Le Système d'Information Theia/OZCAR :

Un portail d'accès à l'ensemble des données in-situ des surfaces continentales

Equipe projet

Coordination scientifique:

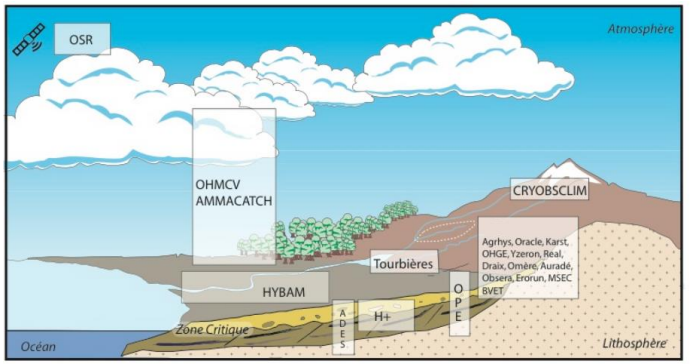
- Sylvie Galle (IRD/IGE,CMI INSU-SIC données in situ, membre BE Theia)
- Isabelle Braud (Irstea Lyon, co-animatrice de l'IR OZCAR)

Equipe Développement:

- Véronique Chaffard (IRD/IGE, responsable technique)
- Charly Coussot (OSUG, ingénieur développement, CDD depuis le 1/10/2017)

Le réseau d'observatoires de la Zone Critique

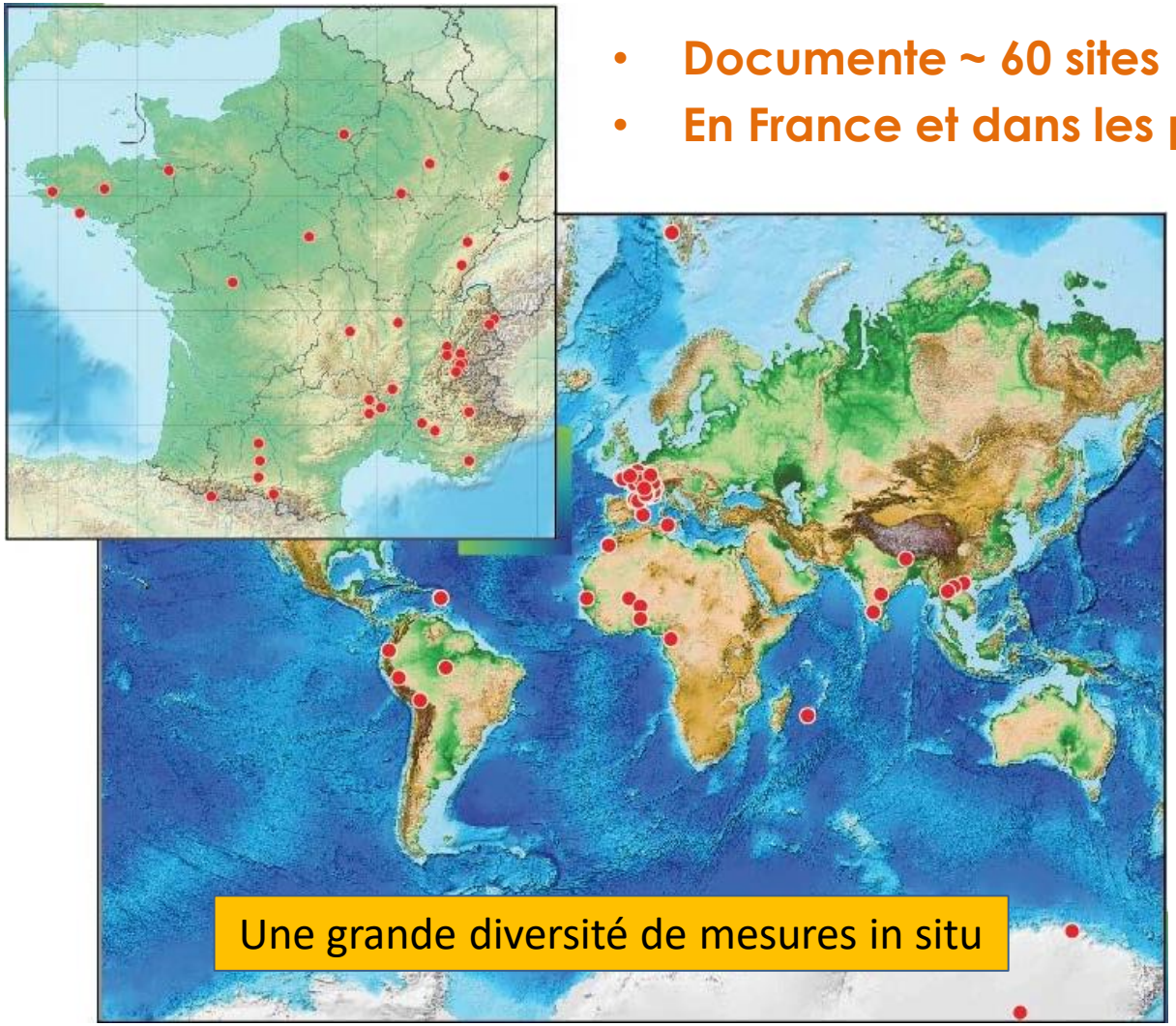
- **Infrastructure de Recherche (IR) française sur la zone critique** créée en 2016.
- OZCAR met en réseau les observatoires de long terme de la zone critique depuis la haute montagne jusqu'à la mer.
- **21 "observatoires élémentaires" labélisés** collectent **>300 variables différentes** (physiques et géochimiques) sur ~**100 sites**



- Chaque observatoire a une histoire différente et a son propre système de de gestion et diffusion des données

- OZCAR fait partie de **l'IR européenne European Long Term Ecological Research** (eLTER-RI 2018)

OZCAR : un réseau de 21 observatoires



- **Documente ~ 60 sites**
- **En France et dans les pays du Sud**

Observations

- Transport de l'eau, des sédiments et des solutés
- Bilans d'énergie de surface
- Exploration géophysique
- Occupation des sols

Objets d'intérêt

Bassins versants, aquifères, rivières, glaciers, permafrost…

De quelles données parlons-nous ?

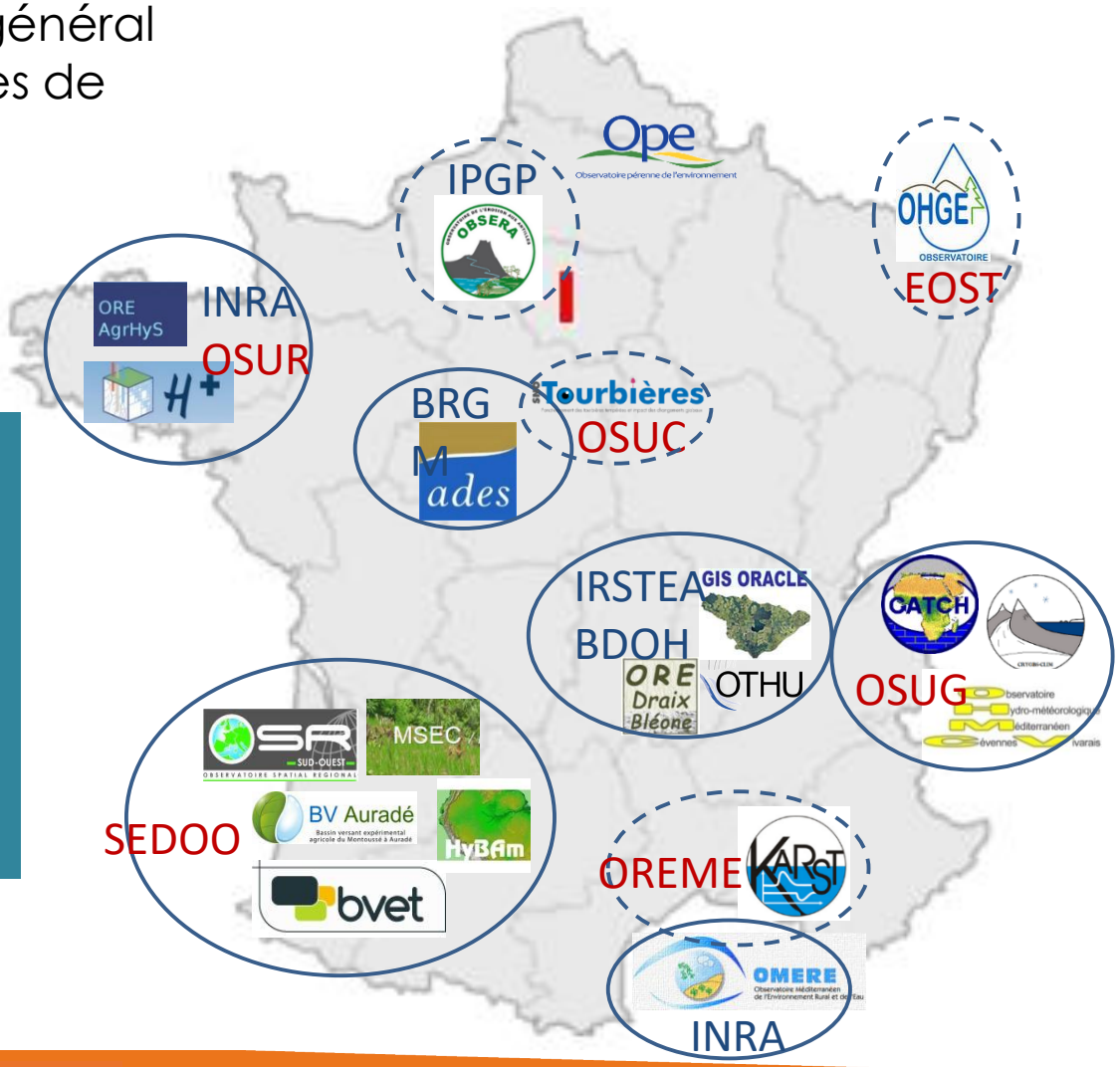
Plus de 300 variables mesurées

- Séries temporelles **ponctuelles** : météorologie, hydrologie, hydrogéologie, glaciologie bilan d'énergie, flux de sédiments, suivi d'éléments chimiques, de contaminants
- **Carottes** de sol
- Coupe géophysique **2D**
- **Cartes** de caractérisation des milieux : occupation des sols, MNT, caractéristiques géophysiques des sols

Une grande diversité de variables

Le panorama des SI des observatoires d'OZCAR

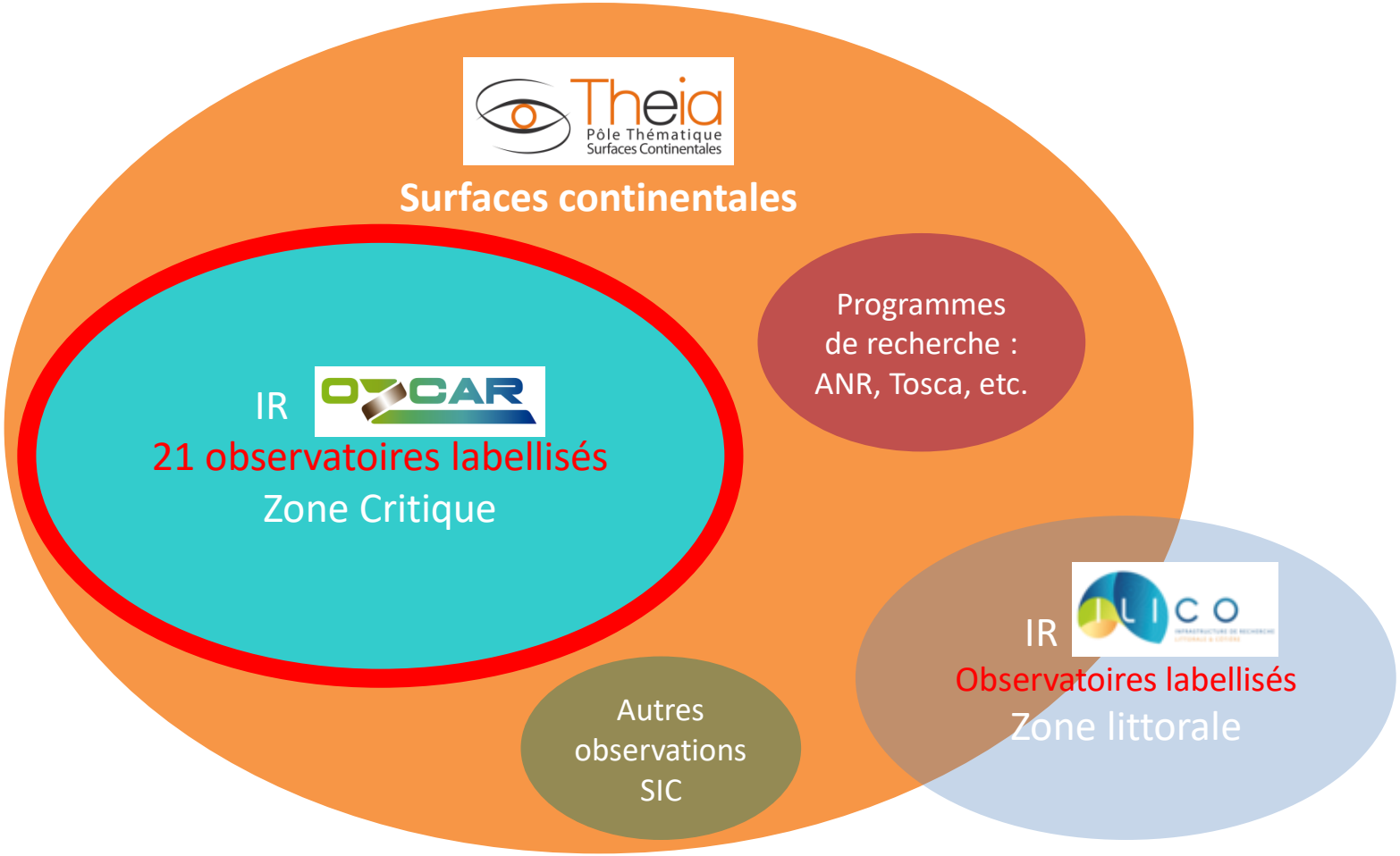
Des observatoires en général rattachés à des centres de données
(OSU ou institutionnels)



Une grande hétérogénéité dans la structuration des données (granularité), dans la maturité des SI, dans les noms de variables

OZCAR / Theia : quelle différence?

OZCAR et Theia n'ont pas le même périmètre



Objectifs du SI Theia/OZCAR

- **Un portail unique** des données d'observation in-situ des surfaces continentales qui permettra un accès **transparent** aux utilisateurs
- Un portail qui ne stocke pas les données qui sont déjà gérées par ailleurs mais qui met en place des flux d'informations qui les collecte
- Un système d'information qui respecte les **critères d'accessibilité et d'interopérabilité** de la Directive Inspire et de standards internationaux
- **Un système interopérable** avec les systèmes d'information français (pôle de données Theia, métacatalogue IR Data Terra) et européen en cours de construction (e-LTER european Long Term Ecological Research)

- Encourager la déclaration de **DOI de données**

Objectifs du SI Theia/OZCAR

Favoriser la **découverte** et **l'exploration** des données, leur **partage** et **réutilisation**, leur **citation**

http://doranum.fr/thematique-identifiants-perennes-pid/

Stratégie de construction:

- Dialogue avec les scientifiques et les informaticiens des observatoires
- Définition d'un **modèle de métadonnées commun (modèle pivot)** pour mettre en place en entrée du SI Theia/OZCAR des **flux de données** avec le SI des producteurs (éléments de métadonnées standards)
- Mettre en place en sortie du SI des **standards d'échange** (web service)
- Utilisation d'un **vocabulaire contrôlé commun pour les noms de variable** (à partir de référentiel existant)

Mise en oeuvre

1. Construction d'un **thésaurus** pour les noms et catégories de variable

2. Travail sur les **métadonnées**: analyse des **standards** et identification des **flux d'informations** à mettre en place

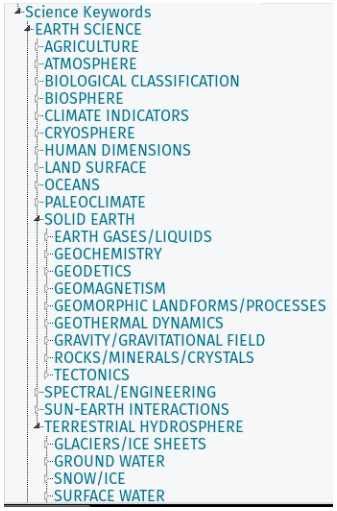
3. Définition d'un **format pivot (data model)** pour l'échange de données

4. Architecture du système et développement d'un prototype de portail

Construction d'un thésaurus pour les noms et catégories de variables

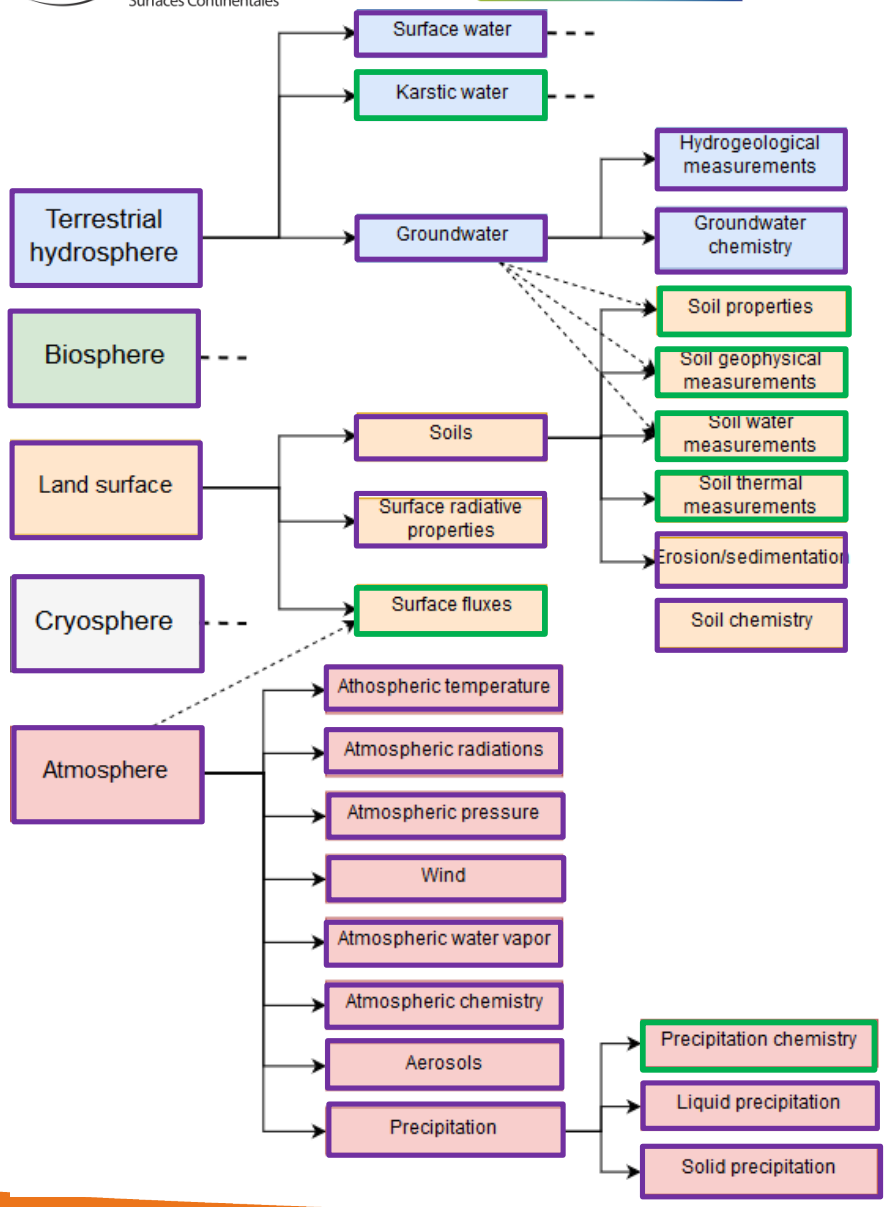
Objectif:
Permettre de rechercher la donnée par un filtrage sur les variables.

Les étapes de construction: démarche bottom-up



1. **Collecte des noms variables producteurs**
Catégorisation/classification de l'ensemble **des noms de variables** vis-à-vis du thésaurus de la **NASA GCMD Science keywords**.
Structure l'information à l'aide de concepts **hiérarchisés**

2. Création d'un **vocabulaire de noms de variables Theia/OZCAR** en se basant sur le thésaurus GCMD et mapping avec nom producteur avec les scientifiques



3. **Hiérarchisation** par **catégories de variables**

La taxonomie GCMD pas toujours adaptée aux besoins Theia/OZCAR (pas assez précise)

-> **nécessité d'ajouter des catégories propres à Theia/OZCAR**

https://wiki.earthdata.nasa.gov/display/CMR/GCMD+Keyword+Access

GCMD categories

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

Added categories

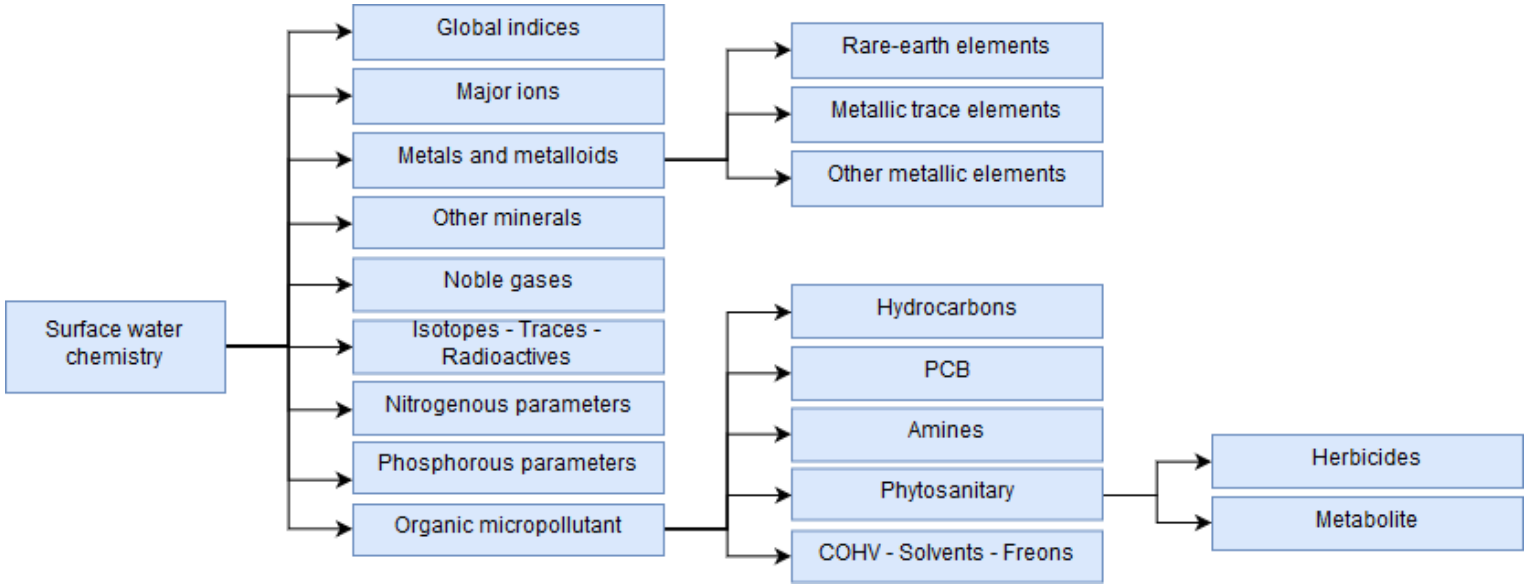
Relevant for OZCAR thematics

3. **Hiérarchisation** par **catégories de variables**

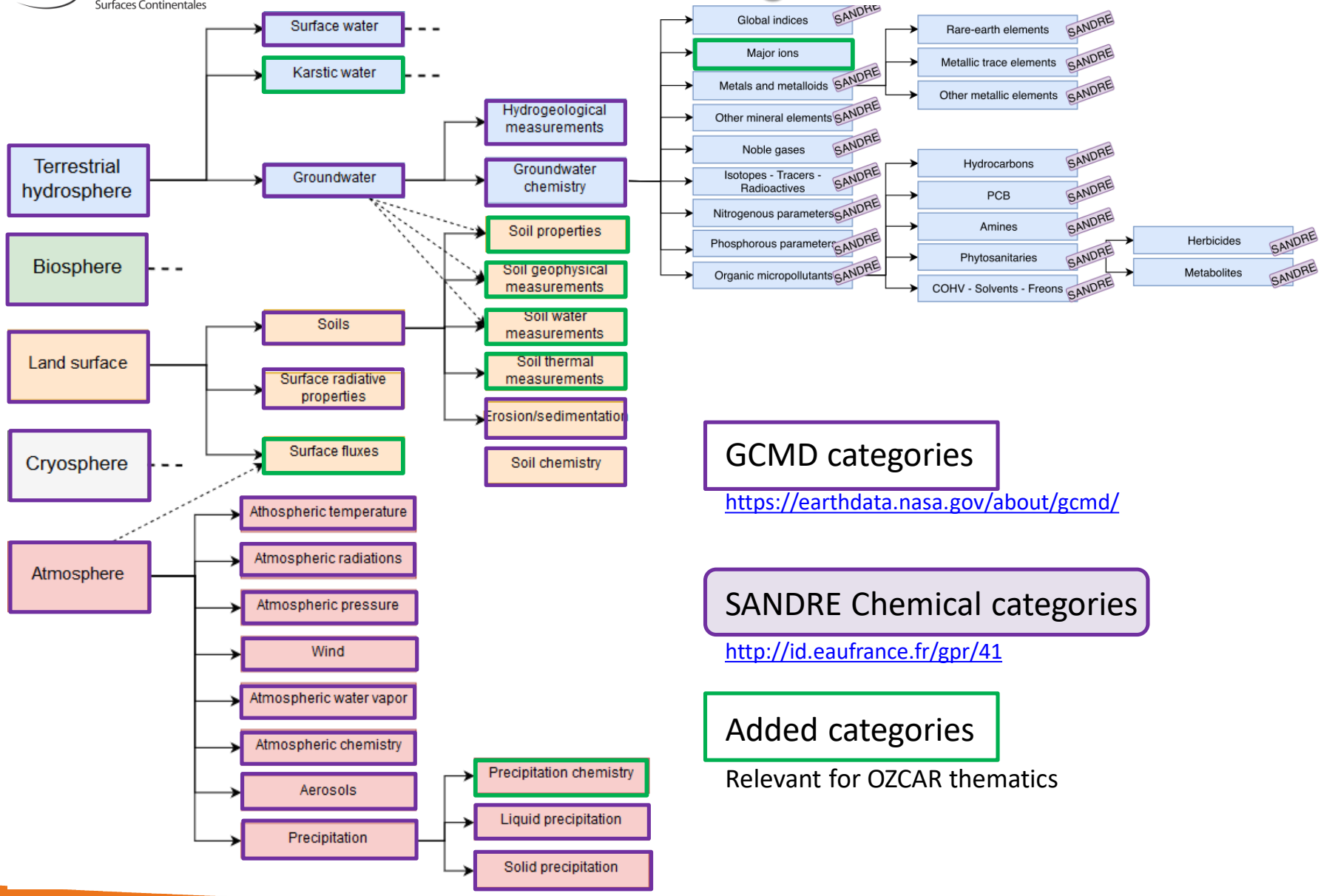
La taxonomie **GCMD** est très **inadaptée** pour **catégoriser** les **variables issues de mesures chimiques**.

Le thesaurus du **SANDRE** est repris pour organiser les variables issues de mesures chimiques.

SANDRE: Service National des Données et Référentiels sur l'Eau



Thésaurus final : catégories de variables



GCMD categories

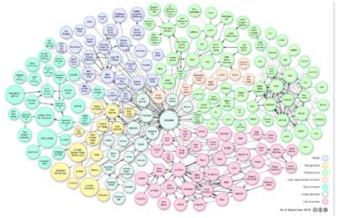
https://earthdata.nasa.gov/about/gcmd/

SANDRE Chemical categories

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

Added categories

Relevant for OZCAR thematics



4. **Publication du thésaurus en Linked Open Data**

- **Alignement des concepts avec** des **thésaurus internationaux du domaine:** AGROVOC (FAO), GEMET (Agence européenne pour l'environnement, EnvThes (LTER), ANAEE,….).
  Outil: OnaGUI

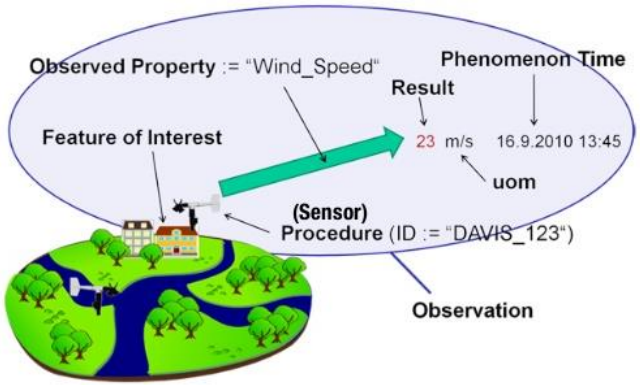
- Formalisation en **SKOS:**
  - définition de prefLabel, relations sémantiques: skos:Broader, propriétés d'alignement skos:ExactMatch, skos:closeMatch, skos:related, collection (catégories variables/noms de variables)
  - Outil: Excell + Skos Play convertisseur RDF/SKOS

- Publication sur le Web: outil Skosmos (s'appuie sur Apache Jena Fuseki SPARQL server)

http://in-situ.theia-land.fr/vocabularies/Skosmos/theia_ozcar_thesaurus/en

Travail sur les métadonnées: analyse des standards et identification des flux d'informations à mettre en place

Métadonnées nécessaires pour mettre en place des **services d'interopérabilité** et des **services** de **citation** ?



Standards analysés:

- Dublin Core
- ISO 19115/Inspire
- DataCite
- O&M et SensorML

- Data Catalog Vocabulary DCAT (schéma RDF)
- Schema.org/dataset (schéma créé Google, Yahoo, Bing, nécessaire pour être référencé dans Google Dataset Search)

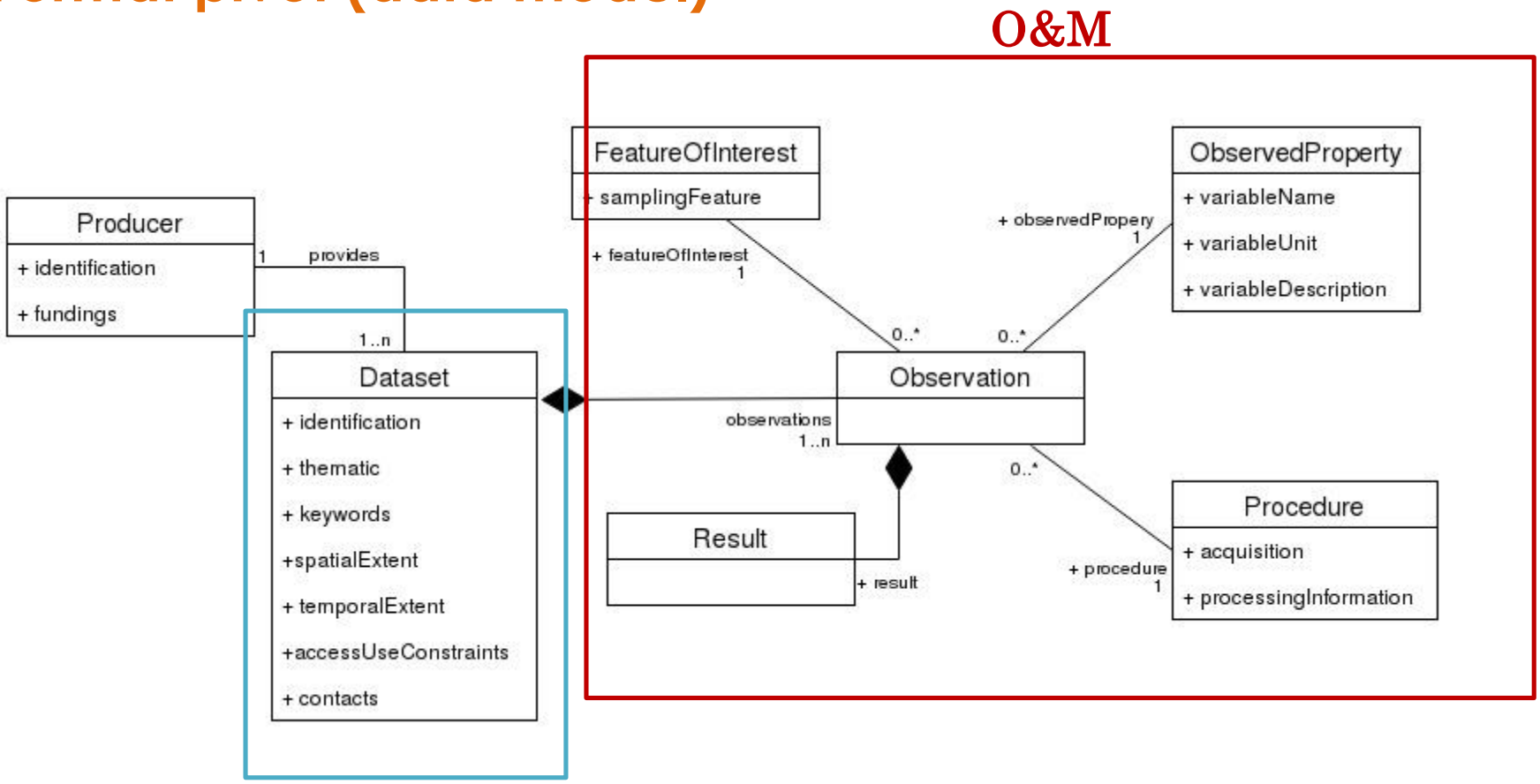
Travail sur les métadonnées: analyse des standards et identification des flux d'informations à mettre en place

- Mapping entre les **différentes normes** (pour générer les flux standards en sortie du SI (webservice OGC CSW &SOS, déclaration de doi)

- Identification d'un **set minimum** de **métadonnées** à échanger

- Choix d'un **format pivot maison** et non le format d'un standard en particulier :
  - Pour contenir toutes les informations nécessaires pour répondre à la fois aux différents standards d'intéropérabilité et de citation et aux fonctionnalités du système
  - Pour offrir une liberté évolutive
  - Pour offrir un format simple (moins verbeux que XML)

Format pivot (data model)



ISO 19115 / Inspire

Modèle conceptuel

Format pivot (data model) 1/2

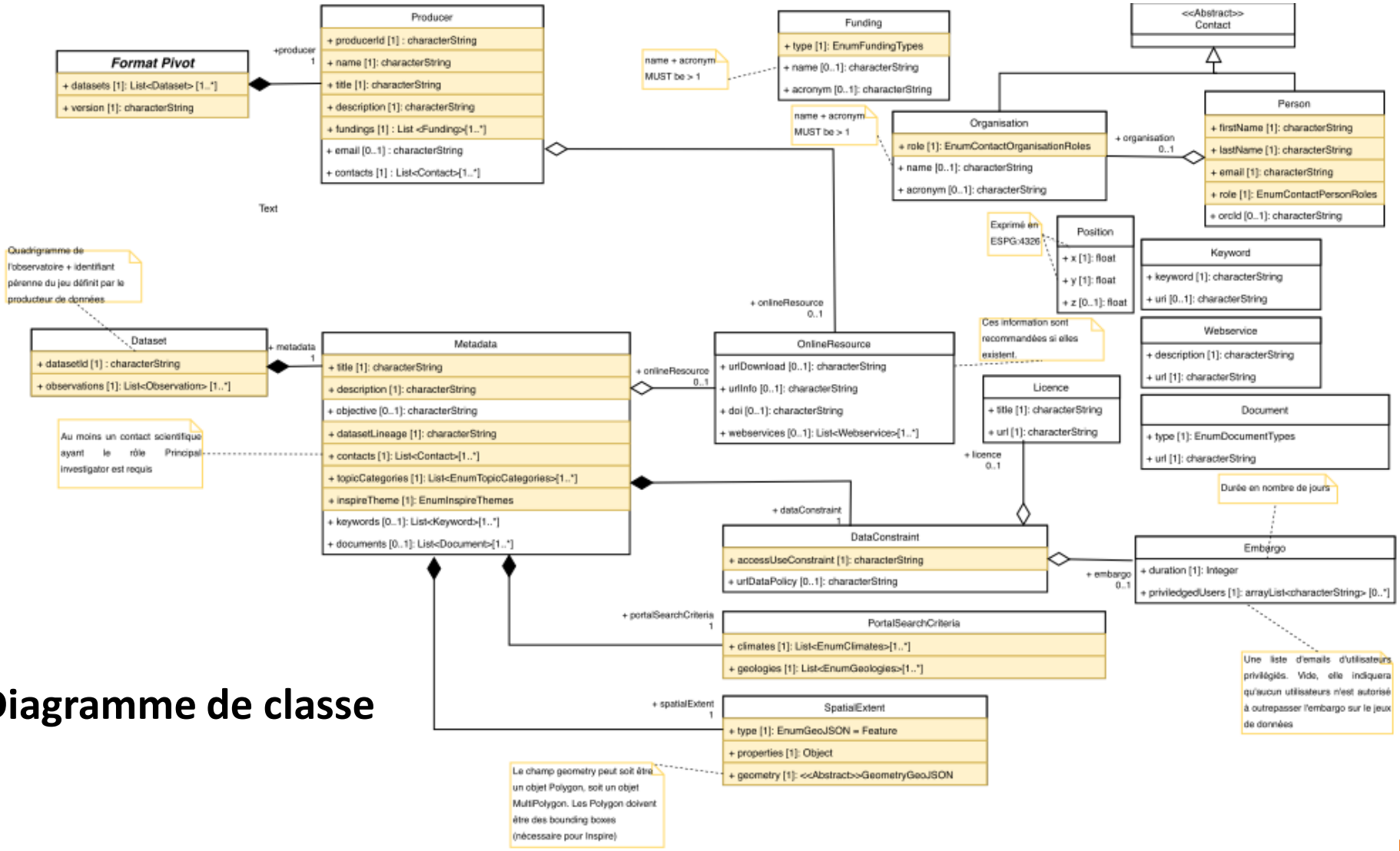
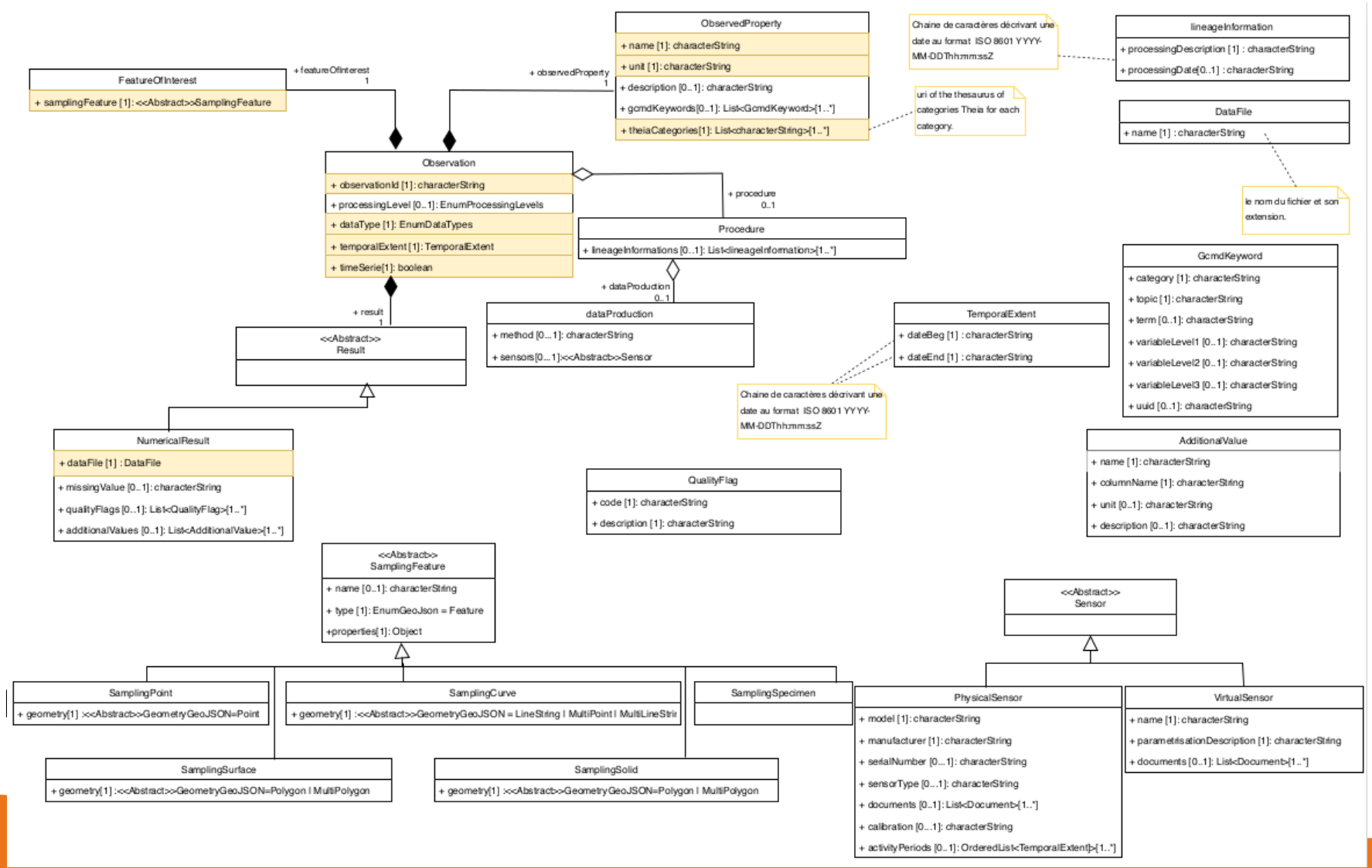
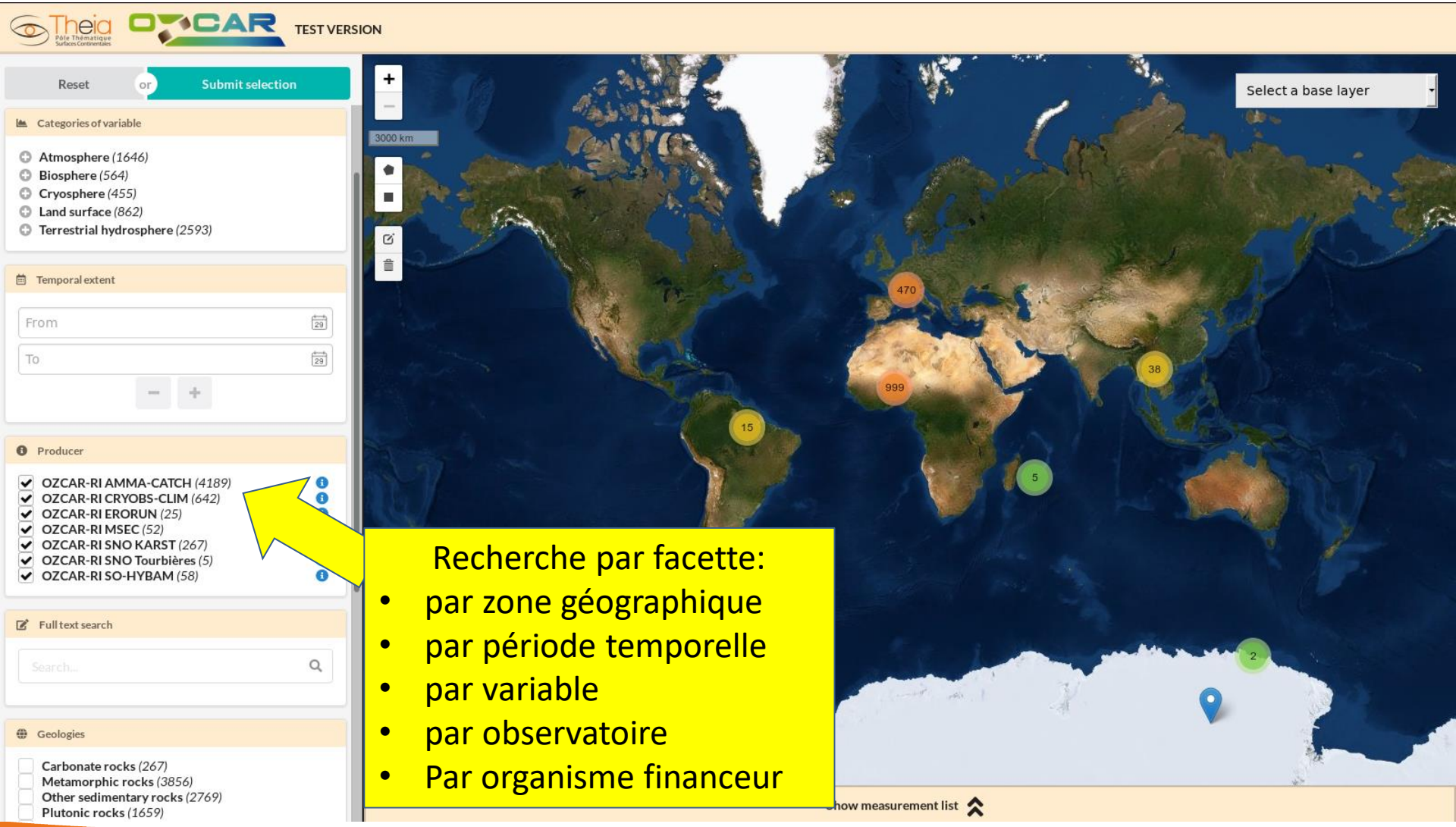


Diagramme de classe

Format pivot (data model) 2/2



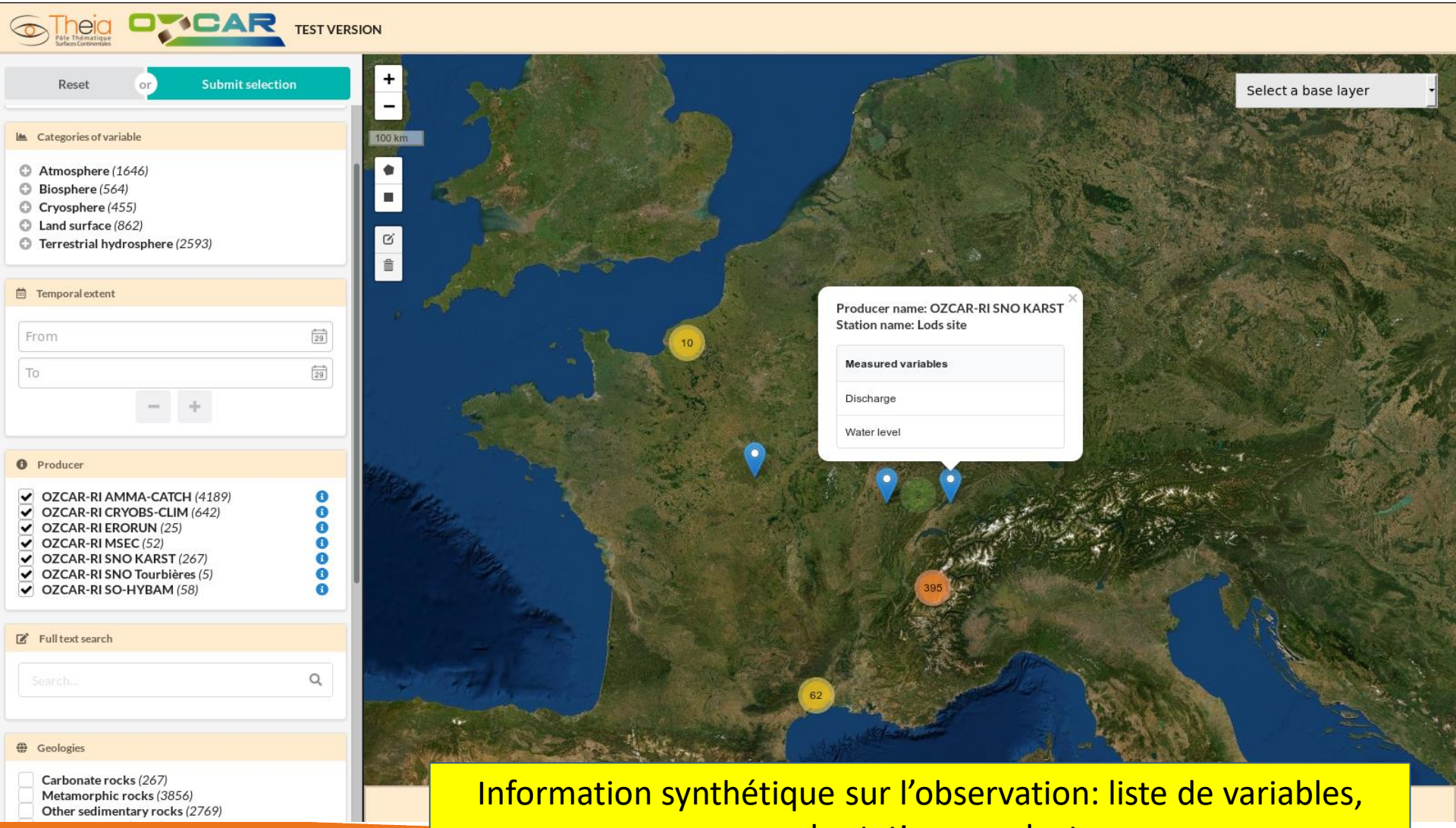
Portail de données



Recherche par facette:

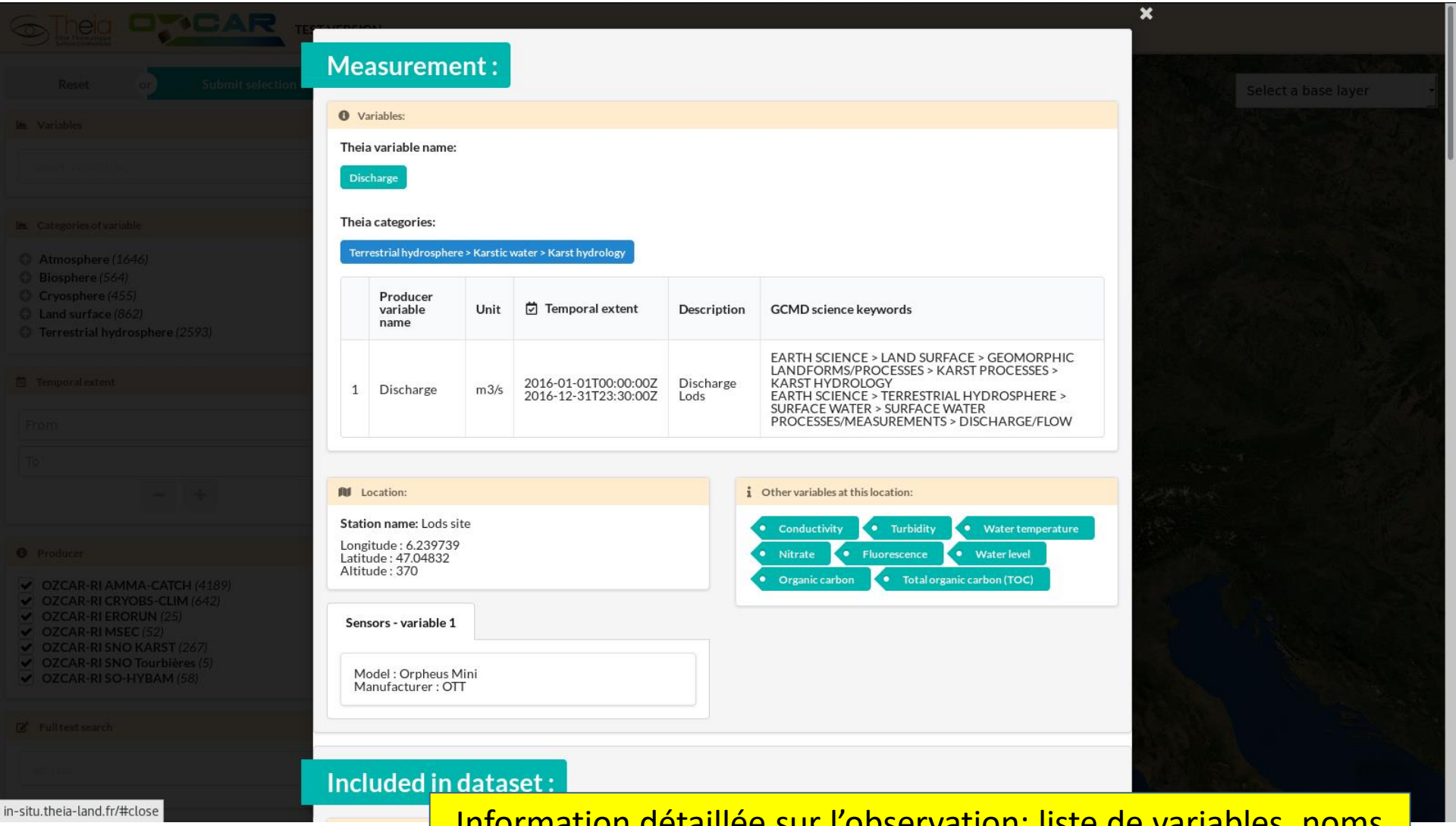
- par zone géographique
- par période temporelle
- par variable
- par observatoire
- Par organisme financeur

Portail de données



Information synthétique sur l'observation: liste de variables, noms de station, producteur

Portail de données



Information détaillée sur l'observation: liste de variables, noms de station, producteur, jeux de données

Prochaines étapes

- Finaliser l'interface du portail
- Finaliser l'architecture de production (déploiement continue docker Kubernetes) et mettre en production (déc. 2019)

2020-2022

- Mettre en place des services d'interopérabilité (web service CSW geonetwork,web service SOS)
- Fournir les données dans différents formats NetCDF, csv
- Continuer à interfacer les SI des producteurs
- Vocabulaire : objet d'intérêt (FeatureOfInterest) sampled Feature

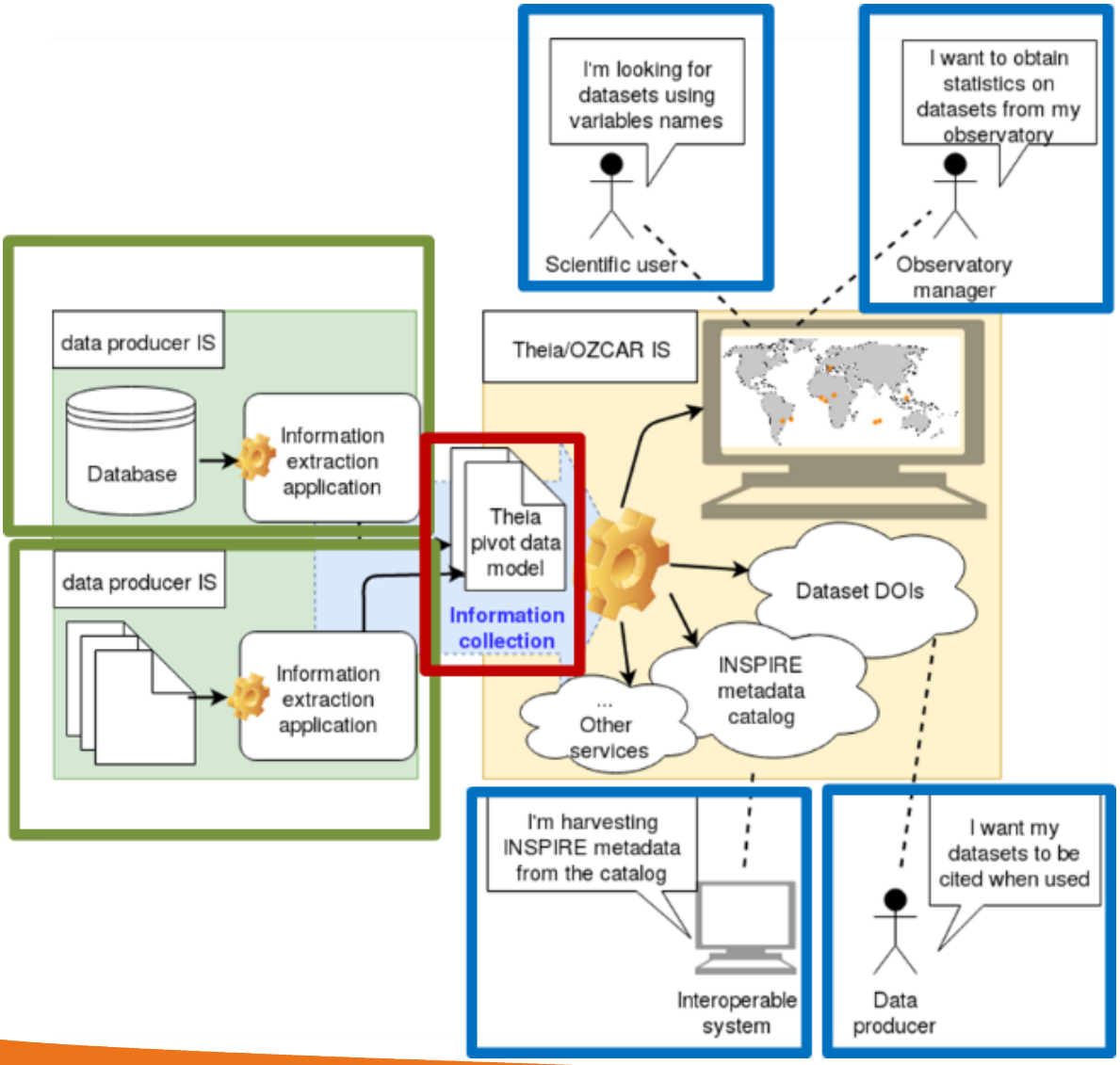
Theia
Pôle Thématique
Surfaces Continentales

OZCAR

Des questions ?

Slides Annexe

Architecture



Construction d'un flux de données :

1) Différents producteurs, différents formats
2) Le **format pivot** permet :
   (i) de collecter ces informations,
   (ii) leur mise à jour en temps réel
3) Le SI Theia/OZCAR répond aux requêtes d'utilisateurs humains ou machine

Architecture du système et développement d'un prototype