

Le Système d'Information Theia/OZCAR :

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

Architecture, Démarche DevOps (infra OSUG)

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)

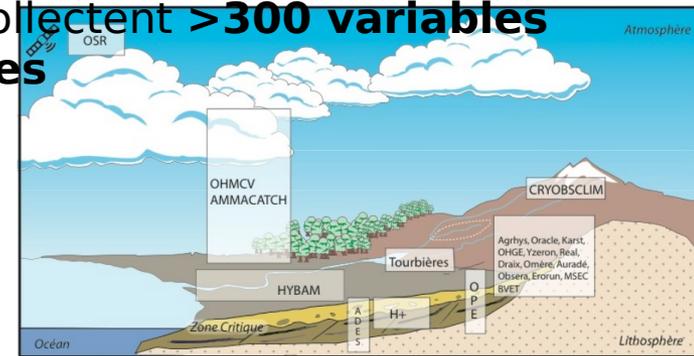
Equipe Développement:

- Véronique Chaffard (IRD/IGE, responsable technique)
- Charly Cousot (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 **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)



: un réseau de 21 observatoires



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



Une grande diversité de mesures in situ

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...

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

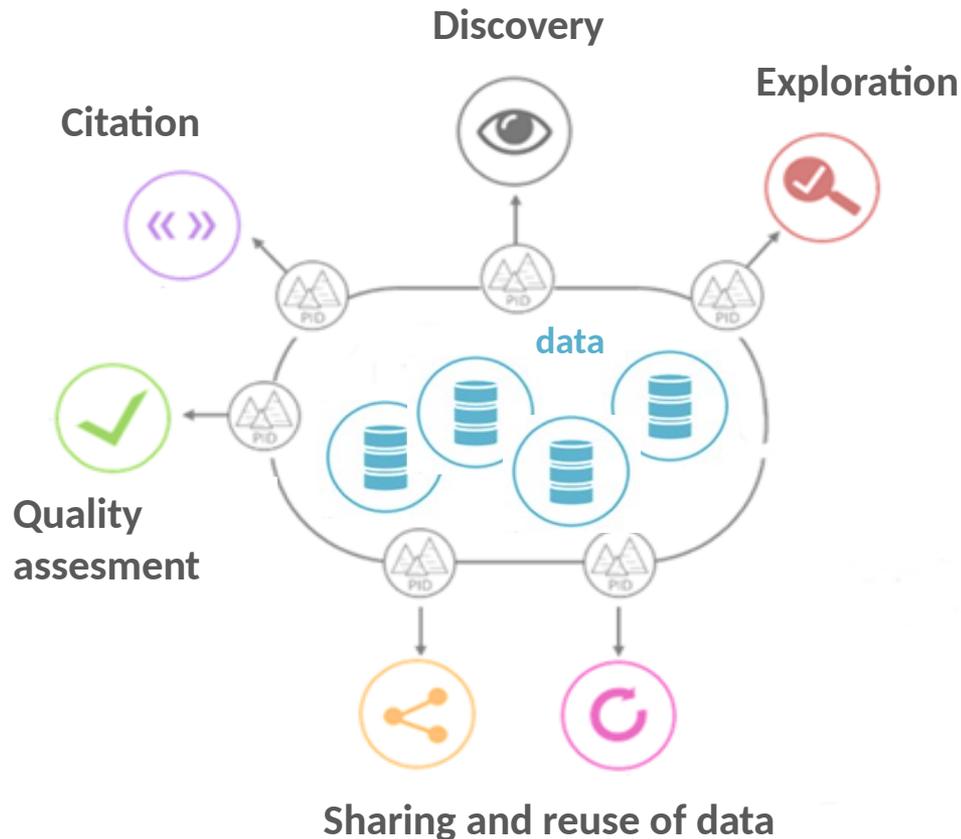


Objectifs du SI Theia/OZCAR

- **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**



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

Les différents services du SI actuel

1. Dépôt des métadonnées/données
2. Portail de métadonnées
3. Thesaurus
4. Interface d'associations des variables

Les différents services du SI actuel

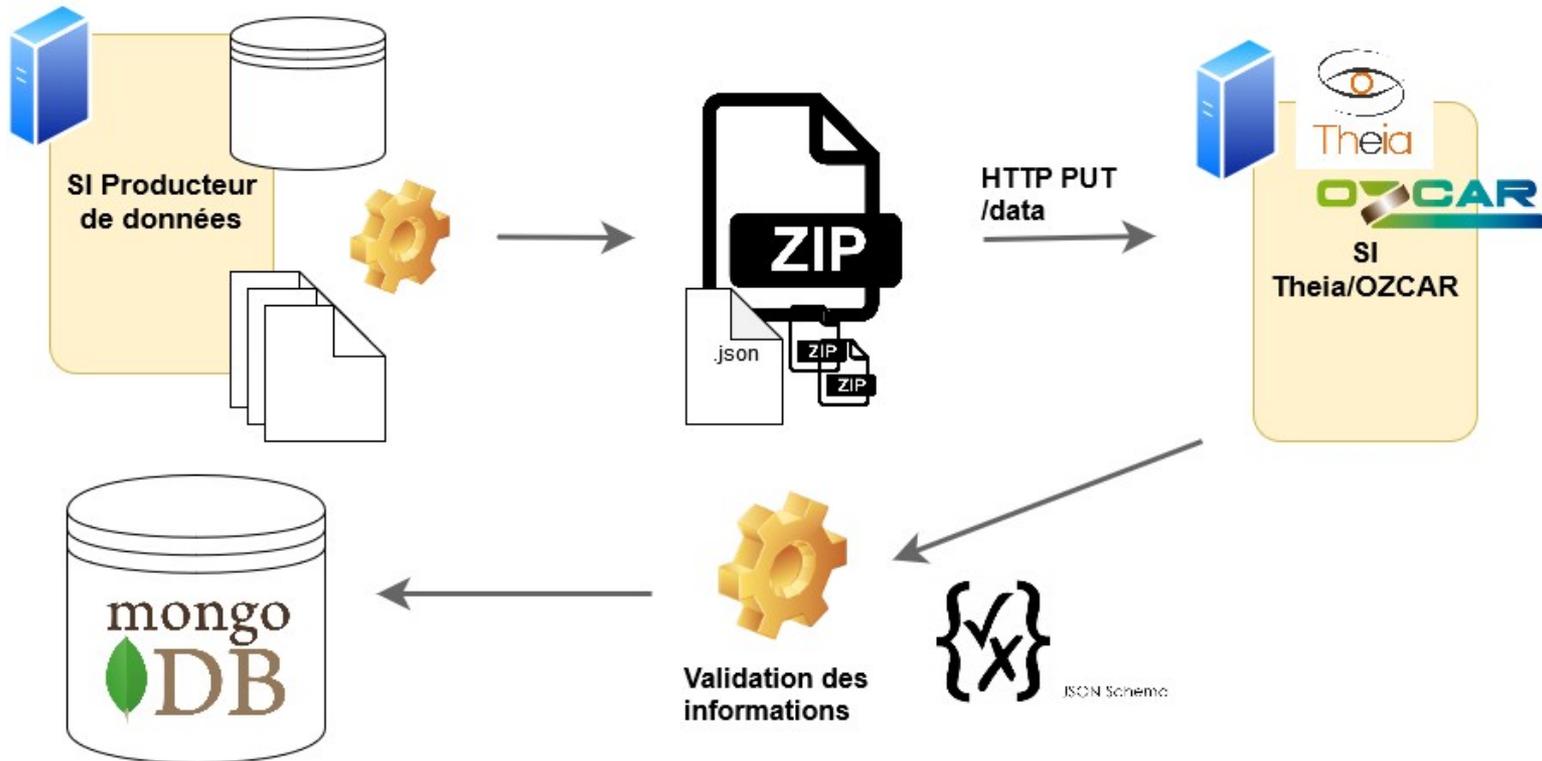
1. Dépôt des métadonnées/données

2. Portail de métadonnées

3. Thesaurus

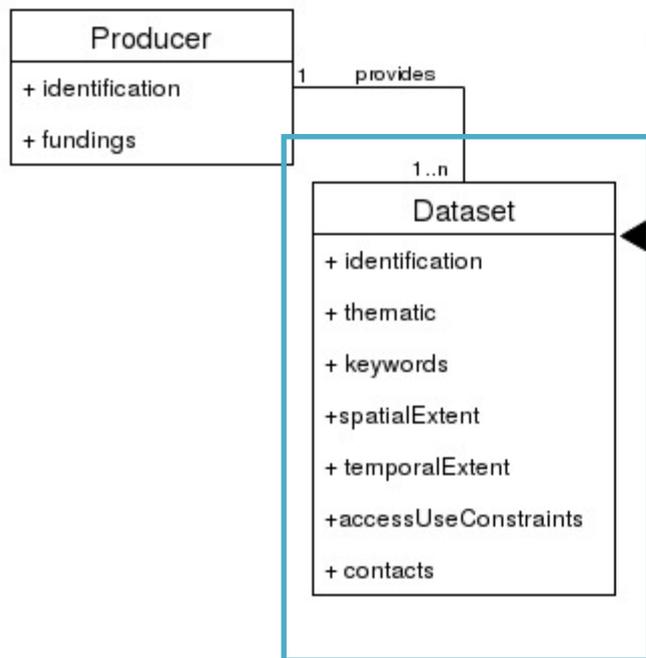
4. Interface d'associations des variables

1 - Dépôt des métadonnées/données

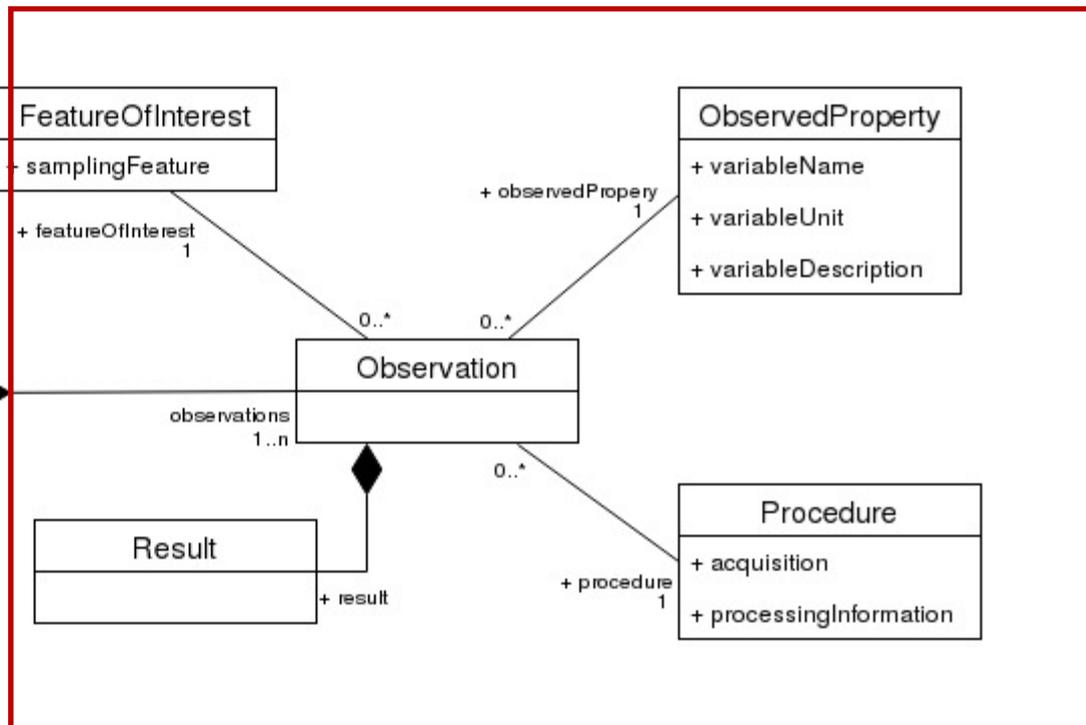


1 - Dépôt des métadonnées/données

Définition d'un modèle pivot
commun basé sur les standards ->
Fichier JSON



O&M



Les différents services du SI actuel

1. Dépôt des métadonnées/données

2. Portail de métadonnées

3. Thesaurus

4. Interface d'associations des variables

2 – Portail de métadonnées

The screenshot displays the OZCAR metadata portal interface. The top header includes the logos for Theia, OZCAR, and OSUG, along with the text "TEST VERSION". The interface is divided into a sidebar on the left and a main map area on the right.

Left Sidebar:

- Reset** and **Submit selection** buttons.
- Variables:** A search box labeled "Select a variable...".
- Categories of variable:**
 - Atmosphere (1646)
 - Biosphere (564)
 - Cryosphere (455)
 - Land surface (862)
 - Soil (805)
 - Erosion/Sedimentation (9) ✓
 - Soil thermal measurements (161)
 - Soil water measurements (635)
 - Surface fluxes (70)
 - Surface radiative properties (2)
 - Terrestrial hydrosphere (2593)
- Temporal extent:** Fields for "From" and "To" with calendar icons and minus/plus buttons.
- Producer:** A list of producers with checkboxes and information icons:
 - ✓ OZCAR-RI AMMA-CATCH (4189)
 - ✓ OZCAR-RI CRYOBS-CLIM (642)
 - ✓ OZCAR-RI ERORUN (25)
 - ✓ OZCAR-RI MSEC (52)
 - ✓ OZCAR-RI SNO KARST (267)
 - ✓ OZCAR-RI SNO Tourbières (5)
 - ✓ OZCAR-RI SO-HYBAM (58)
- Full text search** checkbox.

Main Map Area:

- A world map showing data points as colored circles with numbers: 15 (yellow), 470 (orange), 999 (orange), 38 (yellow), 5 (green), and 2 (green).
- Map controls: zoom in (+), zoom out (-), 3000 km scale bar, home, full screen, and print icons.
- Top right: "Select a base layer" dropdown menu.
- Bottom right: "Show measurement list" button with an upward arrow icon.

Les différents services du SI actuel

1. Dépôt des métadonnées/données

2. Portail de métadonnées

3. Thesaurus

4. Interface d'associations des variables

3 – Thesaurus

Objectif:

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

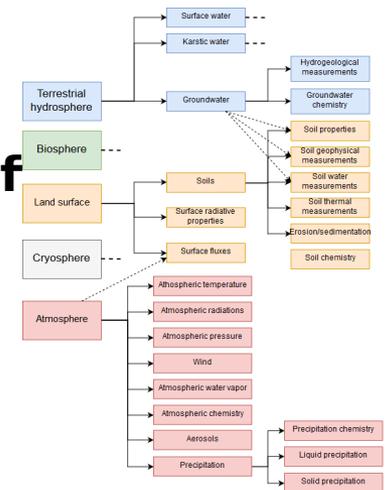
Les étapes de construction: démarche bottom-up

1. **Collecte des noms variables producteurs et classif**
par catégories en se basant sur thésaurus
[NASA GCMD Science keywords](#).

2. Création d'un **vocabulaire de noms de variables**
Theia/OZCAR

3. **Publication du thésaurus en Linked Open Data:**

- Alignement des concepts (lien sémantique) avec des thésaurus internationaux du domaine
- Formalisation en SKOS et Publication sur le Web: outil [Skosmos](#) (s'appuie sur Apache Jena Fuseki SPARQL server)



3 – Thesaurus

Skosmos

Vocabularies About Feedback Help

OZCAR-Theia thesaurus

Content language English Search

Alphabetical Hierarchy Groups

- Nitrogenous parameters (surface water)
- Noble gases (surface water)
- Organic micropollutants (surface water)
- Other mineral Elements (surface water)
- Phosphorous parameters (surface water)
- Variables
 - 3-4-dichloroaniline
 - Absolute humidity
 - Absorbance at 254nm
 - Absorbance at 280nm
 - Aclonifen
 - Actinothermal index
 - Actual evapotranspiration
 - Aerosols chemistry
 - Aerosols radioactive isotope
 - Aerosols size distribution
 - Air pressure
 - Air temperature
 - Albedo
 - Alkalinity
 - Aluminium (Al)
 - Ametryn
 - Amino acid G
 - Aminotriazole
 - Ammonium
 - AMPA
 - **Antimony (Sb)**

... > Surface water > Surface water chemistry > Metals and metalloids (surface water) > Metallic trace elements (surface water) > Antimony (Sb)
Variables > Antimony (Sb)

PREFERRED TERM

Antimony (Sb)

BROADER CONCEPT

[Metallic trace elements \(surface water\)](#)
[Variables](#)

BELONGS TO GROUP

[Variables](#)

URI

<https://w3id.org/ozcar-theia/variables/antimonySb>

Download this concept:

[RDF/XML](#) [TURTLE](#) [JSON-LD](#)

EXACTLY MATCHING CONCEPTS

| | |
|---|--|
| Antimony | AnaEE Thesaurus |
| antimony | EnvThes |
| antimony | EnvThes |
| http://aims.fao.org/aos/agrovoc/c_510 | aims.fao.org |
| http://id.agrisemantics.org/gacs/C15047 | id.agrisemantics.org |
| http://linkeddata.ge.imati.cnr.it/resource/EARTh/63240 | linkeddata.ge.imati.cnr.it |

Les différents services du SI actuel

1. Dépôt des métadonnées/données
2. Portail de métadonnées
3. Thesaurus
- 4. Interface d'associations des variables**

4 – Interface d'association des variables



Variables association administration interface

Selected producer: HYBA

[Back to producer selection](#)

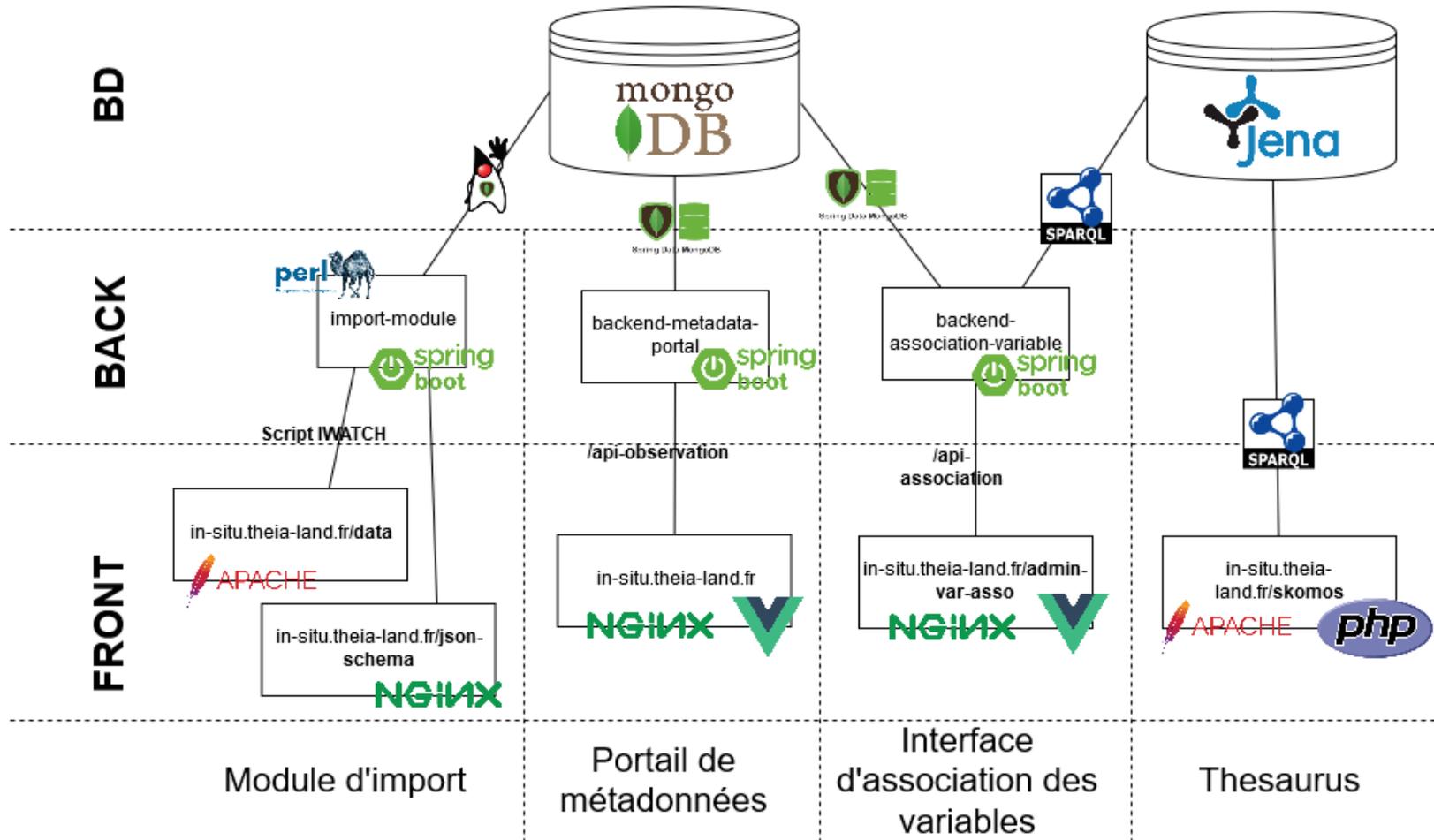
Variable table

| <input type="checkbox"/> | Non-associated variables ▾ | Unit | Theia Categories | Theia variable associated |
|--------------------------|--|------|---|---------------------------|
| <input type="checkbox"/> | Césium | µg/l | https://w3id.org/ozcar-theia/surfaceWaterMetallicTraceElements | |
| <input type="checkbox"/> | Mercure | µg/l | https://w3id.org/ozcar-theia/surfaceWaterMetallicTraceElements | |
| <input type="checkbox"/> | Satellite-derived surface sediment concentration | mg/l | https://w3id.org/ozcar-theia/surfaceWaterGlobalIndices | |
| <input type="checkbox"/> | Surface suspended sediment conc. | mg/l | https://w3id.org/ozcar-theia/surfaceWaterGlobalIndices | |

Number of variables selected for association: 0

[Associate the selection](#)

Les choix techniques



Les choix techniques



MongoDB (base de données orientée document):

- Documents stockés en JSON binaire (BSON)
- La granularité des documents correspond aux objets requêtés (1 document = une mesure à un endroit donné sur une période)
- Schema-less = Flexibilité pour un data model qui tend à évoluer



Apache Jena Fuseki (triple store):

- Open-source
- Full text search via Lucene
- Recommandé avec Skosmos

Les choix techniques



Spring Framework (framework Java):

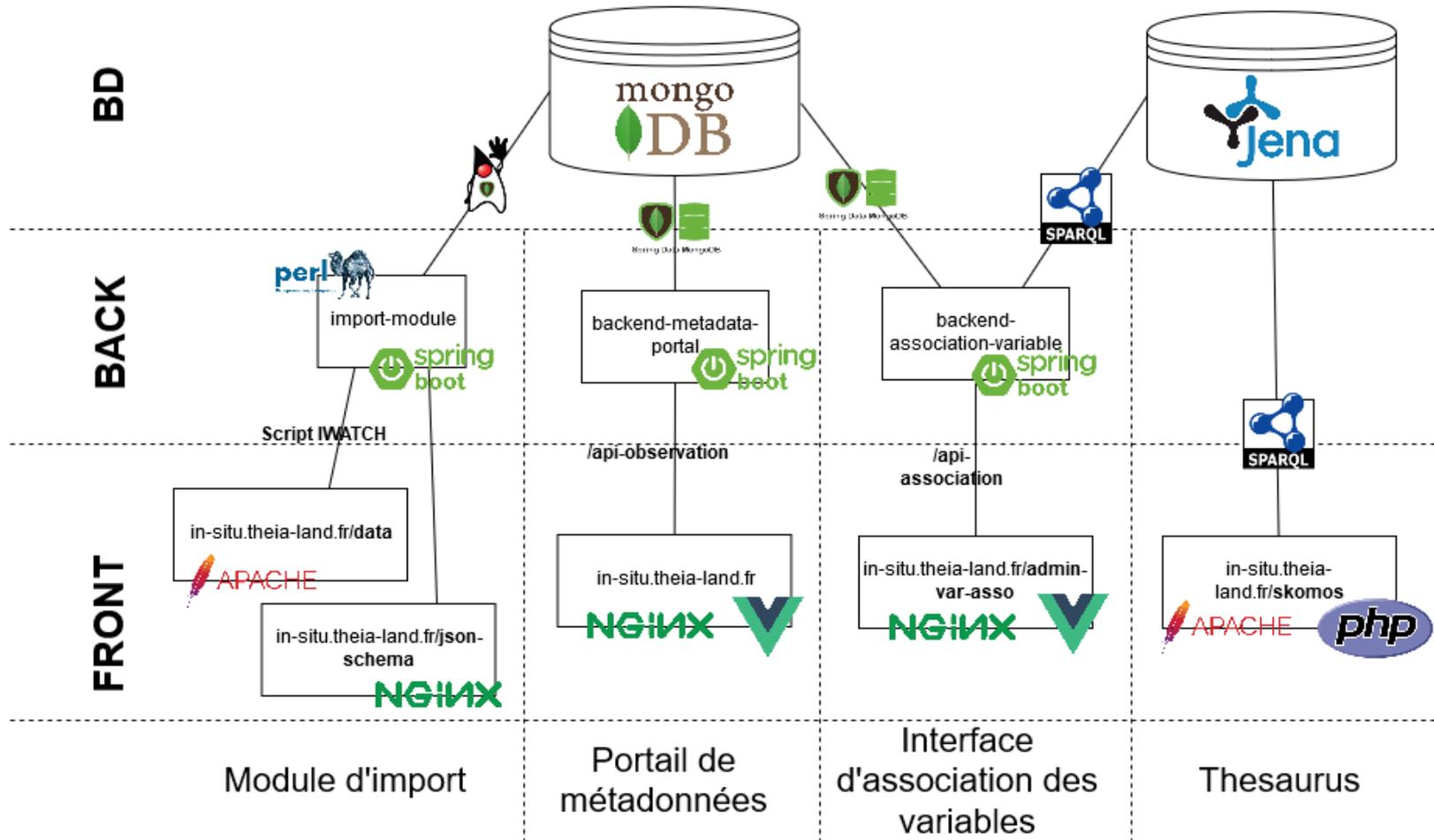
- Historique: langage Java utilisé dans les SI des observatoires à l'IGE
- Framework Java le plus populaire
- Facilite la création d'application standalone



Vue.js (framework javascript):

- Utilisé pour le développement du portail du pôle de données AERIS (IR Data Terra)

Les choix techniques



Le déploiement

Manuel sur une VM de l'OSUG?

- Création des services pour chacun des exécutables
- Beaucoup de configuration / documentation (allocations des ports, mise en place d'un reverse proxy...)
- Pas de haute dispo

Un peu fastidieux...

La solution: **Conteneurisation et orchestration**

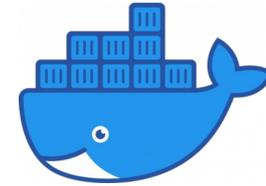


Création et stockage
des images Docker à
l'aide de la forge
GRICAD

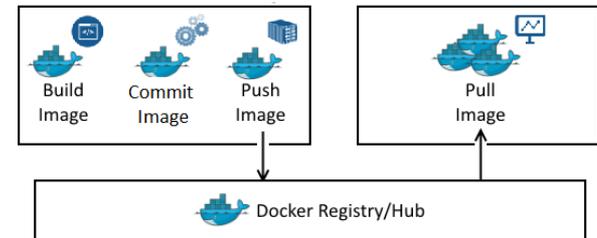
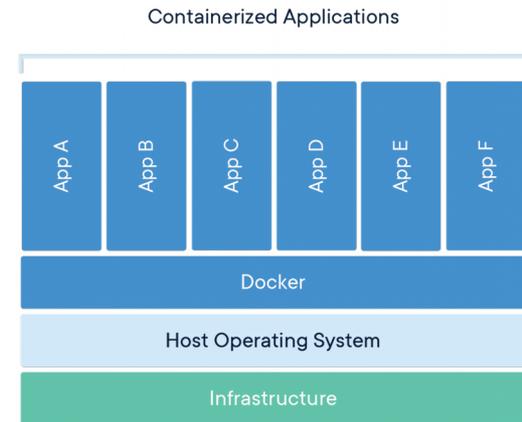


Orchestration sur le
cluster Kubernetes de
l'OSUG

Docker et les conteneurs

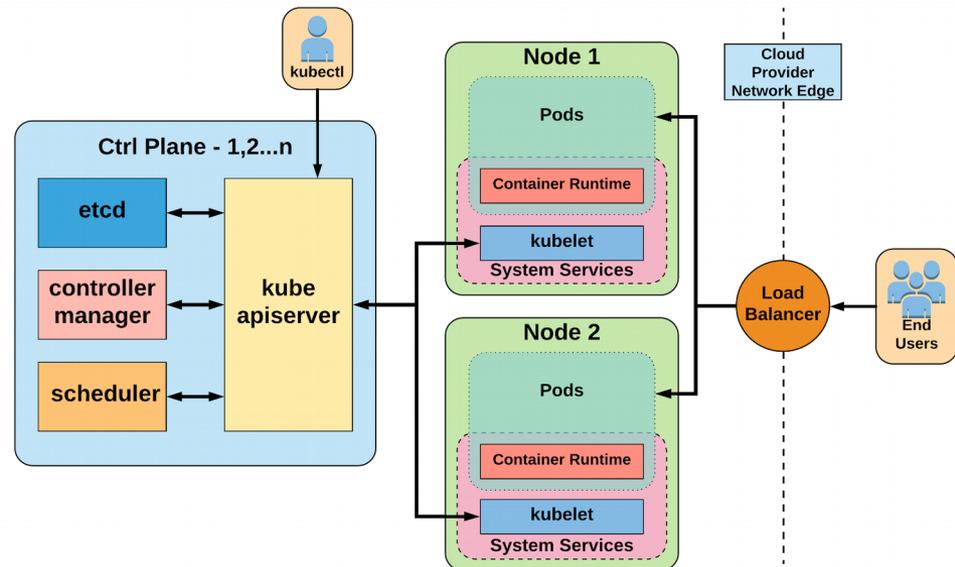


- Docker, présenté à la PyCon 2013
- Isolation en utilisant les capacités du noyau Linux
- Les image de conteneurs : des paquets qui embarquent l'application et ses dépendances
- Un mécanisme d'héritage
- Un système de publication des images



Kubernetes

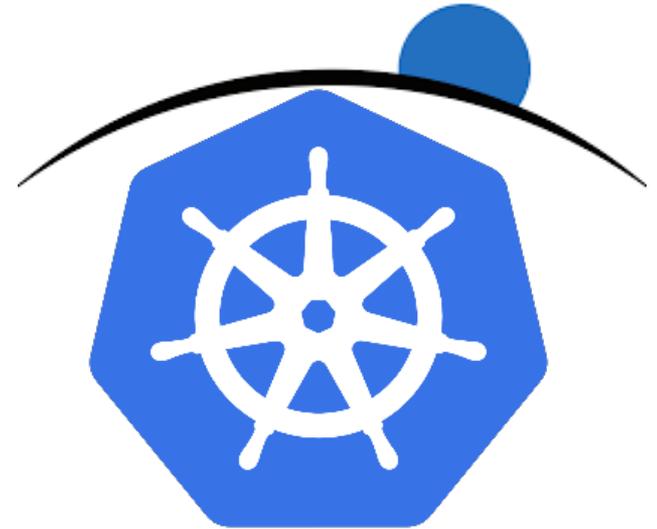
- Projet Google puis Cloud Native Computing Foundation (Linux Foundation)
- Orchestration de conteneurs
- Une API centrale et des composants faiblement couplés
- Bonne intégration aux infrastructures sous-jacentes.





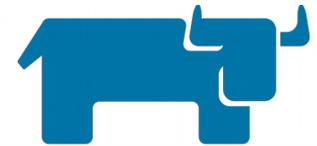
Kubernetes à l'OSUG

- Des compétences depuis fin 2018, production timide en interne depuis début 2019.
- Trois chantiers en cours : JMMC, METABASE et THEIA
- Trois clusters :
 - test, bac à sable pour les mises à jours et test des nouveaux composants.
 - dev, pour les mises au point, réservé à un usage interne.
 - prod, pour la préprod et la prod.



Kubernetes, détails techniques (attention, name-dropping)

- Sur l'infrastructure vSphere OSUG.
- Nœuds avec Saltstack/SaltCloud.
- Cluster avec Rancher Kubernetes Engine.
- Utilisation de Kustomize pour factoriser la configuration des différents clusters.
- Canal (CNI), MetalLB (LB L2), Traefik (LB L4/L7), Prometheus.
- Stockage sur vSAN et NFS (Summer).



Mise en œuvre

- Gitlab (GRICAD) pour les sources et les registry Docker.
- Gitlab CI pour la construction des images.
- Briques tierces :
 - MongoDB, image officielle.
 - Fuseki et Skosmos, pas d'images officielles → images construites par nos soins.



Conteneurisation

- Préparation (« twelfefactorisation) des applications :
 - Logs sur sortie standard
 - Découplage des différentes briques
 - Utilisation de variables d'environnement pour la configuration
- Construction des conteneurs avec JIB (Maven) ou avec des Dockerfile

```
1 FROM node:11.14-stretch as build
2 #Install package dependencies
3 WORKDIR /app
4 COPY package*.json ./
5 RUN yarn install
6
7 #Build the JS bundle
8 COPY . .
9 RUN yarn run build
10
11 #Serve the static content using nginx
12 FROM nginx:1.15
13
14 RUN apt-get update &&\
15     apt-get install -y gettext-base
16
17 COPY --from=build /app/public /usr/share/nginx/html
18 COPY entrypoint.sh /
19
20 EXPOSE 80
21 ENV API_URL "http://localhost/api"
22 ENTRYPOINT ["/entrypoint.sh"]
23 CMD ["nginx", "-g", "daemon off;"]
```

N... < zcar-ui/Dockerfile doc... 56% 13/23 1 : 1

Kubernetesisation

- Un dépôt pour la configuration des déploiements Kubernetes
- Utilisation de kustomize pour test/préprod/prod
- Persistance : utilisation du vSAN via vsphere-cloud-provider
- Sauvegarde : CronJob Kubernetes sur summer
- Difficulté : CORS et Ingresses tordues...

Conclusion

- Conteneurisation facile à mettre en œuvre pour les technos utilisées
- Implique l'externalisation de la configuration
- Kubernetes se prend bien en main malgré quelques zone d'ombre (ingress)!
- Déploiement très facile et rapide
- Application scalable et tolérante aux pannes, mise à jour sans interruption de service
- Visibilité des charges de travail facilitée pour les administrateurs et les développeurs

Futur

- Méthode GitOps (ArgoCD)
- Métriques, logs
- Extension du vocabulaire pour permettre une recherche par objet d'intérêt
- Mise en place des services de catalogages (= nouveaux déploiements/services à intégrer!)
- Développement des services sur les données
- Extension du data-model pour des nouveaux types de mesure in-situ
- Stockage des données?