



DATA  
TERRA



## Compte rendu de l'atelier technique des 23 et 24 mars 2021

CR atelier technique mars 2021



Titre court	
CR atelier technique mars 2021	
Titre long	
Compte rendu de l'atelier technique des 23 et 24 mars 2021	
Auteur principal	
Joël Sudre	
Dissémination	Copyright
Pôle ODATIS	©ODATIS

## Historique

Version	Auteurs	Date	Commentaires
0.1	Joël Sudre	14 avril 2021	Version initiale
0.2	Cécile Nys	15 avril 2021	Relecture et corrections
0.3	Dominique Briand	16 avril 2021	Relecture et corrections
0.4	Christelle Pierkot & Marine Vernet	12 mai 2021	Relecture et corrections
0.5	Cécile Nys	17 mai 2021	Révision et harmonisation des relectures et corrections
0.6	Joël Sudre	18 mai 2021	Relecture
0.7	Cécile Nys	19 mai 2021	Ajout liens & Mise en page finale
1.0	Cécile Nys Caroline Mercier	19 mai 2021	Version diffusable

## Table des matières

<b>1. ACCUEIL ET TOUR DE TABLE DES PARTICIPANTS .....</b>	<b>5</b>
<b>2. INTRODUCTION – (JOËL SUDRE).....</b>	<b>5</b>
<b>3. VRE POUR LA PHYSIQUE .....</b>	<b>6</b>
3.1. PRESENTATION DE L'ENVIRONNEMENT PANGEO/JUPYTERHUB – (FREDERIC BRIOL) .....	6
3.2. TP PRATIQUE ET GUIDE (FREDERIC BRIOL – ANTOINE QUERIC – JOËL SUDRE).....	8
<b>4. FAIR METADATA .....</b>	<b>10</b>
4.1. INTRODUCTION (GILBERT MAUDIRE) .....	10
4.2. WEB SEMANTIQUE (GLENN JUDEAU) .....	10
4.3. METADONNEES SATELLITES AU CERSAT (DOMINIQUE BRIAND).....	12
4.4. L'INTEROPERABILITE SEMANTIQUE DANS L'IR DATA TERRA ET LES POLES DE DONNEES (JEAN-CHRISTOPHE DESCONNETS) .....	13
4.5. EXEMPLES DE FICHES DE METADONNEES DANS LE CATALOGUE ODATIS : SMOS, CORA & ARGO (CHRISTELLE PIERKOT, MARINE VERNET, ROMAIN BOUVIER).....	16
4.6. PREMIERES RECOMMANDATIONS ET PISTES DE TRAVAIL (JEAN-CHRISTOPHE DESCONNETS, CHRISTELLE PIERKOT, MARINE VERNET) .....	19
<b>5. CONCLUSION .....</b>	<b>22</b>
<b>6. ANNEXE .....</b>	<b>23</b>
6.1. QUESTIONS/REPONSES SESSION WEB SEMANTIQUE .....	23
6.2. QUESTIONS/REPONSES SESSION METADONNEES SATELLITES AU CERSAT .....	24
6.3. QUESTIONS/REPONSES SESSION INTEROPERABILITE SEMANTIQUE DANS L'IR DATA TERRA ET LES POLES DE DONNEES .....	25
6.4. EXEMPLES DE FICHES (SMOS & ARGO) .....	26



## Table des illustrations

FIGURE 1. LA PILE LOGICIELLE PANGEO .....	7
FIGURE 2. REPRESENTATION D'UNE STRUCTURE DE DONNEES A N DIMENSIONS AVEC XARRAY .....	7
FIGURE 3. DE LA LISTE A L'ONTOLOGIE FORMELLE .....	14
FIGURE 4. NOTION DE MODELE DE METADONNEES ET LIENS AVEC LES ONTOLOGIES .....	14
FIGURE 5. ONTOLOGIE SOSA (SENSOR, OBSERVATION, SAMPLE AND ACTUATOR) .....	15
FIGURE 6. DIAGRAMME UML DE LA FICHE DU CATALOGUE ODATIS DU PRODUIT SATELLITAL SMOS SUR LE MODELE DATA TERRA FAIR. LES PARTIES EN BLEUES CONCERNENT LE MODELE SOSA ET LES PARTIES EN JAUNE LE MODELE DCAT. ....	17
FIGURE 7. DIAGRAMME UML DE LA FICHE DU CATALOGUE ODATIS DU PRODUIT IN-SITU ARGO SUR LE MODELE DATA TERRA FAIR. LES PARTIES EN BLEUES CONCERNENT LE MODELE SOSA ET LES PARTIES EN JAUNE LE MODELE DCAT .....	18
FIGURE 8. DIAGRAMME UML DE LA FICHE DU CATALOGUE ODATIS DU PRODUIT SATELLITAL SMOS SUR LE MODELE DATA TERRA FAIR. LES PARTIES EN BLEUES CONCERNENT LE MODELE SOSA ET LES PARTIES EN JAUNE LE MODELE DCAT. ....	26
FIGURE 9. DIAGRAMME UML DE LA FICHE DU CATALOGUE ODATIS DU PRODUIT IN-SITU ARGO SUR LE MODELE DATA TERRA FAIR. LES PARTIES EN BLEUES CONCERNENT LE MODELE SOSA ET LES PARTIES EN JAUNE LE MODELE DCAT .....	26
TABLEAU 1. LISTE DES PARTICIPANTS A L'ATELIER ODATIS #9 DES 15 ET 16 DECEMBRE 2020 .....	5
TABLEAU 2. ENSEMBLE DE LIGNES DE CODES PERMETTANT UNE VISUALISATION DES ANOMALIES DE SST DANS LA BOITE NIÑO3.4 DU PACIFIQUE. ....	8



## 1. Accueil et tour de table des participants

Tableau 1. Liste des participants à l'atelier ODATIS #9 des 15 et 16 décembre 2020

Liste des participants à l'Atelier Technique #9	
ALVISET Guillaume (IFREMER)	LAUS HEYNDRICKX Céline (CNRS)
BRIAND Dominique (IFREMER)	LE ROUX Jean-François (IFREMER)
BRIOL Frédéric (CLS)	LEBOURG Laurence (IUEM)
BOUVIER Romain (IRD)	LIBES Maurice (CNRS)
CALVAT Pascal (CNRS)	MAUDIRE Gilbert (IFREMER)
CAMPISTRON Laetitia (CNRS - IUEM - FEIRI)	MAUPETIT Charly (IFREMER)
CARVAL Thierry (IFREMER)	MENDES Fabrice (CNRS)
CAUQUIL Pascal (CNES)	MERCIER Caroline (CNES - prestataire)
CHAMPONNOIS Victor (CNRS)	MORENO Richard (CNES)
CHAUVEL Cédric (UBO - IUEM - FEIRI)	NYS Cécile (OceanScope / IFREMER)
DARRAS Sabine (CNRS)	PHILIPPE Anne-Charlotte (IFREMER)
DEBARD Stéphane (IRD)	PIERKOT Christelle (CNRS)
DEQUIDT Mickaël (IFREMER)	PIOLLE Jean-François (IFREMER)
DESCONNETS Jean-Christophe (IRD)	QUERIC Antoine (IFREMER)
DRU Philippe (CNRS - IMEV)	SCHMECHTIG Catherine (CNRS)
GUIMBARD Sébastien (OceanScope)	SELLE Arnaud (Formater)
HARSCOAT Valérie (IFREMER)	SOUDARIN Laurent (CLS)
HOEBEKE Mark (CNRS)	SUDRE Joël (CNRS - UMS CPST)
HOUBIN Céline (CNRS - SBR)	VAN ISEGHEM Sylvie (IFREMER)
JUDEAU Glenn (IFREMER)	VERNET Marine (IFREMER)
KHVOROSTYANOV Dimitry (CNRS)	

## 2. Introduction – (Joël Sudre)

JS présente l'ordre du jour (voir : [Agenda et accès aux présentations](#)<sup>1</sup>). Cet atelier est composé de deux demi-journées en visioconférence. La première demi-journée est dédiée à la prise en main du **Virtual Research Environment (VRE) pour la physique**, qui est implanté sur HAL (le HPC du CNES) avec une présentation de l'environnement Pangeo/JupyterHub suivie d'un TP pratique et guidé. La seconde demi-journée est consacrée aux **principes FAIR sur les métadonnées**.

<sup>1</sup> <https://www.odatis-ocean.fr/activites/ateliers-techniques/atelier-technique-mars-2021>

### 3. VRE pour la physique

#### 3.1. Présentation de l'environnement Pangeo/JupyterHub – (Frédéric BRIOL)

FB présente la plateforme PANGEO ainsi que sa pile logicielle associée. PANGEO est une plateforme communautaire dédiée au « big data » en géosciences. Elle se compose d'un écosystème de logiciels pour les données massives (ou mégadonnées), d'une communauté internationale de développeurs et d'une infrastructure partagée sur le « Cloud ». Sa mission est d'entretenir un écosystème dans lequel la prochaine génération d'outils, d'analyse « open source » pour les géosciences appliquées à de gros volumes de données peut être développée, distribuée et maintenue. Sa vision est :

- D'avoir un développement ouvert et collaboratif,
- De développer des outils de mise à l'échelle des calculs pour des ensembles de données de petite à très grande taille,
- D'être une structure pour intégrer l'analyse scientifique aux données,
- D'encourager une culture du développement accueillante et participative.

La plateforme possède un :

- Site web : <https://pangeo.io/><sup>2</sup>,
- Github : <https://github.com/pangeo-data><sup>3</sup>,
- Blog : <https://medium.com/pangeo><sup>4</sup>,
- Gitter : <https://gitter.im/pangeo-data><sup>5</sup>.

La pile logicielle disponible sur PANGEO possède plusieurs librairies (voir Figure 1) dont Dask, Xarray et Jupyter qui vont être utilisés principalement dans les travaux pratiques de cet atelier. Les données sont stockées et cataloguées sur un système de stockage distribué accessible à l'échelle globale.

---

<sup>2</sup> <https://pangeo.io/>

<sup>3</sup> <https://github.com/pangeo-data>

<sup>4</sup> <https://medium.com/pangeo>

<sup>5</sup> <https://gitter.im/pangeo-data>





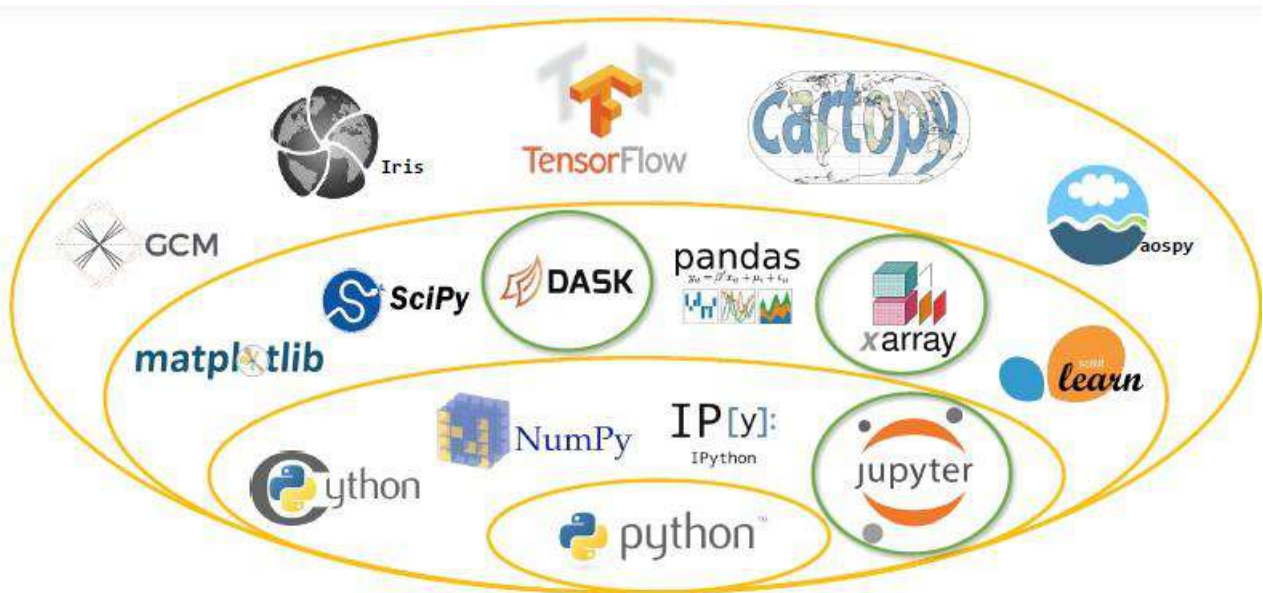


Figure 1. La pile logicielle PANGEO

**Jupyter** est un environnement de calcul qui permet un accès interactif sur des systèmes distants. Il possède un format de document reproductible avec du code, du texte (en Markdown), l'écriture d'équations (en LaTeX) et de la visualisation.

**Xarray** fournit des structures de données pour des tableaux en n dimensions. C'est une API inspirée de Pandas et qui a la particularité d'être interopérable avec les outils standards pydata (voir Figure 2).

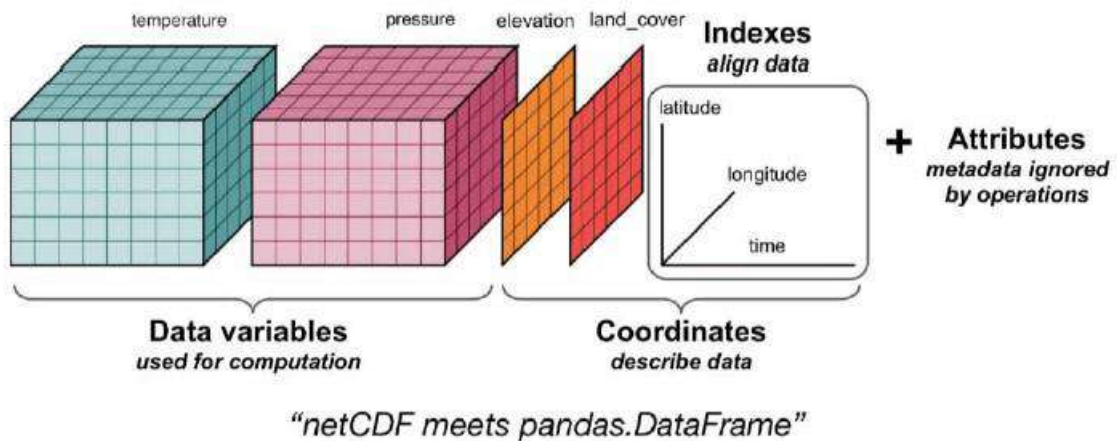


Figure 2. Représentation d'une structure de données à n dimensions avec Xarray

Voici un exemple simple écrit avec Xarray et permettant en quelques ligne de codes (Tableau 2) de faire une visualisation des anomalies de SST dans la boîte Niño3.4 du Pacifique (i.e. : Boite entre 5°N – 5°S et 190°E – 240°E classiquement utilisée pour calculer l'index ENSO) :

Tableau 2. Ensemble de lignes de codes permettant une visualisation des anomalies de SST dans la boîte Niño3.4 du Pacifique.

```
sst_clim = sst.groupby('time.month').mean(dim='time')
sst_anom = sst.groupby('time.month') - sst_clim
nino34index= (sst_anom.sel(lat=slice(-5,5), lon=slice(190,240))
.mean(dim=('lon','lat'))
.rolling(time=3).mean(dim='time'))
nino34_index.plot()
```

**Dask** permet de communiquer avec les nœuds de calcul et de leur donner des instructions. Dask prend en compte les systèmes de calcul parallèle qui sont construits sur Kubernetes ou HPC et permet de paralléliser simplement la lecture/écriture et les calculs sur les nœuds de calcul. Dask va donc permettre d'envoyer des instructions aux ordonnanceurs (« Schedulers ») de manière synchrone, multi-threadés (plusieurs threads sur un cœur unique) et/ou multicœurs mais aussi de manière distribué. Dask permet à partir d'un code de haut niveau écrit avec les bibliothèques telles que NumPy, Pandas, Scikit-Learn, de transformer le code en graphe et ensuite de l'exécuter sur le HPC, une machine multicœurs ou multi-threadés. Lors de son exécution sur une des machines précédemment citées, il est possible au moyen de **Jupyterlab** de suivre l'exécution sur les theards ou les cœurs.

### 3.2. TP pratique et Guidé (Frédéric BRIOL – Antoine QUERIC – Joël SUDRE)

A la suite de la présentation de FB sur Jupyter, Xarray et Dask, les participants à l'atelier sont séparés en trois groupes distincts afin de rendre les TP plus conviviaux. Ces groupes sont animés par FB, AQ et JS. L'objectif est de prendre en main l'IHM (Interface Homme Machine) Jupyter sur le cluster HAL et ensuite d'exécuter pas à pas différents notebooks. Les notebooks ont été mis à disposition sur le site ODATIS ([Pangeo – Notebooks Dask en exemple](#)<sup>6</sup>) afin de les télécharger après les travaux pratiques.

Trois notebooks (ouvert à tous publics : débutant ou non) permettent de découvrir les concepts présentés plus haut. Les différentes cellules présentées doivent juste être exécutées et une explication de chaque cellule est donnée par l'animateur :

- Présentation des concepts de Dask : ***pangeo\_01\_dask.ipynb***
- Analyse des grilles DUACS : ***pangeo\_02\_duacs\_l4.ipynb***
- Xarray : ***pangeo\_03\_xarray.ipynb***

<sup>6</sup> <https://www.odatis-ocean.fr/ressources/logiciels-et-moyens-de-caculs/environnement-de-developpement/dask>



Quatre notebooks (plus axés pour les non débutants) permettent d’approfondir les concepts manipulés par Dask : graphes, tableaux, ordonnanceurs. Les différentes cellules présentées doivent juste être exécutées, et il a quelques exercices en plus. Ces notebooks bien qu’ils nécessitent une connaissance plus approfondie de Python sont tout de même utilisables par un débutant :

- Les graphes de calcul : ***01\_dask\_delayed.ipynb***
- Les ordonnanceurs : ***02\_distributed.ipynb***
- Les tableaux : ***03\_array.ipynb***
- Analyse des collections : ***04\_bag.ipynb***



## 4. FAIR metadata

### 4.1. Introduction (Gilbert MAUDIRE)

Le catalogue ODATIS est constitué de fiches de métadonnées qui sont saisies pour le catalogue, mais aussi de fiches de métadonnées qui sont moissonnées dans d'autres catalogues existants. Il a été observé des disparités entre ces différentes fiches. De plus, il est aussi nécessaire qu'au niveau de l'IR DATA TERRA avec l'ensemble des pôles et le PNDB, les fiches des catalogues soient les plus homogènes possibles. L'idée de cet atelier est donc d'améliorer les recommandations pour remplir les fiches de catalogue afin de rendre le catalogue le plus homogène possible y compris avec les autres pôles.

Il est aussi nécessaire d'être connecté aux travaux européens qui sont fait autour de la FAIRisation des métadonnées. Mais aussi de travailler sur la hiérarchisation des fiches du catalogue afin d'avoir au premier niveau (niveau découverte) un peu moins de fiches et un second niveau (niveau détaillé) plus étoffé.

Enfin, cet atelier doit permettre de faire un état des lieux des petits problèmes constatés sur certaines fiches afin de montrer comment les résoudre.

### 4.2. Web sémantique (Glenn JUDEAU)

GJ présente le web sémantique (voir [202103\\_ODATIS\\_Atelier\\_GJudeau\\_web\\_semantique.pdf](https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202103/202103_ODATIS_Atelier_GJudeau_web_semantique.pdf)<sup>7</sup>) avec ces principes et quelques cas concrets qui ont été implémentés à l'Ifremer. Le but du web sémantique est d'assurer l'interopérabilité (i.e. : la capacité que possède un produit ou un système à fonctionner avec d'autres produits ou systèmes existants ou futurs – un système étant aussi bien un être humain ou une machine - W3C). Ce web sémantique s'articule donc autour du « linked data » qui est composé de 3 grand principes :

- Des URI HTTP (Uniform Resource Identifier) pour identifier des objets qui peuvent être du :
  - Texte, HTML, XML, JSON, son, vidéo, application, PDF, etc. ;
- De renvoyer des informations structurées quand ces URIs sont interrogées :
  - Selon des ontologies<sup>8</sup> de référence ;
  - Avec un modèle d'échange privilégié comme le RDF (Resource Description Framework) ;
  - Contenu compréhensible par les humains et les machines ;
- D'intégrer dans ces contenus des liens vers des URIs d'autres objets ce qui garantit le lien entre des données différentes.

Le RDF permet de modéliser la donnée sous forme de triplets (Sujet – Prédicat – Objet)

Ex : Un flotteur Argo héberge une sonde :

- Sujet : Flotteur Argo,
- Prédicat : héberge,
- Objet : la sonde.

---

<sup>7</sup>[https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier\\_202103/202103\\_ODATIS\\_Atelier\\_GJudeau\\_web\\_semantique.pdf](https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202103/202103_ODATIS_Atelier_GJudeau_web_semantique.pdf)

<sup>8</sup> Ontologie : modèle formel de la connaissance qui peut définir des concept, propriétés, relations, règle logiques, etc.



Un ensemble de triplets se nomme un graph et l'ontologie est la définition de la sémantique des données, cela va définir le type de nœuds et la nature des prédicats. A l'origine, les données sont stockées dans des bases de données relationnelles et donc pour formaliser ces données en RDF, il va falloir les transformer en triplets. Cette formalisation en triplets va permettre de structurer la donnée et de donner du sens aux ressources. Cet effort permet de rendre compréhensible la donnée et de lui donner un sens. L'avantage d'exposer les données au format RDF, c'est que celles-ci sont de ce fait liées entre elles, ce qui permet de faire des recherches de donnée fédérée. L'utilisation d'ontologie permet quant à elle d'avoir un vocabulaire normalisé et partagé. Ceci permet *in fine* d'avoir des applications avec des solutions transdisciplinaires.

Une fois les données exposées au format RDF, il est possible de les interroger au moyen du langage SPARQL (SPARQL Protocol And RDF Query Language) qui est un langage de requête et de protocole pour rechercher, ajouter, modifier ou supprimer des données RDF. Cette interrogation se fait au travers d'un SPARQL Endpoint qui est un point d'accès HTTP permettant d'exécuter des requêtes SPARQL. Ce SPARQL Endpoint permet d'exposer les données à la fois pour un être humain, pour une machine et ceci de façon structurée (ce qui correspond au principe FAIR). Le « linked data » est donc une démarche globale qui peut être montré via le lien <https://lod-cloud.net><sup>9</sup> qui expose 1255 jeux de données avec plus de 16 100 liens.

Ce « linked Data » est par exemple utilisé par Google. Lorsque vous effectuez une recherche google sur Steven Spielberg par exemple, ses œuvres vont être répertoriés en haut de la page et sur la droite vous avez l'utilisation que fait google du « linked data » (un autre exemple est montré à la planche 11 de la présentation).

Les principes généraux pour implémenter des données sont de partir de la source de données, de les extraire, de les convertir en triplets aux moyens de modèles d'ontologie et ensuite de charger ces triplets dans un triple-store qui est l'équivalent d'une base de données optimisée pour gérer des triplets. Et enfin d'exposer ce triple-store via un SPARQL Endpoint.

Dans le cas de SeaDataNet (SDN) pour les « Cruises Sea Report » (CSR), les données sont stockées dans une base Oracle. Via des requêtes SQL et des procédures Java, il est possible de créer des fichiers turtles (.ttl) qui stockent des triplets. Ces fichiers servent d'intermédiaire pour ingérer les triplets dans Jena (API/Fuseki) qui permet de les exposer dans un SPARQL Endpoint.

Dans le cas des données Argo, les fichiers ont été générés via deux APIs (ArgoDataSelection et FleetMonitoring) et des requêtes en HTTP. Puis au moyen de Python ces fichiers ont été converti au format turtle pour ensuite être ingérés dans Jena.

Dans le cadre du Service d'Administration des Référentiels (SAR) une preuve de concept a été menée au niveau du Système d'Information du Milieu Marin (SIMM) et qui a consisté à interroger deux triple-stores différents pour créer le référentiel de Taxon. Elle a donc été implémentée pour faire un référentiel taxons basé sur 2 SPARQL Endpoint contenant le triple store du MNHN et du SANDRE. Une synchronisation des 2 triple-stores vers le triple store du SIMM est effectuée chaque nuit. Il est aussi nécessaire dans ce cas de faire un rapprochement des ontologies entre celle du MNHN et celle du SANDRE. Cette preuve de concept est toujours actuellement en cours d'implémentation.

---

<sup>9</sup> <https://lod-cloud.net>



Des SPARQL microservices qui permettent d'exposer les données au format RDF existent aussi mais n'ont pas été proposées lors de cette présentation.

*Les questions/réponses qui ont suivi cet exposé se trouve dans l'annexe 6.1.*

### 4.3. Métadonnées satellites au CERSAT (Dominique BRIAND)

DB présente les métadonnées satellites à l'Ifremer et plus particulièrement au CERSAT (voir [202103\\_ODATIS\\_Atelier\\_DBriand\\_metadonnees\\_sat.pdf](https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202103/202103_ODATIS_Atelier_DBriand_metadonnees_sat.pdf)<sup>10</sup>). Le CERSAT est le Centre d'Exploitation et de Recherche SATellitaire de l'Ifremer. Le CERSAT intervient dans la spécification de certains capteurs embarqués sur certaines missions satellites. C'est un centre de traitement de données en « Near Real Time » (NRT – temps quasi-réel) et en « Delayed Mode » (DM – temps différé). C'est aussi un support à la qualité et à l'expertise de certaines données satellites qui intervient dans la CalVal (Calibration et Validation des données) satellitaires. Le CERSAT produit des données évoluées, mais aussi des applications pour la visualisation et l'analyse de ces données en lien avec la communauté utilisatrice. Enfin, le CERSAT produit des séries longues et homogènes afin de servir la communauté scientifique et plus particulièrement celle qui s'occupe du climat (température de surface de la mer ; vagues ; états de mer ; vents ; salinité ; etc.). Ces données représentent ~60To de données réceptionnées/mois, ~200 To d'archive sur bande, ~2Po sur disque et ~25000 tâches/jour (soit ~100h/j). L'ensemble de ces données sont accessibles par différents outils de diffusion et de découverte : FTP, HTTP(S), accès direct, THREDDS (OpenDAP, WMS, WCS), OpenSearch (<https://opensearch.ifremer.fr><sup>11</sup>). L'outil de catalogage au CERSAT est l'outil Sextant (aussi utilisé pour catalogue ODATIS et est basé sur Geonetwork – Norme ISO 19115, protocole OGC) <https://sextant.ifremer.fr><sup>12</sup>.

Pour le profil de métadonnées dans le catalogue CERSAT, la norme utilisée est l'ISO 19115-3 :2018, cette version de la norme est une évolution de la version -2 qui permettait déjà de cataloguer les données de type « image ». La version 3 est utilisée car implémentée dans Geonetwork qui utilise nativement la norme ISO19115-3. Un profil de métadonnées spécifique « données satellites » a été créé pour homogénéiser la façon de compléter et visualiser les données satellites. Ce profil contient :

- Titre, résumé, identifiant, DOI, résolutions ;
- Contacts (Helpdesk, PI, financeur) ;
- Accès et Usage : format, licence, citation ;
- Historique : mises à jour, problèmes, versions ;
- Aperçu de la donnée ;
- Liens de téléchargement et services ;
- Liens vers la documentation (PUM, Validation, Algorithmes, etc.).

---

<sup>10</sup>[https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier\\_202103/202103\\_ODATIS\\_Atelier\\_DBriand\\_metadonnees\\_sat.pdf](https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202103/202103_ODATIS_Atelier_DBriand_metadonnees_sat.pdf)

<sup>11</sup> <https://opensearch.ifremer.fr>

<sup>12</sup> <https://sextant.ifremer.fr>



Pour la plupart des champs ci-dessus, un vocabulaire contrôlé est utilisé via des thésaurus qui sont soit internes, soit partagés par les communautés. Afin de mieux FAIRiser la métadonnée, il faudrait que l'ensemble des thésaurus internes soient partagés par la communauté.

Pour enrichir les métadonnées, le CERSAT utilise la base de donnée du système d'information du « Committee on Earth Observation Satellites » (CEOS – <http://ceos.org><sup>13</sup>) qui regroupe l'ensemble des satellites, ainsi que leurs capteurs, mais aussi d'autres informations et d'autres mots clés en particulier les Essential Climate Variables (ECV).

*Les questions/réponses qui ont suivi cet exposé se trouve dans l'annexe 6.2.*

#### 4.4. L'interopérabilité sémantique dans l'IR DATA TERRA et les pôles de données (Jean-Christophe DESCONNETS)

JCD présente l'interopérabilité sémantique dans l'IR DATA TERRA et les pôles de données (voir [202103\\_ODATIS\\_Atelier\\_JCDesconnets\\_interoperabilite.pdf](#)<sup>14</sup>). L'interopérabilité sémantique permet d'associer une signification aux données, et de les positionner dans un domaine de connaissance. Ceci inclut le développement de vocabulaires et de schémas pour décrire les données et les liens entre les données. Il est aussi nécessaire de décrire les données avec des métadonnées et que, ces métadonnées soient annotées avec des vocabulaires formalisés et partagés. Ces vocabulaires proviennent d'ontologies ou de thésaurus.

Les principes FAIR peuvent être utilisés pour mettre en œuvre l'interopérabilité sémantique:

- Les (méta)données doivent utiliser un langage de représentation des connaissances formel, accessible, commun et ayant un vaste champ d'application → **Interopérabilité syntaxique**,
- Les (méta)données doivent utiliser des vocabulaires qui suivent les principes FAIR,
- Les (méta)données doivent inclure des références vers d'autres (méta)données → **Interopérabilité sémantique**.

Il y a plusieurs degrés de formalisation (Figure 3) qui partent du plus bas par des listes ou des vocabulaires contrôlés (qui sont des listes de termes qui ne sont pas forcément organisés et pas forcément partagés) puis ensuite il y a des hiérarchies taxonomiques qui permettent de mettre de relation entre les termes. Au-dessus il y a les thésaurus qui sont des ensembles de termes qui ont été formalisés et sur lesquels des communautés se sont accordées. Les thésaurus sont organisés à travers des relations entre termes qui sont soit hiérarchiques, soit sémantiques. Quant aux ontologies, elles sont encore plus formelles car elles emportent en plus des thésaurus, des relations logiques, des axiomes et des règles. Il y a donc à chaque niveau un ajout de complexité et de structuration qui va partir du vocabulaire contrôlé jusqu'à l'ontologie. Cette structuration va permettre d'avoir une liste de vocabulaire et de terme qui vont à la fois avoir des définitions mais aussi des relations entre ces termes.

---

<sup>13</sup> <http://ceos.org>

<sup>14</sup> [https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier\\_202103/202103\\_ODATIS\\_Atelier\\_JCDesconnets\\_interoperabilite.pdf](https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202103/202103_ODATIS_Atelier_JCDesconnets_interoperabilite.pdf)



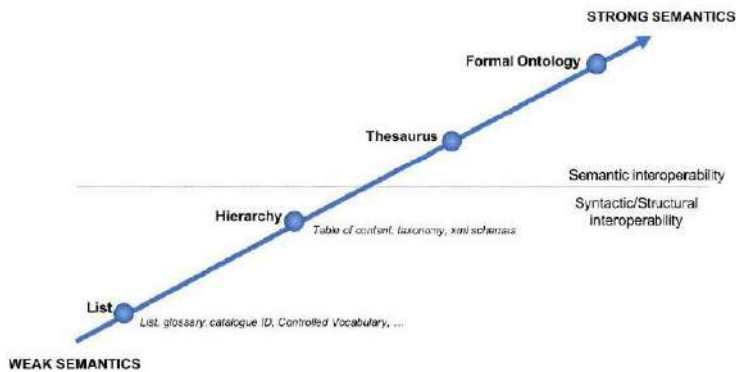


Figure 2: Semantic artefact spectrum. Derived from Leo Obrst, 2010

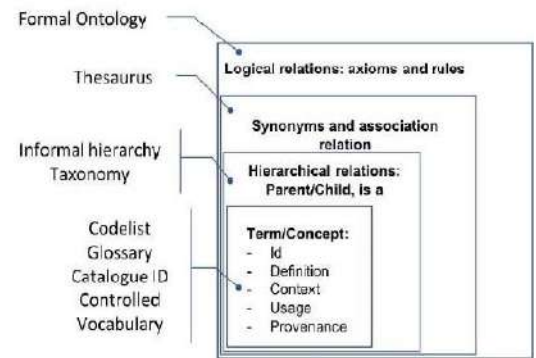


Figure 3: From list to formal ontology, a transformation path.

Figure 3. De la liste à l'ontologie formelle

Le schéma UML de la Figure 4 permet de résumer la notion de modèle de métadonnées et les liens avec les ontologies.

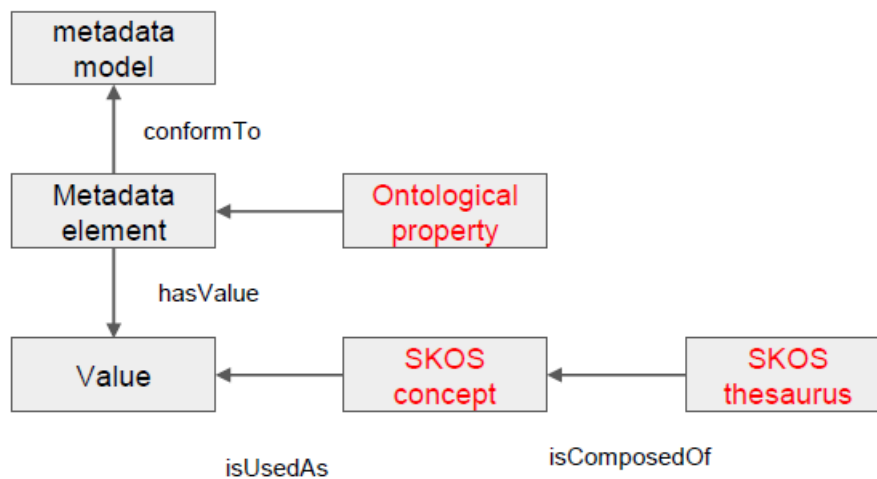


Figure 4. Notion de modèle de métadonnées et liens avec les ontologies

En ce qui concerne les besoins de l'IR DATA TERRA et des pôles de données, il est nécessaire de mettre en place des outils afin de découvrir les données, les services et les traitements de chaque compartiment du système terre sans avoir à les connaître. Cela implique d'avoir une vue de l'ensemble des données et des services pour qu'ils puissent être interrogés et exploités de manière interopérable et transparente pour l'utilisateur final. Pour cela, il a été retenu certains principes pour la découverte et l'accès dans IR DATA TERRA. A savoir, que pour le catalogue, il n'y aura pas un mais plusieurs catalogues, au sein d'une fédération, basés sur un modèle de métadonnées commun sémantiquement riche. Ces catalogues auront deux niveaux de recherche (niveau collection et niveau granule). Il faut aussi qu'il y ait une mise en correspondances (sémantiques) entre les jeux de données qui sera opérées grâce à une standardisation et un alignement des vocabulaires disciplinaires (ce qui implique un contrat d'interopérabilité avec les pôles).



Le slide de 11 de la présentation [202103\\_ODATIS\\_Atelier\\_JCDesconnets\\_interoperabilite.pdf](https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202103/202103_ODATIS_Atelier_JCDesconnets_interoperabilite.pdf)<sup>15</sup> présente un cas d’usage qui illustre l’approche expliquée ci-dessus et une démonstration sur le site <https://dataterra.geomatys.com/><sup>16</sup> est effectué par JCD afin de montrer cette approche en direct.

Derrière le portail de donnée montré précédemment, il y a donc un modèle de métadonnées centré utilisateur qui est l’ontologie SOSA (Sensor, Observation, Sample and Actuator) retenue pour IR DATA TERRA (Figure 5). Cette ontologie est simple mais complètement adaptée à nos données et elle peut être utilisée comme un modèle de métadonnée. En effet une observation porte sur un objet d’intérêt, l’observation est faite par capteur, l’observation porte sur une propriété observée ou une variable et enfin l’observation utilise une procédure. Ce modèle SOSA permet de décrire à peu près toutes nos données et le gros intérêt est cette vision utilisateur qui fait entrer l’utilisateur par des objets qu’il connaît comme les sujets d’étude, les variables ou les capteurs.

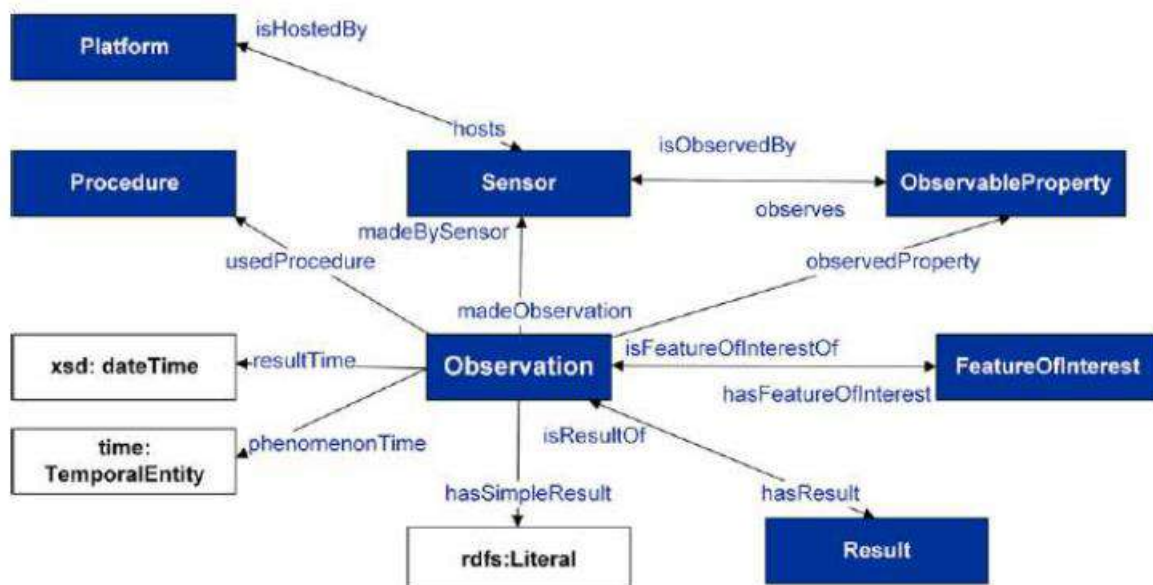


Figure 5. Ontologie SOSA (Sensor, Observation, Sample and Actuator)

De ce fait, il va falloir annoter nos données avec les concepts clés de l’ontologie SOSA, enrichir ces concepts avec les vocabulaires disciplinaires et exploiter les alignements entre termes pour naviguer sur les concepts aux interfaces des compartiments.

Afin de mettre en place ce modèle, un état des lieux des services interopérables des pôles, du PNDB et de CLIMERI par les GTs de l’interpôles a donc été effectué mi-2020. Le but fut d’avoir une vision du niveau d’interopérabilité des catalogues des pôles et de l’IR DT par un inventaire des thésaurus, de leur utilisation dans les catalogues, mais aussi des API d’accès à la donnée, etc. L’objectif fut aussi d’identifier les hétérogénéités des catalogues pour identifier les travaux nécessaires de transformation. Cette réunion fut aussi l’occasion de collecter des métadonnées et données échantillons pour valider le modèle pivot. Cet état des lieux a permis de faire un inventaire et une synthèse sous le prisme des principes FAIR pour avoir les éléments qualitatifs pour analyser,

<sup>15</sup>[https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier\\_202103/202103\\_ODATIS\\_Atelier\\_JCDesconnets\\_interoperabilite.pdf](https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202103/202103_ODATIS_Atelier_JCDesconnets_interoperabilite.pdf)

<sup>16</sup> <https://dataterra.geomatys.com/>

prioriser les actions de FAIRisation des métadonnées et des vocabulaires. Cet inventaire a porté sur les formats de données, les modèles de métadonnées, les API d'accès aux données et métadonnées, les vocabulaires contrôlés et les thésaurus mais aussi sur l'organisation des données (inventaire disponible en suivant ce lien [docs\\_google\\_724670971](https://docs.google.com/spreadsheets/d/10s_uTinEtlCxAl3s0Gf87xfhgeTs6ca594EUoCbd3z8/edit#gid=724670971)<sup>17</sup>).

Au sein du pôle ODATIS, il a été développé des listes de vocabulaire sur les variables, les centres de données, les jeux de données, les thématiques, et certains thésaurus sont déjà utilisés comme WORMS, TAXREF, SeaDataNet (BODC), INSPIRE. La maturité FAIR des pôles a aussi été évaluée au cours de cette revue et il est disponible en suivant le lien : [évaluation\\_FAIR\\_des\\_pôles](#)<sup>18</sup>. Il est important de voir que la maturité FAIR des vocabulaires n'est pas très aboutie et il est nécessaire de les aligner avec les autres communautés, car elles ne sont pas encore réutilisables et interopérables avec les autres communautés des autres pôles. Par contre les thésaurus eux couvrent quasiment l'ensemble des principes FAIR. Un autre point d'amélioration au sein d'ODATIS est d'utiliser les principes FAIR au sein des métadonnées. Dans nos métadonnées, il est nécessaire d'utiliser des mots clés qui permettent de pointer vers des URI afin que le mot clé et sa définition ne soit pas statique. Cela permet aussi d'avoir accès au mot clé (et sa définition), mais aussi aux relations qu'il entretient avec d'autres concepts.

*Les questions/réponses qui ont suivi cet exposé se trouve dans l'annexe 6.3.*

## 4.5. Exemples de fiches de métadonnées dans le catalogue ODATIS : SMOS, CORA & ARGO (Christelle PIERKOT, Marine VERNET, Romain BOUVIER)

Afin de montrer comment bien remplir une fiche de métadonnées dans le catalogue ODATIS et de la rendre plus conforme aux principes FAIR, CP et MV ont repris deux fiches de métadonnées (voir [202103\\_ODATIS\\_Atelier\\_Pierkot\\_SMOS\\_ARGOS.pdf](#)<sup>19</sup>). Ces deux fiches du catalogue ODATIS sont :

- La fiche du produit SMOS (produit satellite),
- La fiche du produit ARGO (produit in-situ).

Ces fiches ont été modifiées afin d'utiliser le modèle SOSA. Ce sont des fiches utilisées comme exemple pour les projets EOSC Pillar et Phidias.

A partir de la fiche de métadonnée ODATIS, qui est au format ISO 19115, un *mapping* a été, dans un premier temps, effectué avec le modèle DATA TERRA basé sur SOSA et sur DCAT. Une fois ce *mapping* effectué, un enrichissement sémantique de la fiche a été mis en place pour la rendre plus interopérable et d'obtenir une fiche enrichie avec des URIs vers les vocabulaires et les thésaurus associés. Cette méthode permet d'intégrer complètement la fiche dans le modèle DATA TERRA et de pouvoir y naviguer comme JCD l'a montré dans sa présentation précédente. Enfin la dernière étape est de voir comment enrichir la fiche initiale (au format ISO 19115) pour prendre en compte toutes ces modifications pour la rendre plus FAIR.

<sup>17</sup> [https://docs.google.com/spreadsheets/d/10s\\_uTinEtlCxAl3s0Gf87xfhgeTs6ca594EUoCbd3z8/edit#gid=724670971](https://docs.google.com/spreadsheets/d/10s_uTinEtlCxAl3s0Gf87xfhgeTs6ca594EUoCbd3z8/edit#gid=724670971)

<sup>18</sup> <https://drive.google.com/file/d/1SUyEOUnl31PvfIVMw2NHSNBZ3lcQaX4P/view>

<sup>19</sup> [https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier\\_202103/202103\\_ODATIS\\_Atelier\\_Pierkot\\_SMOS\\_ARGOS.pdf](https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202103/202103_ODATIS_Atelier_Pierkot_SMOS_ARGOS.pdf)

Pour transformer la fiche SMOS dans le modèle Data Terra, il est nécessaire de voir la terminologie qui a été utilisée pour établir cette fiche (dans les mots clés et le titre par exemple). On retrouve dans cette fiche, du vocabulaire associé, en particulier les variables et les thématiques appartenant au vocabulaire ODATIS, et des mots clés écrits en texte libre. Cette écriture en texte libre des mots clés limite énormément l'interopérabilité et n'est pas conforme aux principes FAIR. La Figure 6 représente sa traduction en UML de la transformation de cette fiche dans le modèle SOSA.

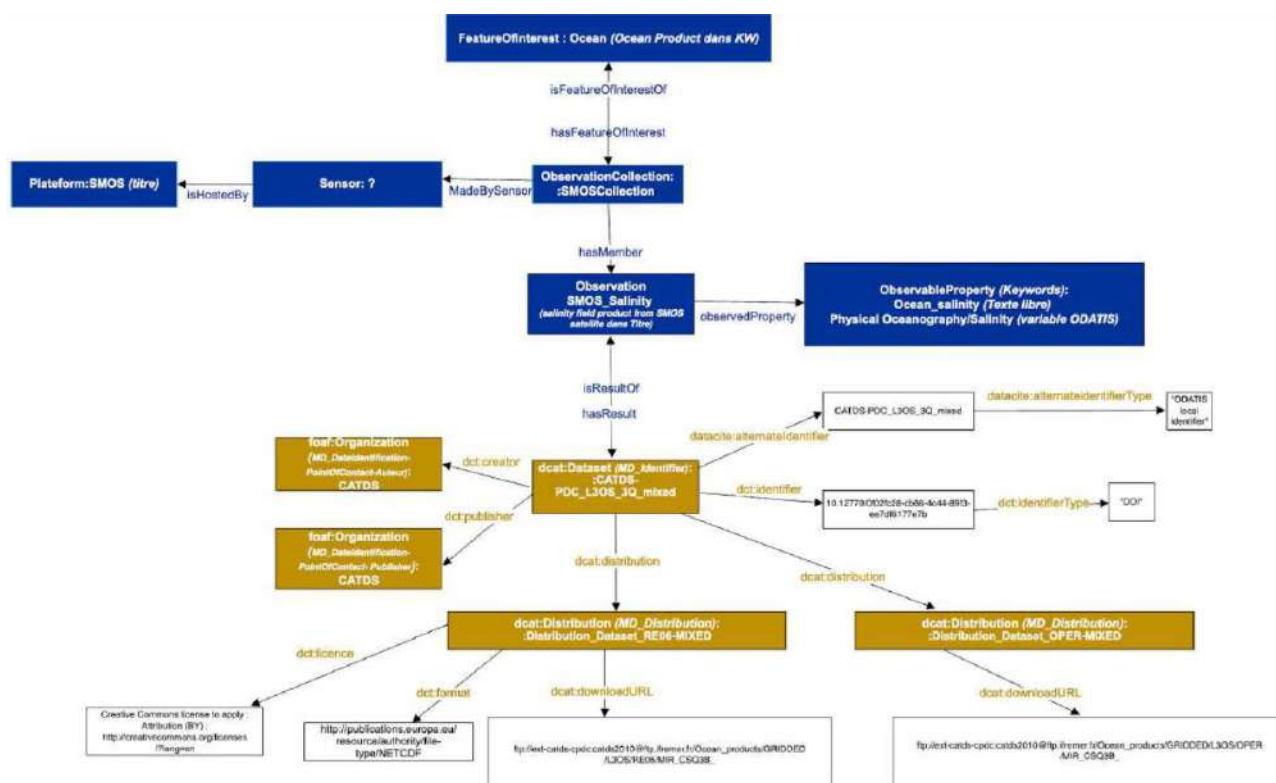


Figure 6. Diagramme UML de la fiche du catalogue ODATIS du produit satellite SMOS sur le modèle DATA TERRA FAIR. Les parties en bleues concernent le modèle SOSA et les parties en jaune le modèle DCAT.

On constate, qu'il y a des informations manquantes, comme par exemple le capteur associé qui n'est pas référencé dans la fiche ODATIS, mais qui est nécessaire dans le modèle SOSA. Il faut également trouver la bonne terminologie pour décrire les propriétés observées, les objets d'intérêt et l'observation du modèle SOSA. Ces informations se trouvent soit dans le titre de la fiche de métadonnées, soit dans les mots clés mais ne sont pas issus de thesaurus pour la plupart. Pour remplir les parties concernant plus le jeu de données, le modèle DCAT est utilisé et on constate qu'il est assez facile de remplir les informations nécessaires au modèle DCAT à partir de la fiche initiale ODATIS. Ce jeu de données a bien été publié par le CATDS et toutes les informations de distribution y sont présentes avec les formats et les informations pour pouvoir télécharger les données. Il est aussi possible grâce au DOI de bien référencer les jeux de données et de le lier à son entrepôt (ici SEANOE). On peut noter qu'il y a deux moyens de télécharger le jeu de données. Il est important de préciser que GeoNetwork permet de faire une sortie DCAT mais ce n'est pas une sortie GeoDCAT-AP telle que présentée dans les présentations précédentes. En effet GeoNetwork permet une sortie GeoDCAT-1 et donc cette sortie se traduit par une perte d'information pour le GeoDCAT-AP nécessaire. Des travaux sont actuellement en cours pour faire des sorties GeoDCAT-AP dans GeoNetwork. Les planches 5, 6 et 7 présentent les modifications qui sont apportées au niveau du

modèle SOSA pour coïncider avec les informations que l'on peut trouver sur le site SMOS (<https://www.salinity-pimep.org/data/input.html><sup>20</sup> – ex : modification de la mean-sea-surface-salinity au lieu de salinity), et dans les vocabulaires utilisés par la communauté (SeaDataNet, EOVS, ...). Une fois que le bon vocabulaire est trouvé pour tous les « termes » (propriétés, variables, « features of interest », etc.), il est nécessaire de trouver les URI associées dans des vocabulaires et des thésaurus recommandés par ODATIS (SDN, EOVS, etc.) qui vont permettre de rendre interopérable la fiche. L'étape suivante est donc de faire de l'enrichissement sémantique (voir planche 7) en insérant les URI du P07 de SDN et de faire des liens sémantiques entre les différents concepts dans la fiche ISO19115 du site ODATIS (voir planche 8). On remplace donc les textes libres de la fiche par des URI sur les thésaurus associés (GCMD, SeaDataNet, CF ou EOVS, etc.) et on fait l'enrichissement sémantique en ajoutant des liens vers d'autres termes ou d'autres vocabulaires.

Cette démarche d'enrichissement sémantique de la fiche ODATIS a aussi été effectuée sur la fiche du produit ARGO (données in-situ). Le *mapping* sur le modèle DATA TERRA FAIR de cette fiche est représenté en Figure 7 et les planches de 10 à 13 montrent l'enrichissement sémantique qui a été effectué sur cette fiche.

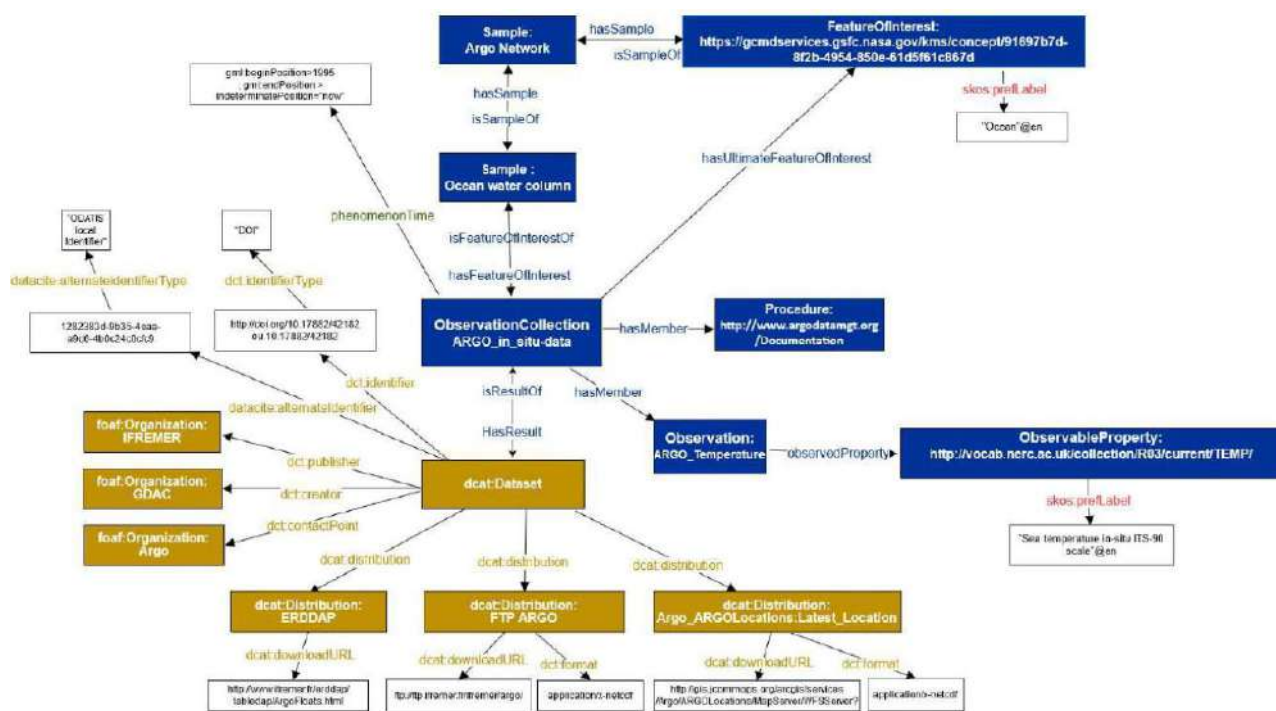


Figure 7. Diagramme UML de la fiche du catalogue ODATIS du produit in-situ ARGO sur le modèle DATA TERRA FAIR. Les parties en bleues concernent le modèle SOSA et les parties en jaune le modèle DCAT

<sup>20</sup> <https://www.salinity-pimep.org/data/input.html>



Voici sous forme synthétique les bonnes pratiques qui sont nécessaires pour améliorer les fiches de métadonnées du catalogue ODATIS :

- Prendre la fiche de métadonnées ISO 19115 du catalogue ODATIS et l'éditer (en XML, en html, etc.) ;
- Faire un *mapping* de la fiche avec le modèle DATA TERRA : SOSA (pour l'observation et ses propriétés) et GeoDCAT-AP (pour le jeu de données). (cf. le modèle DATA TERRA avec les deux modèles SOSA et GeoDCAT-AP en annexe 6.4) ;
- Vérifier l'ensemble des termes utilisés dans le *mapping* en partant des renseignements fournis par la fiche de métadonnée ou trouvés sur les sites de production des produits s'ils existent ;
- Si des termes sont manquants dans le *mapping* SOSA/GeoDCAT-AP, les trouver, si possible dans les thésaurus utilisés par la communauté ;
- Rechercher l'ensemble des termes utilisés/trouvés dans les vocabulaires et thésaurus recommandés par ODATIS et faire les liens de mapping entre les concepts ;
- Mettre à jour l'ensemble de ces informations dans la fiche ISO19115 en remplaçant tous les termes et texte libre de la fiche ISO 19115 par les URIs associés qui ont été trouvés lors de la recherche précédente.

*Les questions/réponses qui ont suivi cet exposé se trouve dans l'annexe 6.4.*

## 4.6. Premières recommandations et pistes de travail (Jean-Christophe Desconnets, Christelle PIERKOT, Marine VERNET)

JCD présente les premières recommandations et pistes de travail qui se trouvent en deuxième partie de la présentation [202103\\_ODATIS\\_Atelier\\_JCDesconnets\\_interoperabilite.pdf](#)<sup>21</sup> (voir à partir de la planche 24).

Les trois pistes de travail, pour améliorer la FAIRisation des catalogues et l'enrichissement des métadonnées, sont de :

- Créer, structurer et enrichir les vocabulaires existants pour travailler sur les enjeux de découverte et d'accès aux données d'ODATIS vers l'IR DATA TERRA ou vers d'autres catalogues comme les catalogues européens ;
- Améliorer la FAIRisation des vocabulaires, ce qui passe par le fait qu'elles soient servies et exposées via des protocoles et des formats standards et également pérennisées ;
- Améliorer l'utilisation des vocabulaires dans les métadonnées des catalogues de données.

Chaque discipline a adapté son vocabulaire à ses besoins, il n'est dès lors pas possible d'imposer un vocabulaire mais il est plutôt nécessaire de rendre ces vocabulaires accessibles via des URI avec des protocoles et des formats standards pour qu'ils puissent être traités et interprétés automatiquement. Grâce à cela, un producteur de données va pouvoir aller sur ces services qui vont mettre à disposition ces vocabulaires, via des thésaurus, pour voir les définitions qui correspondent aux produits qu'il a développés, et que les machines vont aussi pouvoir se connecter pour accéder à

---

<sup>21</sup>[https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier\\_202103/202103\\_ODATIS\\_Atelier\\_JCDesconnets\\_interoperabilite.pdf](https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202103/202103_ODATIS_Atelier_JCDesconnets_interoperabilite.pdf)



ces terminologies. Il y a donc à la fois un travail d'enrichissement des vocabulaires des thésaurus à faire pour intégrer de nouveaux termes et de nouvelles définitions, mais aussi un travail d'alignement des vocabulaires et des thésaurus afin de faire des liens entre des thésaurus et des termes.

A propos des bonnes pratiques, à mettre en place au sein d'ODATIS pour la création de nouvelles terminologies, en acceptant que le paradigme d'observation est celui que l'on doit adopter via le modèle SOSA, il est nécessaire :

- D'utiliser les thésaurus disciplinaires existants et de référence, (ne pas réinventer mais plutôt d'inventorier ce qui existe en premier lieu) ;
- De contribuer, si possible, à enrichir les thésaurus existants en évitant le plus possible que ces thésaurus deviennent des thésaurus locaux. Ceci permet qu'ils soient partagés par le plus grand nombre. (A noter que c'est une bonne pratique déjà adoptée dans plusieurs CDS d'ODATIS par l'enrichissement du vocabulaire du BODC via SDN par exemple) ;
- D'enrichir ces thésaurus en les organisant autour de la vision d'observation : Variables, Plateforme/Capteur, Objet étudié.

Pour rendre les terminologies FAIR, il y a d'abord une nécessité :

- De les formaliser, ce qui implique un volet scientifique de la part des experts scientifiques du domaine qui doivent faire cet effort de formalisation ;
- De les maintenir et de les faire évoluer, qui est également une partie du volet scientifique ;
- De les aligner, ce qui implique une partie scientifique car il faut une bonne connaissance du domaine, mais aussi une partie technique car il est nécessaire de faire ces alignements via des outils et des formats qui vont permettre de les décrire et de les maintenir ;
- De les préserver (ce qui est plutôt du domaine technique) ;
- De les partager et de les exposer (ce qui est aussi du domaine technique).

Au sein d'ODATIS, la partie scientifique de la FAIRisation des terminologies peut se faire dans les CDS, dans les CES, au cours de certains Ateliers Techniques dédiés aux vocabulaires, thésaurus, et référentiels, mais aussi directement par les scientifiques impliqués dans la production de nouveaux produits et de nouvelles données.





En ce qui concerne les outils techniques, qui peuvent nous aider dans ces activités de FAIRisation, il est nécessaire d'avoir une plateforme assurant la gestion, la consultation et l'exposition interopérable des vocabulaires disciplinaires ainsi que leurs pérennités (plateforme qui devrait être mise en place à l'échelle d'un pôle ou de l'IR DATA TERRA, cette question restant encore en discussion). Il existe des outils (open source, portés et maintenus par de grands opérateurs) pouvant aider à cette mise en place en particulier :

**Voc Bench** : Plateforme open source multilingue pour la gestion collaborative d'ontologies en OWL, de thésaurus en SKOS. Cet outil permet de mettre à jour, d'enrichir, d'aligner les ontologies et les thésaurus ;

**Skosmos** : Outil open source permettant de publier, de naviguer et d'exposer des ressources en SKOS sur le web (SKOS étant un standard de thésaurus du W3C qui permet d'exposer des vocabulaires pour leur consommation par des applications comme GeoNetwork). Cette application permet d'avoir un accès à l'ensemble des propriétés d'un concept qui vont être ses définitions, ses libellés, ses liens hiérarchiques et sémantiques avec d'autres liens. Il propose une API REST pour accéder aux référentiels en RDF ;

**UKGovLD Registry** : Outil de gestion et d'exposition de registres de données liées qui a pour fonctionnalités principales la création, la maintenance et l'évolution des listes de codes et de leurs identifiants (URI). Cet outil est utilisé par le BRGM comme registre de données qui permet de maintenir leurs listes et leurs vocabulaires contrôlés.

Ces trois outils permettent d'exposer l'ensemble des ontologies, thésaurus et vocabulaires de manière FAIR ce qui implique que l'ensemble des principes FAIR sont couverts par ces trois outils.

L'INIST a aussi mis en place un service terminologique pluri disciplinaire qui se nomme LOTERRE (Linked Open TERminology REsources - <https://www.inist.fr/services/analyser/loterre><sup>22</sup>). LOTERRE est une plateforme destinée à l'exposition, l'interrogation et le téléchargement de ressources terminologiques scientifiques, inscrite dans une approche de science ouverte. Elle héberge des terminologies produites par l'INIST et propose aussi ses services à d'autres producteurs de données terminologiques. Ce service est construit sur SKOSMOS.

L'IR OZCAR-THEIA a alignée 3 thésaurus (GCMD, SANDRE et CF) pour ses besoins et ils ont fait le choix de SKOMOS pour mettre en place leur thésaurus alignés ([https://in-situ.theia-land.fr/skosmos/theia\\_ozcar\\_thesaurus/fr](https://in-situ.theia-land.fr/skosmos/theia_ozcar_thesaurus/fr)<sup>23</sup>).

Enfin une fois les thésaurus alignés et mis en place, il est nécessaire d'introduire dans les fiches de métadonnées des catalogues les URI comme il a été montré dans les présentations précédentes.

---

<sup>22</sup> <https://www.inist.fr/services/analyser/loterre/>

<sup>23</sup> [https://in-situ.theia-land.fr/skosmos/theia\\_ozcar\\_thesaurus/fr/](https://in-situ.theia-land.fr/skosmos/theia_ozcar_thesaurus/fr/)



## 5. Conclusion

En guise de conclusion, il est possible de faire quelques recommandations pour ODATIS :

- Il est nécessaire de reprendre les fiches du catalogue ODATIS et de renseigner ces fiches de métadonnées avec des vocabulaires contrôlés car nous avons déjà dans ODATIS ce vocabulaire. Il faut à minima renseigner le niveau générique et si possible aller plus loin (ceci n'étant pas toujours possible sur des fiches et données « historiques » agrégeant plusieurs jeux de données de différents capteurs et de différentes méthodes).
- Il est nécessaire de relier tous les vocabulaires utilisés et mettre tous les URIs à ces vocabulaires utilisés dans les fiches de métadonnées. C'est-à-dire remplacer le texte libre (mots clés, nom des capteurs etc.) dans les fiches de métadonnées par les URIs correspondantes prises dans les thésaurus et ontologies recommandées par ODATIS. Les serveurs de vocabulaires permettant de référencer cela sont maintenant en capacité de le faire et sont en place. Ceci est une des étapes les plus importantes pour la FAIRisation des métadonnées dans les fiches du catalogue ODATIS.
- Dans le catalogue ODATIS, il va être nécessaire d'avoir deux niveaux de description. Un premier très générique qui référence par exemple un satellite comme SMOS et tous les jeux de données qui en sont issues. Un second où chaque jeu de données est décrit. Cela permettra de ne pas multiplier les fiches au premier niveau et d'avoir une notion de catalogue hiérarchisé avec un premier niveau qui représente les grands ensembles de données et un deuxième niveau (qui est toujours dans le catalogue de découverte) avec plus de détail où chaque ensemble de données est détaillé. Cette hiérarchisation permettra aussi d'exploiter les niveaux dans les vocabulaires de manière plus précises et de ne pas mettre toutes les fiches au même niveau. Ceci va aussi permettre d'ingérer de grands ensembles composés de plusieurs dizaines de fiches sans pour autant mettre toutes les fiches au même niveau (en râteau). Cela permettra aussi d'ordonner les fiches et les ensembles en fonction de leur granularité ce qui est actuellement très variable dans le catalogue ODATIS. Il va donc être nécessaire de restructurer le catalogue avec les CDS pour mieux prendre en compte cette notion de hiérarchie.
- Il est aussi nécessaire de mettre un titre court pour chaque fiche de métadonnées car certaines fiches ont des titres faisant parfois plusieurs lignes.
- Continuer d'utiliser les vocabulaires contrôlés pour les variables.



## 6. Annexe

### 6.1. Questions/Réponses session Web sémantique

<p>Exemple : un capteur est intégré au flotteur. “intégré” c’est un prédicat ? Où trouve-t-on la liste de tous les prédicats définis ?</p>	<p>“Intégré” est la traduction du prédicat “hosts” de l’ontologie SOSA (<a href="https://www.w3.org/TR/vocab-ssn/#SOSAhosts">https://www.w3.org/TR/vocab-ssn/#SOSAhosts</a>). Tu peux trouver les prédicats définis par l’ontologie SOSA avec le lien précédent.</p>
<p>Une fois que tous les triplets sont stockés dans la “base” Jena, comment se fait l’exposition sur un front. Le front interroge en direct la base “Jena” avec SparQL?</p>	<p>Pour interroger le triple-store (l’équivalent de la notion de ‘base’), il faut qu’un SPARQL Endpoint soit déployé. JENA est un ensemble d’API permettant de manipuler les triplets (données RDF). Le SPARQL endpoint permet d’interroger la ‘base’ (le triple store) directement. Le front interroge donc la ‘base’ via le SPARQL Endpoint.</p> <p>Thierry Carval à propos de Jena :</p> <p>Quand on passe à plusieurs millions de triplets (ex: Argo : 30 millions), JENA ne sait plus répondre avec des temps acceptables. La bonne pratique est d’interroger un index Elasticsearch (ou autre) plutôt que la base Jena.</p>
<p>On peut avoir ces thesaurus rajoutés a geonetwork? (oui je connaissais ceux de seadatanet que j’utilise) , c’était les autres que j’ai vu passer qui m’intéressaient</p>	<p>Pour SeaDataNet, les vocabulaires sont accessibles ici : <a href="http://vocab.nerc.ac.uk/search_nvs/">http://vocab.nerc.ac.uk/search_nvs/</a></p> <p>On peut télécharger les thesaurus ODATIS. La liste se trouve ici : <a href="https://sextant.ifremer.fr/geonetwork/srv/eng/rdf.metadata.get?uuid=faf24a56-dfec-46a5-94d1-2373cc3e7e29">https://sextant.ifremer.fr/geonetwork/srv/eng/rdf.metadata.get?uuid=faf24a56-dfec-46a5-94d1-2373cc3e7e29</a> (Le lien ne semble plus fonctionnel mais je l’ai déjà utilisé!)</p> <p>Lien pour télécharger le thesaurus des variables ODATIS : <a href="https://sextant.ifremer.fr/geonetwork/srv/eng/thesaurus.download?ref=local.theme.odatis_variables">https://sextant.ifremer.fr/geonetwork/srv/eng/thesaurus.download?ref=local.theme.odatis_variables</a></p>
<p>Il a été présenté que le Sandre est interrogé de façon journalière par votre SparQL endpoint (celui qui fait la synchro si j’ai bien compris). Quelles informations sont récupérées?</p>	<p>Le projet est au statut de POC à l’heure actuelle. Je dois vérifier mais l’ensemble des informations exposées par le endpoint SPARQL du Sandre sont récupérées.</p>



## 6.2. Questions/Réponses session Métadonnées satellites au CERSAT

<p>ENVRI-FAIR et l'Europe mettent en avant la norme DCAT-AP pour les catalogues de données (DCAT Application profile for data portals in Europe). Pourra-t-on passer de ISO19115 à DCAT-AP sans perdre d'information ?</p>	<p>L'ambition du catalogue ENVRI est de type méta-catalogue pour l'ensemble des domaines et dans le but aussi de servir l'EOSC, sachant qu'au niveau de l'EOSC c'est DCAT. Uniquement démonstration du catalogue ENVRI dans le cadre du projet ENVRI-FAIR.</p> <p>On peut passer de ISO19115 à DCAT-AP mais DCAT-AP ne reprend pas l'ensemble d'ISO19115</p> <p>Ce qui est plus adapté à nos données est GEO DCAT-AP</p> <p>Dans le cadre du projet EOSC-Pillar des tests de génération GEO DCAT-AP à partir de Sextant sont prévus</p> <p>Thierry Carval:</p> <p>DCAT-AP:</p> <p><a href="https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/11">https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/11</a></p>
<p>Geonetwork permet une sortie DCAT : <a href="https://geonetwork-opensource.org/manuals/4.0.x/en/api/rdf-dcat.html">https://geonetwork-opensource.org/manuals/4.0.x/en/api/rdf-dcat.html</a></p> <p>Avec perte pour l'instant car ce n'est pas le profil DCAT-AP qui est en sortie, mais DCAT v1. Une implémentation a été proposée ici : <a href="https://github.com/SEMICEu/DCAT-AP/issues/65">https://github.com/SEMICEu/DCAT-AP/issues/65</a></p>	<p>Ce n'est pas une question mais une réponse à la question ci-dessus ;-)</p> <p>Exemple avec la fiche de métadonnées SMOS : <a href="https://sextant.ifremer.fr/Donnees/Catalogue#/metadata/0f02fc28-cb86-4c44-89f3-ee7df6177e7b/formatters/xsl-view?root=div&amp;view=advanced&amp;header=false&amp;related=">https://sextant.ifremer.fr/Donnees/Catalogue#/metadata/0f02fc28-cb86-4c44-89f3-ee7df6177e7b/formatters/xsl-view?root=div&amp;view=advanced&amp;header=false&amp;related=</a></p> <p>Sortie DCAT de geonetwork (avec Perte d'informations) : <a href="https://sextant.ifremer.fr/geonetwork/srv/eng/csw-EOSC_PILLAR?service=CSW&amp;request=GetRecordById&amp;version=2.0.2&amp;outputSchema=http://www.w3.org/ns/dcat%23&amp;elementSetName=full&amp;id=0f02fc28-cb86-4c44-89f3-ee7df6177e7b">https://sextant.ifremer.fr/geonetwork/srv/eng/csw-EOSC_PILLAR?service=CSW&amp;request=GetRecordById&amp;version=2.0.2&amp;outputSchema=http://www.w3.org/ns/dcat%23&amp;elementSetName=full&amp;id=0f02fc28-cb86-4c44-89f3-ee7df6177e7b</a></p>
<p>Pour remplir une fiche (ISO 190015-2 via Geonetwork), est-il prévu de pouvoir reprendre la fiche qu'on avait déjà rempli via SEANOE pour obtenir un DOI ?</p>	<p>Geonetwork n'implémente pas l'ISO19115-2 mais la -3.</p> <p>Il me semble qu'il est possible "d'exporter" la fiche de Seanoe pour mettre les informations dans la norme 19115. Sextant (Geonetwork?) permet aussi de demander des DOI.</p>

### 6.3. Questions/Réponses session interopérabilité sémantique dans l'IR DATA TERRA et les pôles de données

<p>Dans SOSA comment "insérer" les informations de versions des données? Plusieurs versions (liées aux algorithmes appliqués) peuvent correspondre à une seule observation. (Dominique B)</p>	<p>Le modèle SOSA permet de décrire la provenance à travers la classe <code>sosa:procedure</code>. On peut aligner l'ontologie de provenance <code>PROV-O</code> sur <code>sosa:procedure</code> pour permettre d'avoir une description plus formelle et plus précise de la provenance</p>
<p>Est-il possible de naviguer selon les "Unités" (mm par ex.) pour afficher toutes les observations (quel que soit le paramètre et la thématique) ce qui pourrait permettre de "décloisonner" les thématiques. Plus généralement y a-t-il de l'inférence de faite pour lier les données entre elles?</p>	<p>Il ne nous semblait pas pertinent d'entrer dans les données par les unités. Par contre, le modèle de métadonnées Sosa est étendu pour permettre la description des unités</p>
<p>Quelle est la différence entre un thesaurus et une taxonomie ?</p> <p>Merci. Je pense alors que WoRMS est une taxonomie et non un thesaurus (TAXREF référentiel taxonomique français, WoRMS référentiel taxonomique international/mondial)</p>	<p>Une taxonomie est une ontologie particulière dont les concepts sont liés uniquement à travers des relations de hiérarchie</p> <p>Gilbert Maudire:</p> <p>Pour information TAXREF a été couplé à WORMS via le SANDRE dans le cadre des travaux SAR/SIMM (<a href="https://sar.milieufrance.fr/Nos-rubriques/Referentiels-alpha-numeriques/Appellation-de-taxons">https://sar.milieufrance.fr/Nos-rubriques/Referentiels-alpha-numeriques/Appellation-de-taxons</a>)</p>





### 6.4. Exemples de fiches (SMOS & ARGO)

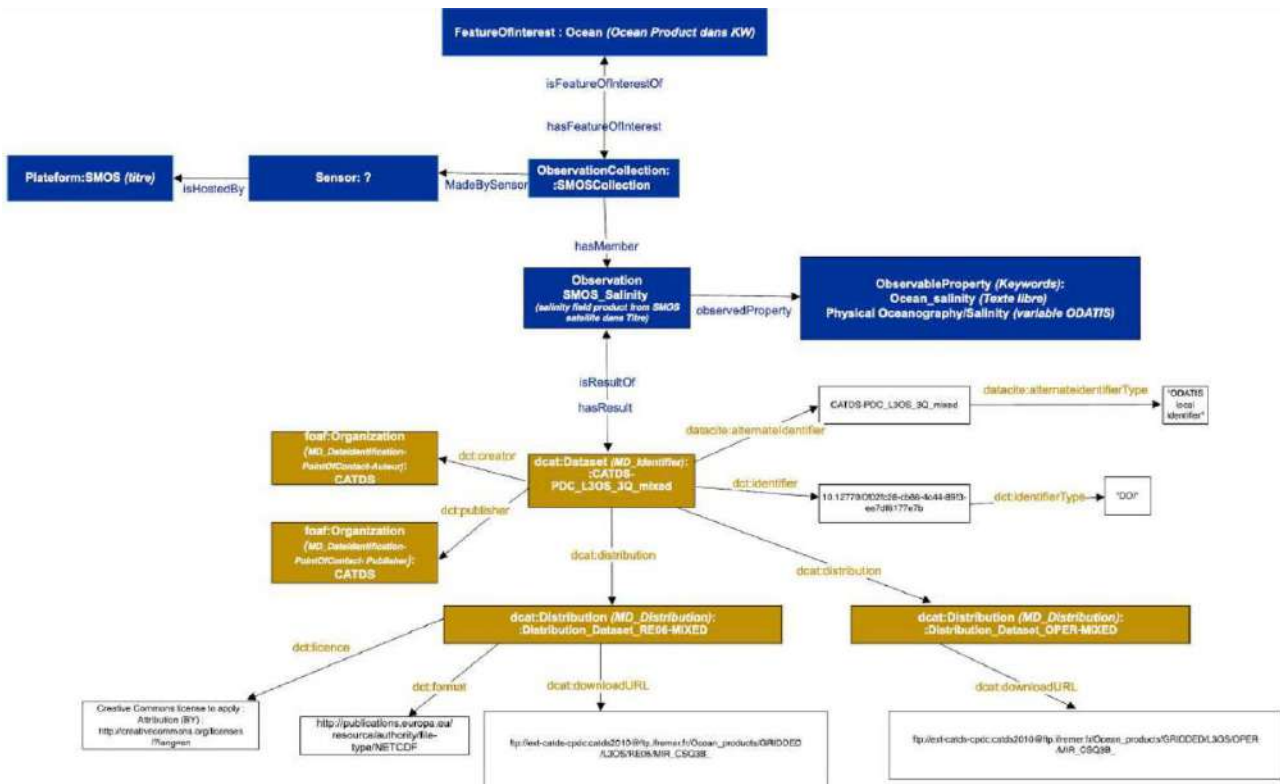


Figure 8. Diagramme UML de la fiche du catalogue ODATIS du produit satellite SMOS sur le modèle DATA TERRA FAIR. Les parties en bleues concernent le modèle SOSA et les parties en jaune le modèle DCAT.

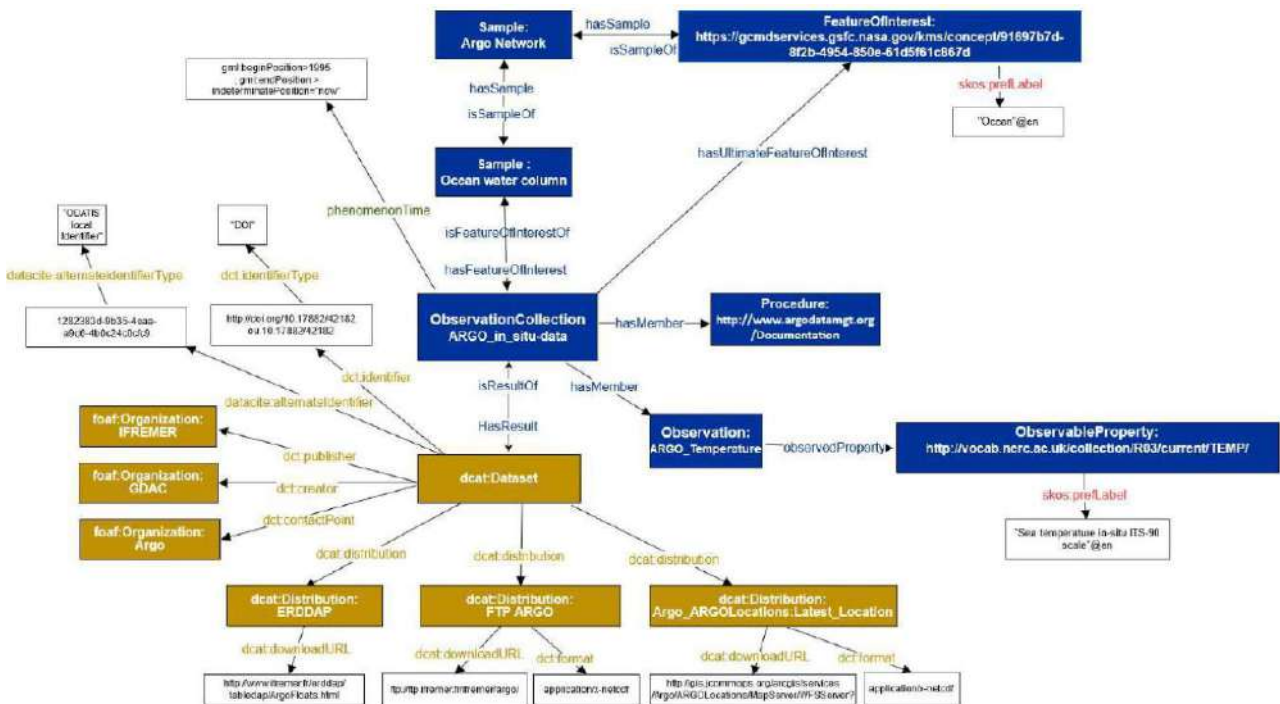


Figure 9. Diagramme UML de la fiche du catalogue ODATIS du produit in-situ ARGO sur le modèle DATA TERRA FAIR. Les parties en bleues concernent le modèle SOSA et les parties en jaune le modèle DCAT





<p>Ce qui aurait été bien c'est que SSS sea surface salinity soit déjà un mot clé dans la fiche geonetwork ?</p> <p>ok je connais mal skos</p>	<p>C'est une possibilité mais on peut aussi grâce a skos:broader ou skos:narrower faire le lien entre le mot clé existant dans le vocabulaire variables ODATIS, et le concept Sea Surface Salinity de CF ou de EOV</p> <p>Lien vers SKOS : <a href="https://www.w3.org/2009/08/skos-reference/skos.html">https://www.w3.org/2009/08/skos-reference/skos.html</a></p>
<p>Du coup, le vrai nom de la variable dans le produit (mean_sea_surface_salinity) disparaît-il? Comment l'expert qui est très familiarisé avec ce nom va-t-il s'y retrouver?</p>	<p>C'est une très bonne question ! Nous n'avons pas trouvé de thesaurus associé à cette variable mais s'il existe nous sommes preneurs d'information.</p>
<p>Dans le produit original, il n'y a pas que mean_sss, est-ce que tous les autres paramètres ou une partie vont être transformés dans ce modèle?</p>	<p>La fiche de MD étudiée ne comportait que mean_sea_surface_salinity comme paramètre (<a href="https://www.odatis-ocean.fr/donnees-et-services/acces-aux-donnees/catalogue-complet#/metadata/0f02fc28-cb86-4c44-89f3-ee7df6177e7b">https://www.odatis-ocean.fr/donnees-et-services/acces-aux-donnees/catalogue-complet#/metadata/0f02fc28-cb86-4c44-89f3-ee7df6177e7b</a>). Mais effectivement, il faudrait trouver un moyen pour que les propriétés observées du modèle Data Terra, soit au plus près de la réalité.</p>
<p>Les URI sont-ils des liens permanents ou peuvent-ils disparaître comme des url ?</p>	<p>Idéalement, un URI devrait être porté par un PID (Thierry)</p> <p>Le PID (Persistant IDentifier) est un identifiant persistant fourni par un fournisseur "indépendant", typiquement une fondation (orcid, DOI, ...)</p> <p>Thierry Carval</p> <p>Idéalement, un URI doit être fournit par un PID (Persistant Identifier) <a href="https://www.openaire.eu/what-is-a-persistent-identifier">https://www.openaire.eu/what-is-a-persistent-identifier</a></p> <p>Jean-Christophe Desconnets (IRD, ESPACE)</p> <p>Ce n'est pas une pratique répandue aujourd'hui, je pense</p> <p>Thierry Carval</p> <p>DATACITE, ORCID utilisent des PID</p> <p>Soudarin Laurent</p> <p>Où créer un PID pour un service desk?</p> <p>Thierry Carval</p> <p>Depuis peu, le BODC expose ses vocabulaires avec</p>





VocPrez <http://vocab.nerc.ac.uk/collection/>

Un fournisseur de PID utilisé par ICOS-RI :  
<https://www.handle.net/>

Cécile Nys

Un projet européen (FREYA - <https://www.project-freya.eu/en/about/mission>) a étudié très largement les PID et a constitué une communauté (PID Forum - <https://www.pidforum.org/>)

