, Data Terra)

Plan

- Enjeux de l’interopérabilité sémantique dans Data Terra
- Illustration de l’approche utilisée
- L’existant dans les pôles
- Premières recommandations et pistes de travail
- Orientations et questions (encore) ouvertes

Intéropérabilité sémantique

Associer une signification aux données, les positionner dans un domaine de connaissance

inclut le développement de vocabulaires et de schémas pour décrire les données et les liens entre les données

décrire les données avec des métadonnées
les annoter avec des vocabulaires formalisés et partagés

Quels schéma de métadonnées, quels vocabulaires utiliser ?

Les pistes pour mettre en oeuvre l'interopérabilité sémantique

Les principes FAIR comme guide

1. Les (méta)données doivent utiliser un langage de représentation des connaissance formel, accessible, commun et ayant un vaste champ d'application

→ **Interopérabilité syntaxique**

1. Les (méta)données doivent utiliser des vocabulaires qui suivent les principes FAIR

1. Les (méta)données doivent inclure des références vers d'autres (méta)données

→ **Interopérabilité sémantique**

Définitions

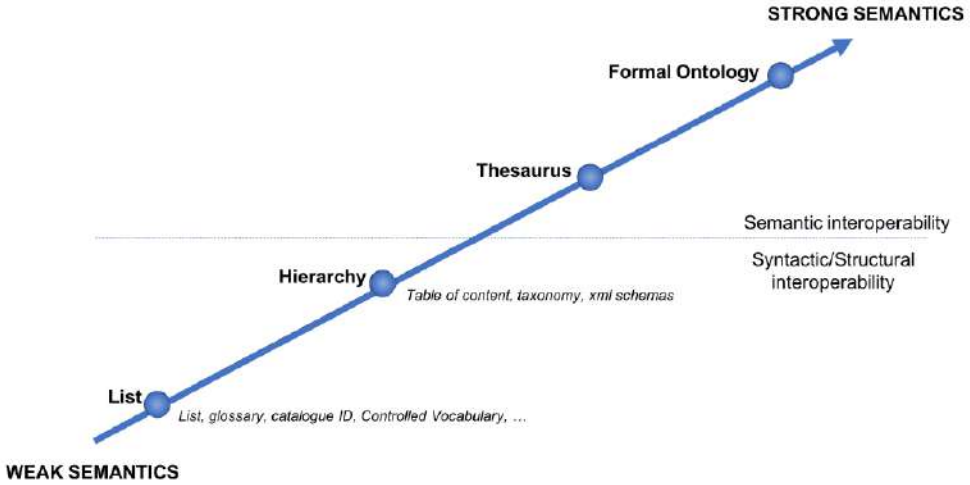


Figure 2: Semantic artefact spectrum. Derived from Leo Obrst, 2010

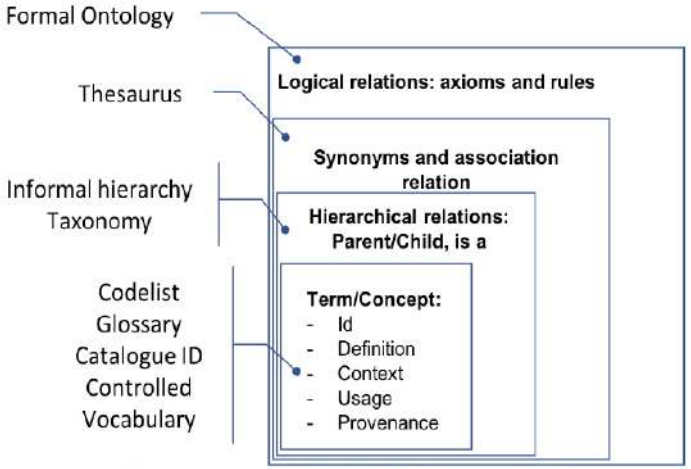
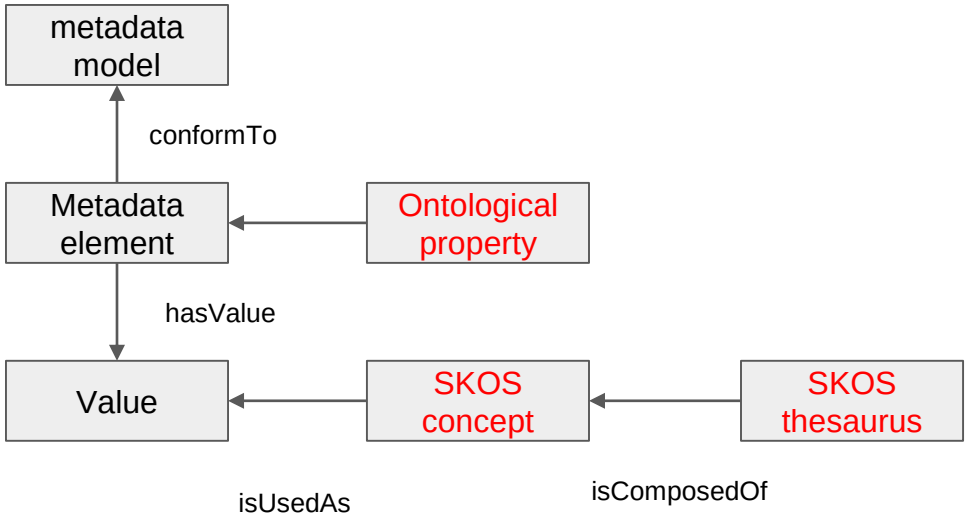


Figure 3: From list to formal ontology: a transformation path.

Définitions

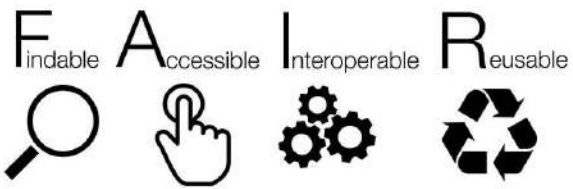


notion de modèle de métadonnées et liens avec les ontologies

Besoins Data Terra et des pôles de données

- **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**



Principes retenus pour la découverte et l’accès

- **Fédération** des **catalogues** sur la base d’un modèle de **métadonnées** commun **sémantiquement riche**
- ***2 Steps-Search***
- **Mises en correspondances** (sémantiques) entre jeux de données opérées grâce à une **standardisation** et un **alignement** des **vocabulaires** disciplinaires
- Définition d’un **contrat d’interopérabilité** avec les **pôles**

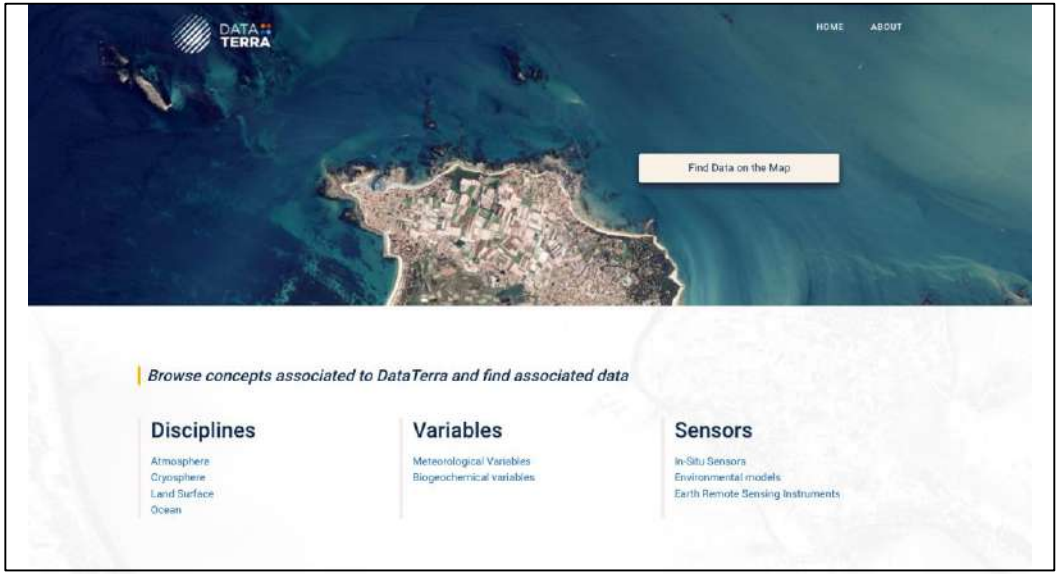
Illustration de l’approche utilisée : découvrir les données en naviguant dans les compartiments de la Terre, les capteurs et les propriétés observées

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

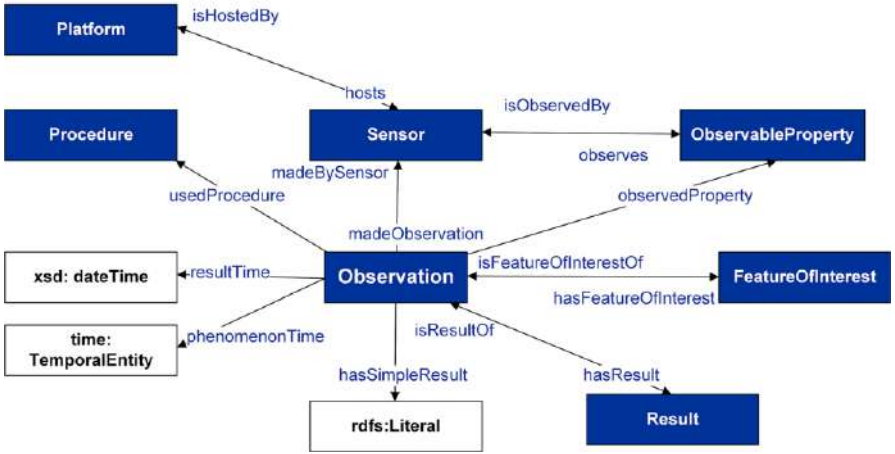
https://dataterra.geomatys.com/



Sous le capot : un modèle de métadonnées centré utilisateur

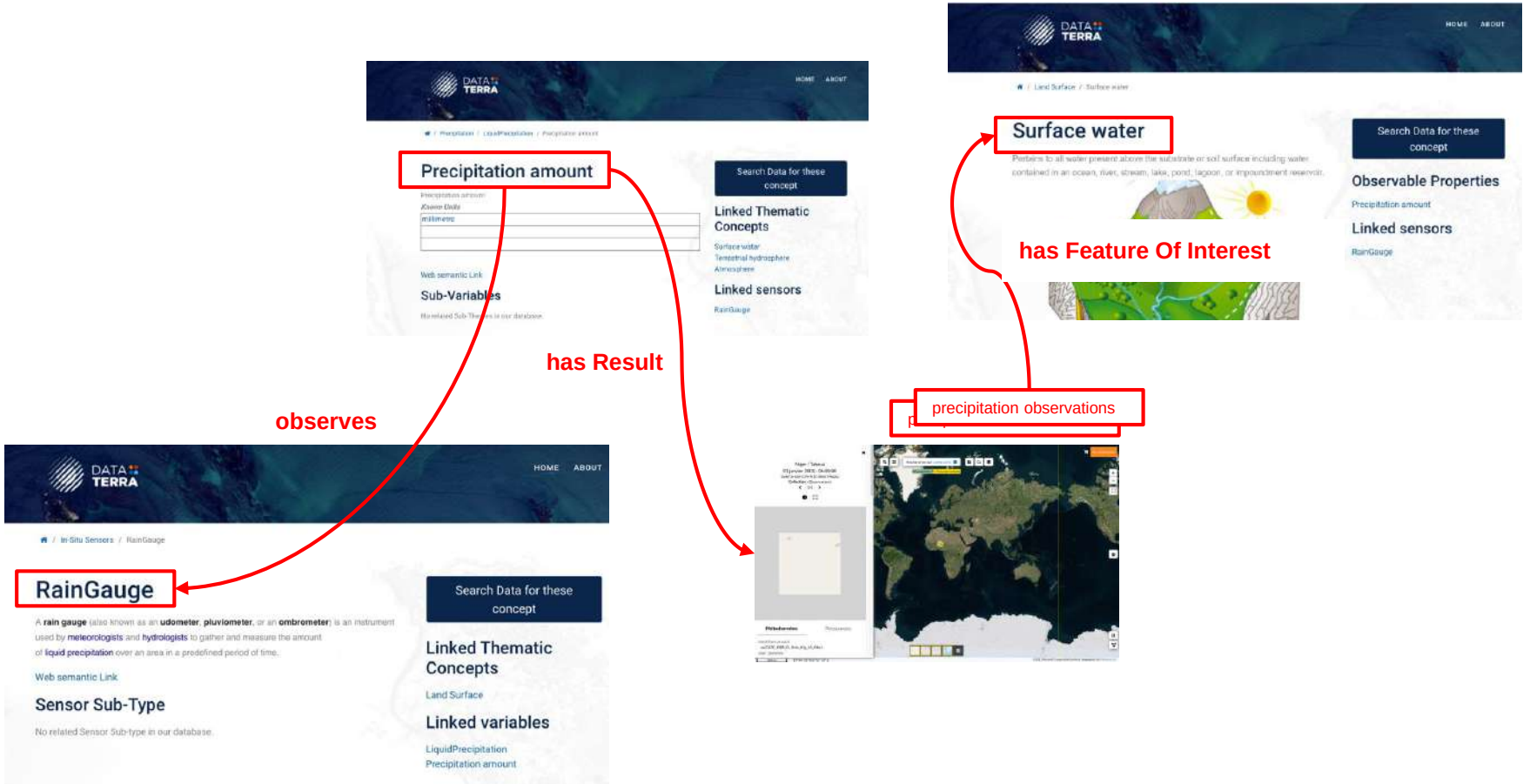
pour lier les données aux interfaces des pôles

- Annoter les données avec les concepts clés de l'ontologie SOSA
- Enrichir ces concepts avec les vocabulaires disciplinaires
- Exploiter les alignements entre termes pour naviguer sur les concepts aux interfaces des compartiments



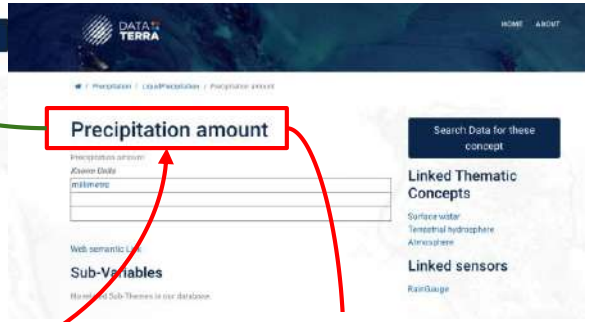
Ontologie SOSA : Sensor, Observation, Sample, and Actuator

Sous le capot : un modèle métadonnées de centré utilisateur...



… enrichi par les vocabulaires disciplinaires

skos:broader



observes

has Result

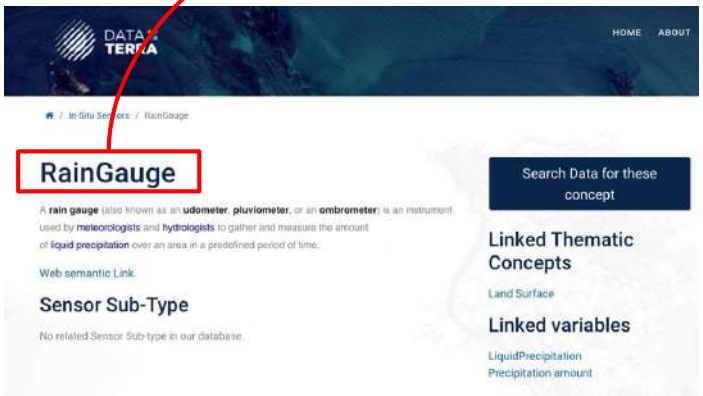
has Feature Of Interest

Land Surface

is Feature Of Interest Of

skos:narrower

precipitation observations



Etat des lieux dans les pôles

Etat des lieux des services interopérables des pôles, PNDB, CLIMERI

- Réunions organisées mi 2020 : CLIMERI, PNDB, AERIS, ICARE, ODATIS, DINAMIS, THEIA, Formater

But

1 - Avoir une vision du niveau d'interopérabilité des catalogues des pôles et IR,

- inventaire des thésaurus, leur utilisation dans les catalogues, standardisation et API d'accès

- identifier les hétérogénéités des catalogues pour identifier les travaux de transformation

2 - Collecter des métadonnées et données échantillons pour valider le modèle pivot

Inventaire et Synthèse sous le prisme des principes FAIR

→ avoir les éléments qualitatifs pour analyser, prioriser les actions de fairisation des métadonnées et des vocabulaires

Inventaire et synthèse

Sur :

- Format de données
- Modèle de métadonnées
- API d”accès aux données et métadonnées

→ **Vocabulaires contrôlés et thésaurus**

- Organisation des données

Etat des lieux des pôles sur les données, services, métadonnées et thésaurus :

https://docs.google.com/spreadsheets/d/10s_uTinEtlCxAI3s0Gf87xfhgeTs6ca594EUoCbd3z8/edit?usp=sharing

Les vocabulaires disciplinaires utilisés (1/2)

	Type	Odatis	Aeris	Theia	Ozcar	Formater	Climeri	PNDB
Vocabulaires propres aux pôles								
odatis_variables	liste							
odatis_centre_donnees	liste							
odatis_type_jeux_donnee	liste							
odatis_thematiques	liste							
ozcar-TheiaThesaurus	thésaurus							
Thesaurus de Form@ter	thésaurus							
Variable CMIP6	liste							

	Utilisation

Les vocabulaires disciplinaires utilisés (2/2)

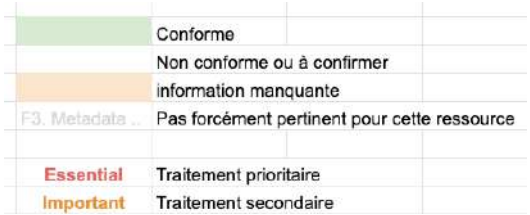
	Type	Odatis	Aeris	Theia	Ozcar	Formater	Climeri	PNDB
Vocabulaires de référence								
WORMS	thésaurus							
TaxRef	taxonomie							
SND-P01 (Voc général)	thésaurus							
SDN-P02 (Decouverte)	thésaurus							
SDN-P07 (mapping SDN/Convention_CF)	thésaurus							
EOV	liste							
Référentiels SANDRE	thésaurus							
GCMD-Instruments	thésaurus							
GCMD-Locations	thésaurus							
GCMD-Plateforms	thésaurus							
GCMD-Science Keywords	thésaurus							
INSPIRE Themes	thésaurus							
EBVs	liste							
CF_Convention (names)	liste							

(Essai d’) évaluation de la maturité FAIR des pôles

Travail préliminaire à utiliser comme une synthèse et une analyse de cet existant

Sur :

- FAIRness des vocabulaires des pôles
- FAIRness des vocabulaires des communautés
- FAIRness des métadonnées des pôles
- FAIRnessdes données des pôles



Niveau de fairisation des données, métadonnées et thésaurus des pôles - Etat d'avancement:

https://drive.google.com/file/d/1SUyEOUnl31PvfIVMw2NHSNBZ3IcQaX4P/view?usp=sharing

Maturité FAIR des vocabulaires

Il est essentiel de connaître la maturité FAIR pour évaluer notre capacité d'annoter nos métadonnées avec ces vocabulaires et produire des métadonnées FAIR

FAIR PRINCIPLES	Priority	Odatis	Aeris	Théia Spatial	Theia Ozcar	Formater	Climeri	PNDB
Findable								
F1. Terminology are assigned a globally unique and persistent	Essential							
F2. Terminology are described with rich metadata (defined by R1 below)	Essential							
F3. Terminology clearly and explicitly include the identifier of the data they		N/A	N/A	N/A	N/A	N/A	N/A	N/A
F4. Terminology are registered or indexed in a searchable resource	Essential							
Accessible								
A1. Terminology retrievable by their identifier using a standardised	Essential							
A1.1 The protocol is open, free, and universally implementable	Essential							
A1.2 The protocol allows for an authentication and authorisation		N/A	N/A	N/A	N/A	N/A	N/A	N/A
A2. Metadata are accessible, even when the data are no longer available		N/A	N/A	N/A	N/A	N/A	N/A	N/A
Interoperable								
I1. Terminology use a formal, accessible, shared, and broadly	Important							
I2. Terminology use vocabularies that follow FAIR principles		N/A	N/A	N/A	N/A	N/A	N/A	N/A
I3. Terminology include qualified references to other (meta)data	Important							

Les vocabulaires de référence à divers stades de maturité FAIR

FAIR PRINCIPLES	Priority	WORMS	TaxRef	SND-P01	SDN-P02	EOVs	SANDRE	GCMD	INSPIRE Themes	EBVs	CF
Findable											
F1. Terminology are assigned a globally unique and persistent identifier	Essential										
F2. Terminology are described with rich metadata (defined by R1 below)	Essential										
F3. Terminology clearly and explicitly include the identifier of the data they											
F4. Terminology are registered or indexed in a searchable resource	Essential										
Accessible											
A1. Terminology retrievable by their identifier using a standardised	Essential										
A1.1 The protocol is open, free, and universally implementable	Essential										
A1.2 The protocol allows for an authentication and authorisation	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
A2. Metadata are accessible, even when the data are no longer available	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Interoperable											
I1. Terminology use a formal, accessible, shared, and broadly applicable language	Important										
I2. Terminology use vocabularies that follow FAIR principles	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
I3. Terminology include qualified references to other (meta)data	Important										

Des usages plus ou moins maîtrisés des vocabulaires au sein des catalogues

FAIR PRINCIPLES	Priority	Odatis	Aeris	Théia Spatial	Theia Ozcar	Formater	Climeri	PNDB
Findable								
F1. Metadata are assigned a globally unique and persistent identifier	Essential							
F2. Metadata are described with rich metadata (defined by R1 below)	Essential							
F3. Metadata clearly and explicitly include the identifier of the data they describe	Essential							
F4. Metadata are registered or indexed in a searchable resource	Essential							
Accessible								
A1. Metadata are retrievable by their identifier using a standardised communications protocol	Essential							
A1.1 The protocol is open, free, and universally implementable	Essential							
A1.2 The protocol allows for an authentication and authorisation procedure, where necessary		N/A	N/A	N/A	N/A	N/A	N/A	N/A
A2. Metadata are accessible, even when the data are no longer available	Essential	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Interoperable								
I1. Metadata use a formal, accessible, shared, and broadly applicable language for knowledge	Important							
I2. Metadata use vocabularies that follow FAIR principles	Important							
I3. Metadata include qualified references to other Metadata	Important							

Premières recommandations et pistes de travail

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

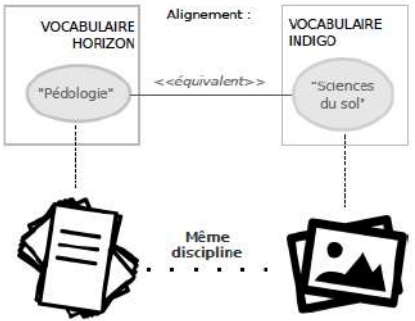
Rendre les vocabulaires standards et accessibles

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

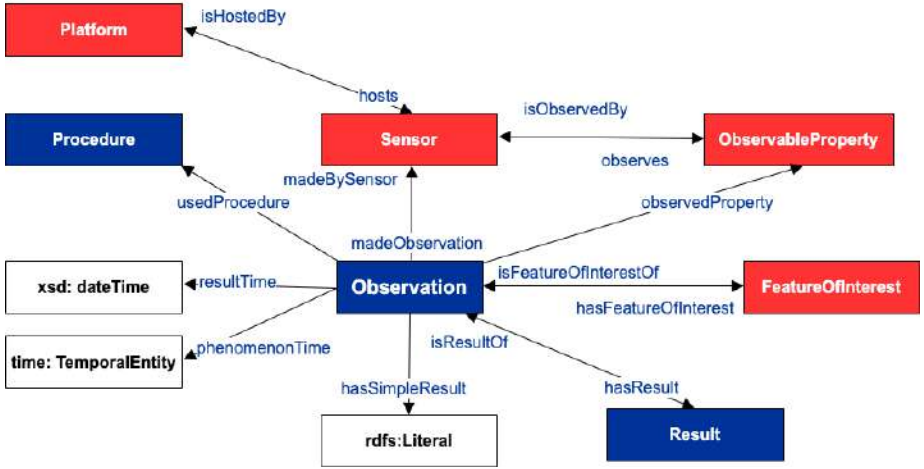


Sur les bonnes pratiques de création des terminologies

1 - Utiliser les **thésaurus** disciplinaires **existants** et de référence

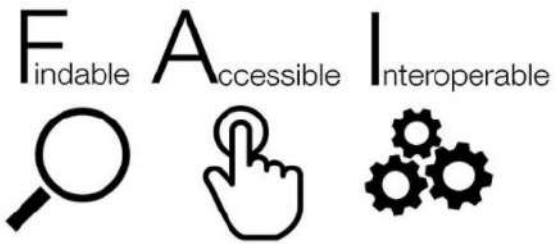
2 - Contribuer si possible à les enrichir - **éviter** le plus possible de produire des **thésaurus "locaux"**

3 - Les **enrichir** par des **alignements** et des **nouveaux concepts** en les organisant autour de la vision d'observation : **Variables, Plateforme-capteur, Objet d'étudié**



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 ?

Le besoin : plateforme assurant la gestion, la consultation et l’exposition interopérable des vocabulaires disciplinaires

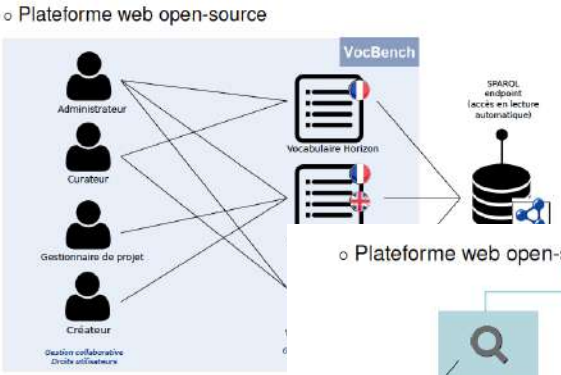
Quels outils pour les pôles de données ?

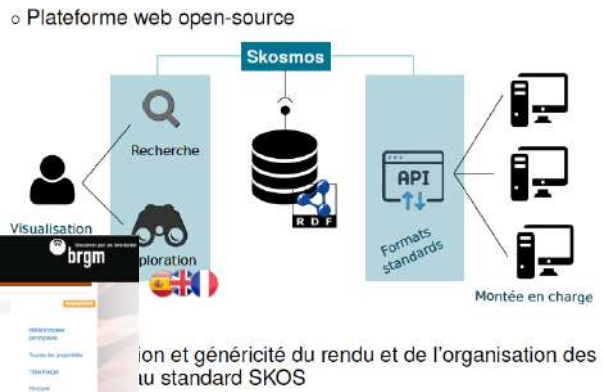
Des outils sur étagère

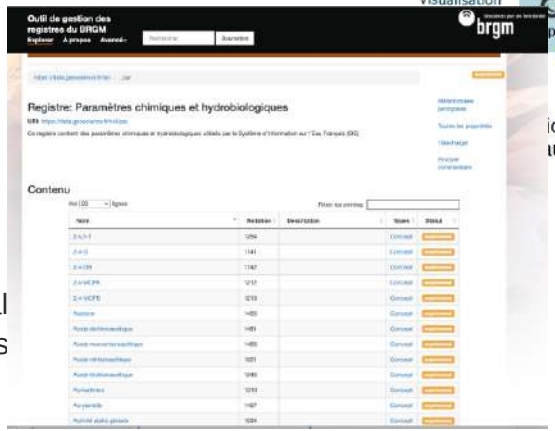
Voc Bench : Plateforme open source multilingue pour la gestion collaborative d'ontologies en OWL, de thésaurus en SKOS et plus généralement

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).



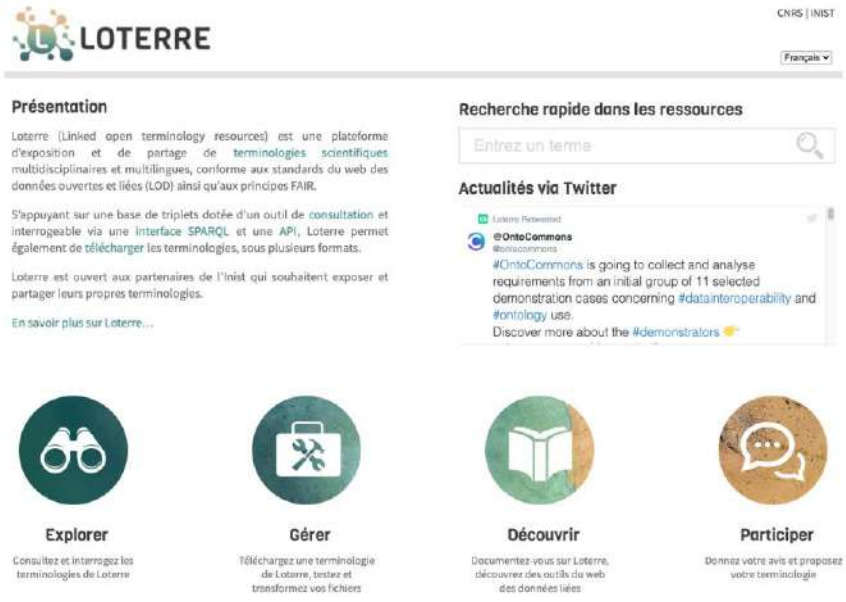


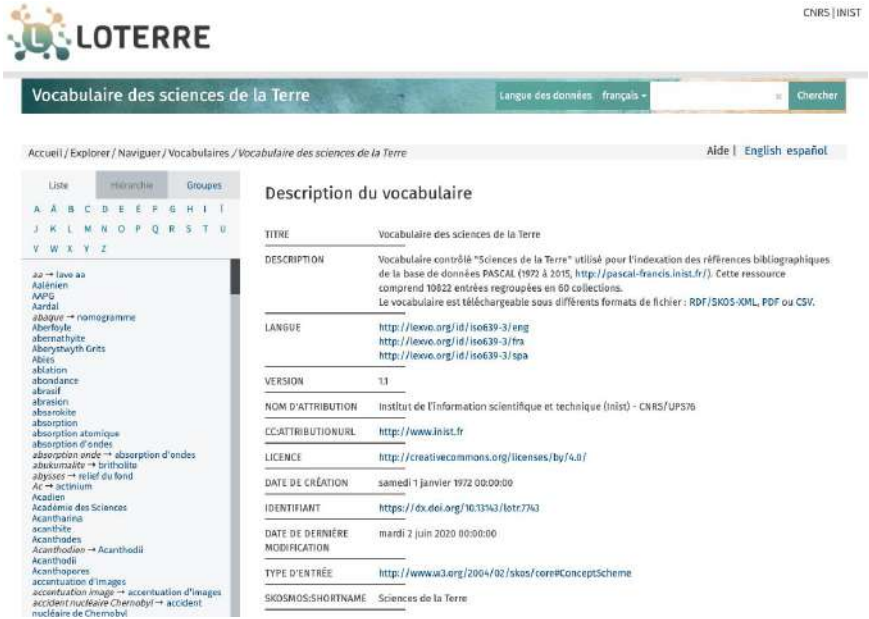


ion et généricité du rendu et de l'organisation des
u standard SKOS

Quels outils pour les pôles de données ?

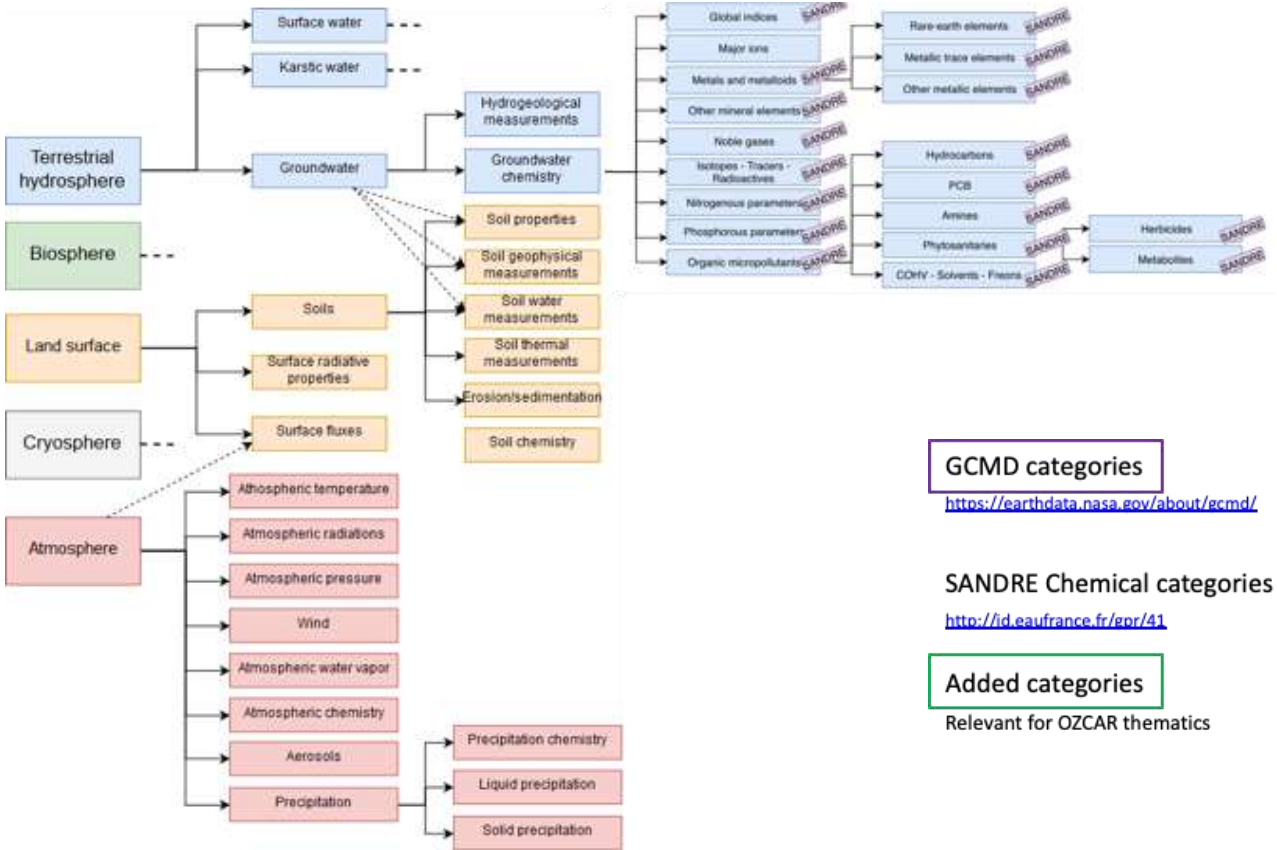
Une service terminologique pluri disciplinaire :
LOTERRE





Exemple de mise en place: IR OZCAR-THEIA (1/3)

sur l'enrichissement de thésaurus existants pour répondre aux besoins d'annotation d'une communauté



GCMD categories

https://earthdata.nasa.gov/about/gcmd/

SANDRE Chemical categories

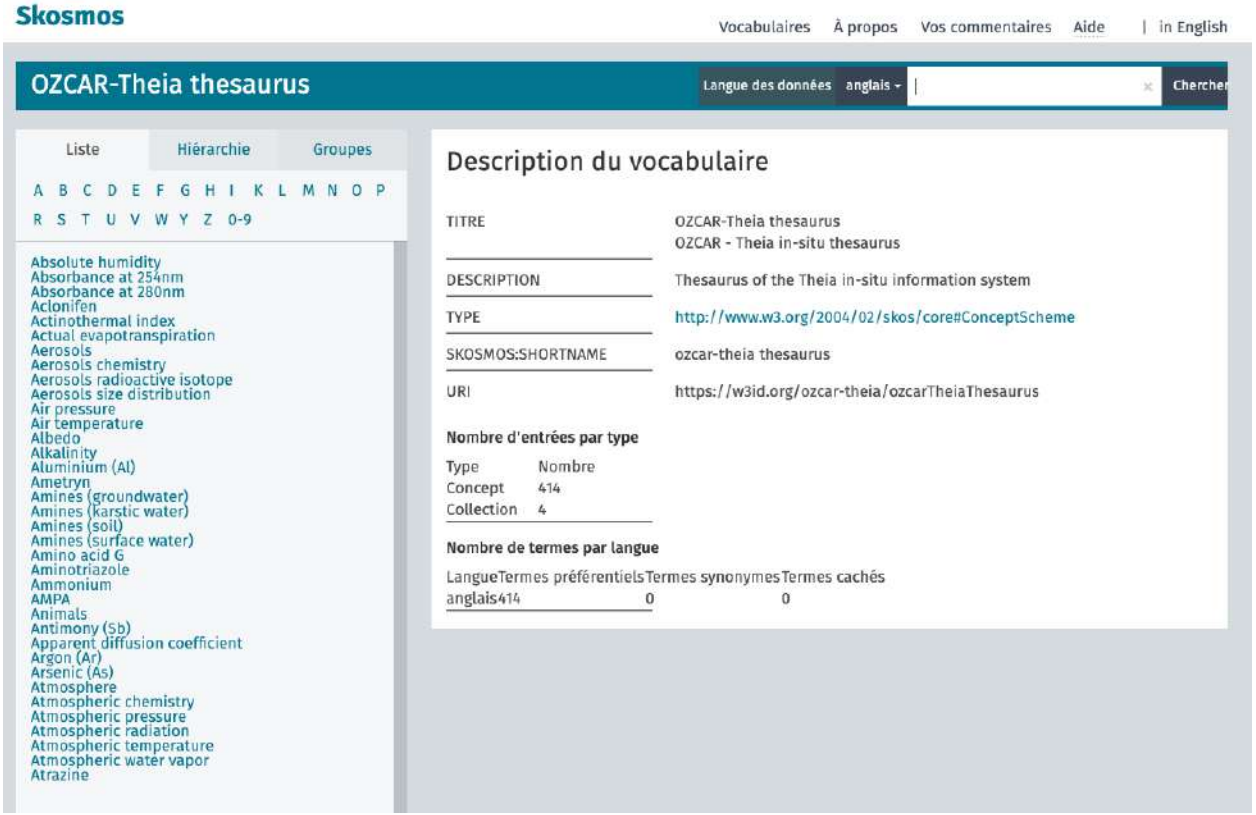
http://id.eaufrance.fr/gpr/41

Added categories

Relevant for OZCAR thematics

Exemple de mise en place: IR OZCAR-THEIA (2/3)

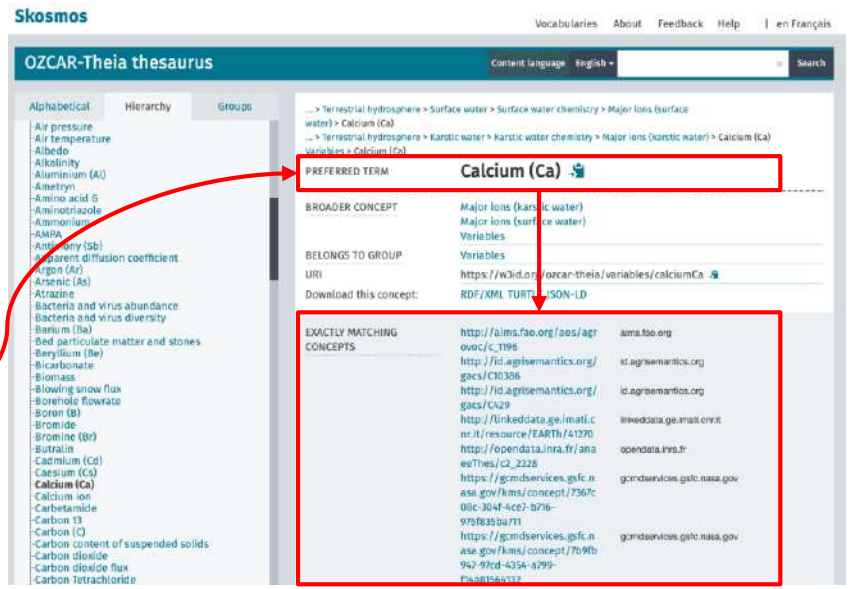
Sur la fairisation des vocabulaires : le service de thésaurus OZCAR-THEIA



Exemple de mise en place: IR OZCAR-THEIA (1/3)

Sur les bonnes pratiques de catalogage : Il ne suffit pas d'avoir des vocabulaires FAIR, pour que les métadonnées le soient…


```
  ],
  "gcmdKeywords": [
    {
      "category": "EARTH SCIENCE",
      "topic": "TERRESTRIAL HYDROSPHERE",
      "term": "SURFACE WATER",
      "variableLevel1": "SURFACE WATER CHEMISTRY",
      "variableLevel2": null,
      "variableLevel3": null,
      "uuid": null
    },
    {
      "category": "EARTH SCIENCE",
      "topic": "TERRESTRIAL HYDROSPHERE",
      "term": "WATER QUALITY/WATER CHEMISTRY",
      "variableLevel1": null,
      "variableLevel2": null,
      "variableLevel3": null,
      "uuid": null
    }
  ],
  "theiaCategories": [
    "https://w3id.org/ozcar-theia/surfaceWaterMajorIons"
  ],
  "theiaVariable": {
    "uri": "https://w3id.org/ozcar-theia/variables/calciumCa",
    "prefLabel": [
      {
        "lang": "en",
        "text": "Calcium (Ca)"
      }
```

extrait catalogue OZCAR-THEIA : curl -X GET https://in-situ.theia-land.fr/apiobservation/observation/initFacets

https://in-situ.theia-land.fr/skosmos/theia_ozcar_thesaurus/fr/