Compte rendu de l'atelier technique ODATIS du 8 et 9 octobre 2019

CR atelier technique octobre 2019

contact@odatis-ocean.fr
www.odatis-ocean.fr

Titre court
CR atelier technique Octobre 2019
Titre long
Compte rendu de l'atelier technique ODATIS du 8 et 9 octobre 2019
Auteur
Joël Sudre

Dissémination	Copyright
Publique via le site www.odatis-ocean.fr	Pôle de Données Océan – ODATIS

Historique

Version	Auteurs	Date	Commentaires
0.1	Joël SUDRE	25 octobre 2019	Version initiale
0.2	Cécile NYS	31 octobre 2019	Relecture et corrections
0.3	Joël SUDRE & Cécile NYS	06 février 2020	Insertions des recommandations finales

Table des matières

1. Accueil et tour de table des participants 4
2. Point d'avancement - (Gilbert Maudire) 4
2.1. Retours du Comité de Direction (CODIR) 4
2.1.1. Sur le Conseil Scientifique (CS) 4
2.1.2. Sur la Convention de Pôle (CP) 5
2.1.3. Sur la demande de poste « Coordination des CDS pilotés par le CNRS » 5
2.1.4. Sur le périmètre des données gérées 5
2.1.5. Sur la demande d'adhésion de l'Agence Française pour la Biodiversité (AFB) 5
2.2. Retours et décisions du Bureau Exécutif (BE) 6
2.2.1. Sur l'amélioration du site internet 6
2.2.2. Sur les noms des CDS 6
2.2.3. Sur les interfaces avec les IR d'observation 7
2.3. Projets menés par l'IR DT et le Pôle 8
2.3.1. Projets en cours 8
2.3.2. Nouveaux projets (automne 2019) 8
2.4. Retours sur les Consortiums d'Expertise Scientifique (CES) 9
2.4.1. CES Oxygène (coord : V. Garçon, J. Sudre, et S. Schmidt) 9
2.4.2. Atelier couleur de l'eau 9
3. Traitement des données ARGO avec Dask et Parquet -(Frédéric Briol) 10
4. Référentiels (PXX) de SeaDataNet (SDN) – (Michèle Fichaut) 11
4.1. Présentation SDN et outils associés 11
4.2. Outils SDN et formats de donnée SDN 12
5. Netcdf, convention Climate and Forecast (CF), fichier ASCII et métadonnées – (Joël Sudre) 13
6. Système d'Information du Milieu Marin (SIMM) du MTES – (Clémence Rabevolo et Steven Piel) 14
7. Recommandations ODATIS pour les CDS 17
7.1. Format des métadonnées [Obligatoire]: 17
7.2. Attributs de paramètres [Fortement recommandée] 17
7.2.1. Quelle convention pour quelle discipline ? [Obligatoire]: 17
7.2.2. Convention paramètres et données de SeaDataNet (SDN) [Obligatoire]: 18
7.2.3. Convention paramètres et données du NetCDF-CF [Fortement recommandée] 18
7.3. Paramètres de positionnement [Obligatoire] : 19

1. Accueil et tour de table des participants

Liste des participants à l'atelier ODATIS :

- Frédéric Briol (*CLS*) – FB,
- Gérald Dibarboure (*CNES*) – GD,
- Wendy Diruit (IMEV - CDD ODATIS) – WD,
- Michèle Fichaut (*IFREMER*) – MF,
- Valérie Hascoat (*IFREMER*) – VH,
- Mark Hoebeke (Station Biologique de Roscoff) – MH,
- Dimitry Khvorostyanov (*LOCEAN*) – DK,
- Stevenn Lamarche (Univ. Brest) – SL,
- Maurice Libes (*OSU Pytheas*) – ML,
- Céline Laus Heyndrickx (DT-INSU) – CLH,
- Didier Mallarino (*OSU Pytheas*) – DM,
- Gilbert Maudire (*IFREMER*) – GM,
- Cécile Nys (*IFREMER*) – CN,
- Steven Piel (*AFB*) – SP,
- Julien Penguen (*OASU – EPOC*) – JP,
- Clémence Rabévolo (*IFREMER*) – CR,
- Catherine Schmechtig (IMEV) – CS,
- Sabine Schmidt (Univ. Bordeaux / *EPOC*) – SS,
- Joël Sudre (*OMP*/*LEGOS*) – JS.

JS présente l'ordre du jour (voir : Agenda et accès aux présentations), en précisant que le compte rendu de l'atelier précédent n'est pas encore disponible et qu'il est en cours de relecture et de finalisation. Ce CR ayant fait remonter les besoins des CDS, il a été noté qu'une demande forte des CDS est que l'atelier technique d'ODATIS apporte des solutions techniques, des mises en pratique (avec une prise en main des solutions) et des recommandations. Cet atelier technique se présente sous un nouveau format en prenant en compte ces besoins.

2. Point d'avancement – (Gilbert Maudire)

GM présente (voir 201910_ODATIS_Atelier_Gmaudire_retour_codir.pdf) un résumé des événements qui ont marqué le pôle ODATIS et l'IR Data Terra depuis l'atelier technique de juin 2019 afin d'informer l'ensemble des CDS.

2.1. Retours du Comité de Direction (CODIR)

2.1.1. Sur le Conseil Scientifique (CS)

Actuellement le CS est composé de 20 représentants. Ce nombre important de représentants rend complexe la prise de décision du CS et la recherche de solution adéquate pour le Pôle. De plus l'IR Data Terra (IRDT) envisage de mettre en place aussi CS. La question se pose donc de faire vivre deux

CS en parallèle (celui du Pôle et celui de l'IRDT). Le CODIR a donc demandé au Bureau Executif Restreint (BER) de faire une proposition d'évolution du CS du Pôle et de conforter les liens avec l'observation (IR ILICO, IR Flotte, etc.).

2.1.2. Sur la Convention de Pôle (CP)

Le service juridique de l'Ifremer a demandé un nouveau tour de table avant la signature de la CP. Ce tour de table se justifie par la non conformité présumée avec la loi actuelle du droit sur les données et l'interprétation qui en est faite sur l'« Open data » (respect du protocole de Nagoya et du protocole sur les ressources génétiques). Il est donc nécessaire avant la signature de la CP de mettre toutes les exceptions dans la convention et en particulier sur les aspects internationaux. Il est aussi nécessaire de citer dans cette convention les articles de loi faisant référence aux zones sensibles, la protection des données des pays tiers, des ressources génétiques, l'embargo sur les données, ainsi que la possibilité pour les navires français de faire des levées dans des zones maritimes non française.

Les correspondants juridiques de chaque partenaire du Pôle ODATIS doivent donc se mettre d'accord pour harmoniser leurs politiques de gestion des données avant la fin de l'année 2019 (date butoir décidée par le CODIR à laquelle la convention de pôle doit être signée).

2.1.3. Sur la demande de poste « Coordination des CDS pilotés par le CNRS »

Une demande de « Nouveaux Emplois offerts à la Mobilité Interne » (NOEMI) a été déposée en septembre pour l'UMS 2013 (UMS CPST Coordination des Pôles du Système Terre de l'IRDT) afin de mettre en place un poste de coordinateur des CDS CNRS pour le Pôle ODATIS.

2.1.4. Sur le périmètre des données gérées

Le CODIR a demandé que le périmètre des données gérées soient mieux précisé avec des liens avec les différentes IR participantes à la collecte de ces données. En particulier il est nécessaire de préciser le périmètre sur :

- les aspects biologiques et génomiques,
- la méta-génomique (incluse ou non),
- la désagrégation des jeux de données entre les pôles.

Pour répondre aux questions sur la génomique et la méta-génomique, un CES méta-génomique est envisagé pour réunir les différents partenaires impliqués dans cette thématique afin de préciser le périmètre de ces données.

2.1.5. Sur la demande d'adhésion de l'Agence Française pour la Biodiversité (AFB)

L'Agence Française pour la Biodiversité (AFB) produit des données dans les parcs marins. Elle apporte aussi une aide financière à certaines campagnes à la mer et est une interface avec le Ministère de la Transition Ecologique et Solidaire (MTES). Dans ce cadre, l'AFB a fait une demande d'adhésion comme nouveau partenaire du Pôle ODATIS.

Le CODIR a émis un accord de principe sur ce nouveau partenariat, sous réserve d'acceptation de la part de l'AFB :

- de la convention du pôle,
- de la participation aux instances et à la vie du Pôle dans son ensemble.

A noter qu'il a été demandé à l'AFB d'accepter cet accord avant la mise en place de l'Office Français de la Biodiversité (OFB) qui reprend les missions de l'AFB et de l'Office National de la Chasse et de la Faune Sauvage (ONCFS).

2.2. Retours et décisions du Bureau Exécutif (BE)

2.2.1. Sur l'amélioration du site internet

Le BE a demandé de modifier le site internet du Pôle ODATIS en particulier sur la page d'accueil afin qu'elle soit plus homogène avec les autres pôles (Aeris en particulier). Ces modifications ont déjà été apportées à la page d'accueil par l'ajout de nouveaux boutons permettant l'accès direct aux produits et jeux de données, aux animations techniques et aux animations scientifiques du Pôle. L'agenda comporte maintenant plus d'entrées et une navigation chronologique. Enfin des liens vers les sites des autres pôles et l'IRDT ont été rajoutés sur la page d'accueil via les logos de ces sites (voir page d'accueil du Pôle ODATIS).

2.2.2. Sur les noms des CDS

Le BE a demandé de reprendre pour l'ensemble des CDS la nomenclature et la description de ces centres en utilisant les noms historiques car ils sont déjà connus par la communauté scientifique. Il a donc été convenu que les noms des CDS soient ceux de la Figure 1.

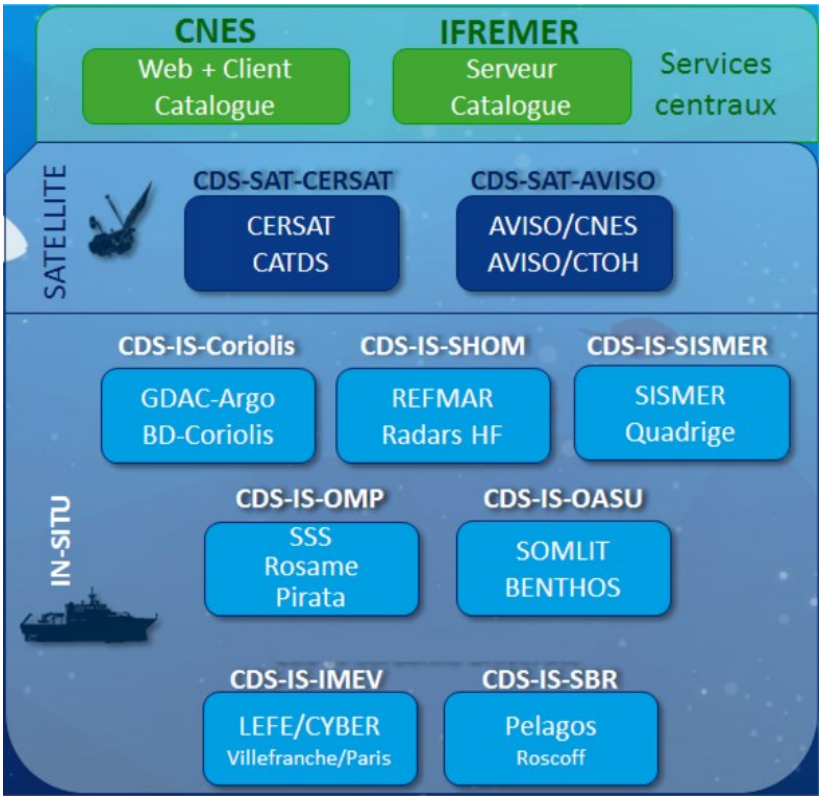

Figure 1. Nomenclatures des CDS du Pôle ODATIS

2.2.3. Sur les interfaces avec les IR d'observation

Afin de trouver un fonctionnement optimal, il est nécessaire pour le pôle ODATIS de travailler en interaction étroite avec les différentes infrastructures de recherche impliquées dans l'acquisition et la création de données ayant attrait à l'Océan. Un exemple de démarche avec l'IR ILICO est donnée par la Figure. 2.

Figure 2. Interaction avec l'IR ILICO

Afin de généraliser et de valider cette démarche d'interaction avec une IR d'observation, il est nécessaire de préciser le rôle du Conseil Scientifique et de prendre en compte les initiatives inter-organismes existantes (ex : Resomar, Coriolis, etc.). Il est aussi important de bien définir le mandat et le périmètre thématique des CDS.

2.3. Projets menés par l'IR DT et le Pôle

Les projets auxquels participe le pôle ODATIS sont répertoriés ci-dessous avec la distinction des projets qui sont déjà en cours actuellement et les projets futurs.

2.3.1. Projets en cours

DG-Recherche : **H2020 SeaDataCloud**

L'ensemble des CDS IS (In Situ) d'ODATIS est impliqué dans ce projet. Ce projet est une source de financement en France supérieur à 500k€ (hors coordination). Cependant, la proposition SeaDataCloud 2 n'a pas été retenue, ce qui est décevant car la France a toujours été bien positionnée et représentée lors des propositions précédentes (SeaDataNet, …).

DG-Recherche : **H2020 Pillar**

Ce projet est le support de l'implémentation d'EOSC, en fédérant les initiatives nationales (les Pôles) et/ou thématiques. Il est basé sur les concepts de l'« Open Science » et du « FAIR data ». Il va permettre de faciliter l'adoption des principes d'EOSC, par des allers retours avec les initiatives existantes.

2.3.2. Nouveaux projets (automne 2019)

ANR Flash « Données Ouvertes » : **Copilote** pour Odatis

Ce nouveau projet a pour objectif d'être un support à l'adoption des principes FAIR dans les CDS et de permettre au CDS qui ont répondu présent lors de sa rédaction de ce projet d'aller vers une certification « Core Trust Seal (CTS)» de la Research Data Alliance (RDA). Les CDS-IS ayant répondu à ce projet sont Coriolis, SISMER, IMEV et SBR (voir Figure 1). A noter que l'Ifremer avec le CDS-IS-Coriolis et le CDS-IS-SISMER ont déjà entrepris une démarche de certification CTS auprès de la RDA. Cette certification va servir d'exemple pour les autres CDS.

DG-Recherche : **H2020 Blue Cloud**

Ce projet est le volet "marin" d'EOSC. Il est constitué de cas d'étude permettant la définition des services du Pôle. Il est à noter que ce projet a un périmètre plus large qu'Odatis car il inclus aussi l'environnement, les pêches, l'aquacullture, la biodiversité (génomique), etc.

DG-Connect : **Phidias**

Ce projet rassemble des études techniques autour de l'implémentation de l'EOSC pour le Pôle ODATIS et AERIS en particulier. Elles sont basées sur des cas d'usages thématiques (Océan & Atmosphère en particulier). Ces études sont en lien avec l'étude menée par le CNES sur la structuration des données (parallélisation, stockage objet…). Ce projet constitue également une base de réflexion pour nos infrastructures d'organismes Datarmor à l'Ifremer, « DataLake » au CNES et la politique INFRANUM nationale.

2.4. Retours sur les Consortiums d'Expertise Scientifique (CES)

2.4.1. CES Oxygène (coordination : V. Garçon, J. Sudre, et S. Schmidt)

Ce CES Oxygène dissous a été mis en place début 2019 afin de mettre en réseau et fédérer les acteurs scientifiques au niveau national, voire international, autour de la thématique de désoxygénation de l'océan hauturier et côtier et d'établir une base exhaustive et qualifiée des données nationales d'oxygène dissous océanique.

Un premier atelier du CES Oxygène s'est tenu les 2 et 3 Juillet 2019 à Paris. Il est à noter que cette atelier est une initiative du pôle ODATIS mais qu'il est commun avec LEFE-CYBER qui a participé financièrement à la mise en place de ce premier séminaire. Le compte rendu est disponible en suivant le lien suivant : CR CES Oxygène.

2.4.2. Atelier couleur de l'eau

Les 28 et 29 mai 2019 s'est tenu un atelier sur la couleur de l'eau avec pour objectif de réunir les communautés qui travaillent autour du thème Couleur de l'eau (traitement image et données in-situ) afin :

- d'échanger autour des expertises nationales,
- d'initier une dynamique nationales,
- de contribuer à la mise en place d'un CES Couleur de l'eau.

La première journée a été consacrée à l'inventaire des usages de chaque communauté (académiques ; bureaux d'étude), des liens à développer avec ODATIS et des besoins de la communauté (SNO, campagnes à la mer, sites instrumentés). Quant au second jour, il a eu pour but de définir les actions possibles dans le cadre d'Odatis, la typologie des produits, les priorités et les échéances. Elle a aussi permis de définir les CES relatifs à la Couleur de l'eau.

Lors de cet atelier plusieurs thèmes de réflexion en relation avec la couleur de l'eau ont émergés, en particulier :

- pertinence de la combinaison données in situ/satellite pour l'évaluation des eaux côtières sur le long terme,
- apport de la haute résolution à l'évaluation de la représentativité spatiale (les SNO ont besoin de quelle résolution ?),
- spatiale de l'observation des écosystèmes, un réseau a en général un nombre limité de site par zone),
- bathymétrie littorale : comparaison des techniques/approches sur des sites communs (LEGOS / EPOC),
- d'autres pistes : salinité, assistance aux campagnes en mer (SPASSO, relais vers ODATIS ?), groupes,
- phytoplanctoniques (PFT, algues toxiques), fluorescence.

3. Traitement des données ARGO avec Dask et Parquet -(Frédéric Briol)

FB présente la pile logiciel PANGEO ainsi que l'application sur les données ARGO avec Dask et Parquet (voir 201910_ODATIS_Atelier_FBriol_Argo.pdf).

PANGEO est un écosystème de librairie écrite en python permettant de manipuler l'ensemble des données satellitaire et in-situ. A ce titre, il est un outil qui peut permettre de traiter aussi bien sur un PC standard que sur un HPC (avec des calculs parallèles massifs) ou un cloud les données avec les mêmes scripts (une adaptation simple du script (via Dask) étant à insérer en début de code pour préciser sur quel moyen de calcul, il est lancé). Les codes sont développés et exécutés via jupyterhub, qui est une interface graphique permettant de déporter l'architecture sur du HPC, du cloud, avec un environnement propre à l'utilisateur. Cela permet donc de travailler simplement sur des machines distantes (massivement parallèle ou pas), sans une trop grande contrainte de modification du code. Cette architecture est testée :

- sur HAL jupyterhub via ressource CNES. Pour le moment il est nécessaire de faire une demande d'ouverture de compte au CNES (voir pour cela avec GD). Une forte contrainte se situe actuellement au niveau des données à disposition car il est nécessaire de demander le téléchargement des données extérieurs (nasa par ex.). L'ouverture vers l'extérieur est envisagée avec la mise en place de contraintes moins importantes.
- sur Datarmor jupyterhub est en place aussi avec des comptes locaux. L'Ifremer a la volonté de mettre en place une identification commune pour avoir accès à l'HPC via la fédération d'identité recherche france (RENATER). Actuellement la pile PANGEO est en phase de test sur Datarmor.

Une VM a été mis à disposition pour l'atelier technique ODATIS pour commencer à utiliser cet outil. A noter que Binder, qui permet de créer des environnements informatiques spécialisés via binderhub open-source (https://binder.pangeo.io/v2/gh/fbriol/pangeo-argo/master), est aussi utilisable avec jupyter (sur le cloud google).

FB a ensuite fait un rappel sur la librairie Parquet et Arrow pour introduire les données Argo qui sont des données tabulaires distribuées en netcdf sous forme de matrice régulière (bcp de matrices creuses…). Parquet permet de mettre des données tabulaires sur disque quant à Arrow, il permet de filtrer et de hiérarchiser de manière très efficace des données provenant des fichiers Parquet.

La principale difficulté pour la conversion des données NetCDF vers les « DataFrame » utilisables avec Parquet/Arrow consiste à créer le schéma de la table (voir p 17 de la présentation).

Sur le jeu de donnée ARGO PR/PF de 1995 à juin 2018, qui représente un volume de 5,6 Go (sur le « DataFrame ») et 94 Go au format NetCDF, les performances sont très correctes avec par exemple une sélection sur une région géographique pour une année en 16 sec ou sur l'intégralité du jeu en 1 min 37 sec (voir p 20 de la présentation pour d'autres exemples) .

A la suite de la présentation FB, a distribué une VM permettant d'utiliser la pile Pangeo. **Cette VM est à conserver sur les postes pour effectuer d'autres travaux pratiques au cours des prochains**

ateliers afin de familiariser les utilisateurs à cet outil. Pour le prochain atelier, il est fortement conseillé de faire quelques tests avec cette VM. A noter qu'il est possible de demander à GD et FB de mettre en place vos propres exemples afin de faire d'autres travaux pratiques appliqués à votre usage et pouvant servir à d'autres participants.

4. Référentiels (PXX) de SeaDataNet (SDN) – (Michèle Fichaut)

4.1. Présentation SDN et outils associés

Suite à l'atelier technique ODATIS de juin 2019, il est apparu urgent que le pôle clarifie sa position sur l'utilisation des référentiels et du vocabulaire associé pour l'ensemble de ses CDS. MF présente les référentiels de SeaDataNet (SDN) pour les paramètres (voir 201910_ODATIS_Atelier_Mfichaut_ref_Seadatanet.pdf).

Pour la description des paramètres mesurés, il est nécessaire d'utiliser plusieurs référentiels SDN avec leurs vocabulaires associés :

- le P08 pour la discipline du paramètre (12 Sélections Possibles),
- le P03 pour la catégorie du paramètre (76 Sélections Possibles),
- le P02 pour faciliter la découverte de paramètres (462 Sélections Possibles)
- le P01 qui rassemble le vocabulaire pour chaque paramètres (42911 Sélections Possibles)

Pour un paramètre donné, il est nécessaire de trouver sa discipline (P08) puis sa catégorie (P03), son paramètre de découverte (P02) et enfin son vocabulaire propre (P01).

Un paramètre est défini par un vocabulaire contrôlé et par un modèle sémantique (définis dans SDN depuis 2004). Ce modèle et ce vocabulaire ont été adoptés par plusieurs projets (nationaux et européens) pour faciliter l'échange de donnée et de créer de l'interopérabilité entres les différents projets. Il est nécessaire de bien comprendre le modèle sémantique adopté par SDN (provenant du British Oceanographic Data Centre -BODC) ainsi que ses composants, sa structure et sa logique pour arriver à trouver aisément le vocabulaire associé aux paramètres recherchés. Le modèle sémantique pour le P01 est constitué de 3 éléments principaux :

- une propriété quantitative ou nominative d'une entité (concentration, abondance,…) (« property »),
- une entité physique, biologique ou chimique (« object of interest »),
- une entité environnementale à laquelle l'entité physique, bilologique ou chimique se rapporte ou dans laquelle elle est intégrée (colonne d'eau, sédiment,…)(« Matrix »),
- des champs optionels (« statistical qualifier, sample preparation, analytical method, processing method »).

Le nom du paramètre est structuré de la façon suivante :

<Property> {statistical qualifier} <object of interest> <Measurement to Matrix relationship> <Matrix> {optional fields}

ex : <Concentration>{standard deviation} of <ammonium{NH4 + CAS 14798-03-9}><per unit of mass of the> <water body> [dissolved plus reactive particulate].

Pour chaque élément constitutif du modèle sémantique un vocabulaire est associé et il est disponible sur le site de SDN (voir slides 11 et 12 de la présentation).

Voici quelques questions clés qui vous permettrons de trouver plus facilement votre paramètre :

- quels sont les éléments essentiels du paramètre?
- est-ce que c'est la propriété d'une substance chimique? D'une entité biologique? D'une entité physique?
- quelle quantité ou propriété est mesurée/reportée?
- dans quel environnement ou substrat est faite la mesure?
- quelle est l'unité?
- comment est faite la mesure?
- est-ce que l'échantillon a été filtré? Si oui, quelle était la taille des pores ou le type de filtre ?
- le paramètre doit-il être réparti en classes comme par exemple les classes de taille des particules ?
- Il existe plusieurs solutions pour trouver son paramètre :
- recherche hiérarchique (site SDN) qui est idéale pour un utilisateur inexpérimenté,
- recherche avec facettes de recherche (site SDN) pour un utilisateur peu familiarisé,
- recherche par mot-clef (site BODC) idéal pour les utilisateurs expérimentés (permet aussi de soumettre un nouveau code P01)

Il est aussi possible de visualiser le contenu des vocabulaires au travers du site web vocab.nerc.ac.uk pour le P02 par exemple, pour changer de référentiels il suffit de changer le P02 de ce lien http://vocab.nerc.ac.uk/collection/P02/current/.

Librairie Java 1.8 est un logiciel développée par l'IFREMER qui permet de stocker localement la liste des vocabulaires et de mettre à jour ces listes sur demande.

Pour toutes questions sur le vocabulaire il est possible d'utiliser les helpdesk suivant :

- BODC : vocab.services@bodc.ac.uk
- SDN : sdn-userdesk@seadatanet.org.

A la suite de cette présentation MF a fait faire quelques exercices pratiques aux participants de l'atelier pour qu'il se familiarise avec les outils de recherche. Une discussion sur les recommandations à donner par le pôle s'en est suivi. Les recommandations se trouve à la fin de ce document afin de les trouver plus facilement.

4.2. Outils SDN et formats de donnée SDN

SDN a défini trois formats majeurs pour gérer des données océanographiques (profils verticaux, séries temporelles, trajectoires,…). Ces formats contiennent des informations spécifiques pour SDN (LOCAL_CDI_ID, EDMO_Code, le vocabulaire SDN (P01,P06 pour les unités, L22 pour les capteurs), possibilité de faire rérérence aux catalogues SDN des campagnes.

Il existe deux formats ASCII : ODV SeaDataNet et MedAtlas SeaDataNet (MedAtlas est obsolète et ne doit plus être utilisé), un format binaire en NetCDF avec la convention SDN qui est compatible avec la convention Climate and Forecast (CF).

Il existe 3 variantes du format ASCII ODV pour gérer les données de biologie, de microplastiques et de cytométrie en flux et une variante en NetCDF pour gérer les données de Radar HF. Ces différents formats sont accessible en suivant ce lien.

MF présente ensuite les outils pour gérer ces formats :

- Nemo qui est l'outil de formatage SeaDataNet pour générer des fichers au formats SDN à partir de fichiers en entrée ASCII,
- Octopus qui est convertisseur pour générer des formats SDN à partir d'un format SDN,
- Mikado qui permet de générer des descriptions des données au format XML ISO-19139.

Afin d'avoir plus de détails sur ces différents outils et la manière de s'en servir le lecteur est encourager à télécharger la présentation suivante : 201910_ODATIS_Atelier_MFichaut_outils_Seadatanet.pdf

5. Netcdf, convention Climate and Forecast (CF), fichier ASCII et métadonnées – (Joël Sudre)

JS présente une introduction au Network Common Data Format (NetCDF - voir 201910_ODATIS_Atelier_Jsudre_netcdf.pdf) qui est issue d'une Action National de Formation (ANF) proposée en 2017 et 2018 par le réseau « Séries Interopérables et Systèmes de traitement des données » (SIST) qui est le réseau technologique des informaticiens gestionnaires de données d'observation du CNRS.

Le format NetCDF a été créé par Unidata (https://www.unidata.ucar.edu/) en 1988 afin d'avoir un format portable et des données indépendantes de la machine, une bibliothèques de procédures libre et gratuite écrite dans de nombreux langages. Ce format a été recommandé par l'Inter-pôles comme un format d'échange standard (avec aussi le format ASCII Tabular Separated Variable).

Le format NetCDF est un format riche qui embarque données et métadonnées dans un même fichier. C'est un format flexible, normalisé, bien adapté au stockage de tableaux de nombres multidimensionnels, permettant la mise en place de convention. Il a aussi l'avantage de pouvoir être interopérable et il accepte la standardisation.

La présentation permet de découvrir :

- le modèle de données classique (NetCDF 3),
- le modèle de données NetCDF 64-bit Offset Formation,
- le modèle de données NetCDF 4 (qui est le format actuellement recommandé).

Cette présentation aborde aussi les conventions associées, et en particulier la convention Climate and Forecast (CF) très largement répandue dans les communautés Océanographique et Atmosphérique. La dernière partie de la présentation permet de découvrir des outils gratuits et

facilement installable sur toutes les plateformes permettant de lire, écrire, modifier simplement des fichiers NetCDF.

6. Système d'Information du Milieu Marin (SIMM) du MTES – (Clémence Rabevolo et Steven Piel)

CR et SP présente les Système d'Information fédérateurs, le Système d'Information du Milieu Marin (SIMM) ainsi que le Service d'Administration des Référentiels (SAR).

Suite à la loi Biodiversité, le décret de création de l'Agence Française pour le Biodivertité (AFB) stipule (Art. R. 131-34 du code de l'environnement) que l'AFB doit assurer l'animation et la coordination technique :

- du SI sur l'Eau (SIE), les milieux aquatiques et les services publics d'eau,
- du SI sur la biodiversité, dont le SI Nature et Paysage (SINP),
- du SI sur le Milieu Marin (SIMM).

Ces 3 SI ont été créés afin de répondre aux politiques publiques (DCSMM, DCE, Halieutique,...). Les enjeux stratégiques de ces SI fédérateurs sont aux niveaux fonctionnels :

- de tendre à l'open data sur les données avec une compatibilité INSPIRE,
- de faciliter les multiples usages de la donnée « primaire »,
- de répondre aux besoins, pilotages et projets locaux,
- de rationaliser les modalités de rapportage européen et national.

Du point de vue structurel, ces SI doivent permettre de :

- mettre en place des processus de production de la donnée (protocoles, etc.),
- structurer les données, pour les rendre interopérables,
- fiabiliser et bancariser les données,
- organiser l'ensemble des projets informatiques du milieu marin.

Associé à ces 3 SI, un nombre important de SI métier ont été créés comme le SI des aires marines protégées, le SI DCSMM, SI halieutique, SI récifs et mangroves etc. (voir la présentation pour une liste plus exhaustive). Chaque SI métier est constitué de une ou plusieurs base(s) de données comme par exemple le SI DCSMM avec la base Quadrige^2, la base Bruit, la base mégafaune, la base du SIH. Ces bases étant alimentées par différents partenaires (ex : Quadrige^2 est alimentée à la fois par la DREAL, Pomet, Ifremer, le Cedre, etc.).

Les instances de gouvernance de SIMM sont (voir Figure 3) :

- une instance stratégique assurée par le Comité National des Directives Milieux Marins (CNPDMM),
- une instance technique assurée par le Comité de coordination Technique (CT), chargée de rédiger et valider le schéma national des données sur le milieu marin et des documents juridiques sur le droit des données dans le cadre du SIMM,

- des groupes « spécialisés » (ex : Urbanisation, Langage commun, portail internet milieumarinfrance.fr créé en juillet 2019),
- des groupes de pilotage technique de chacun des projets informatiques.

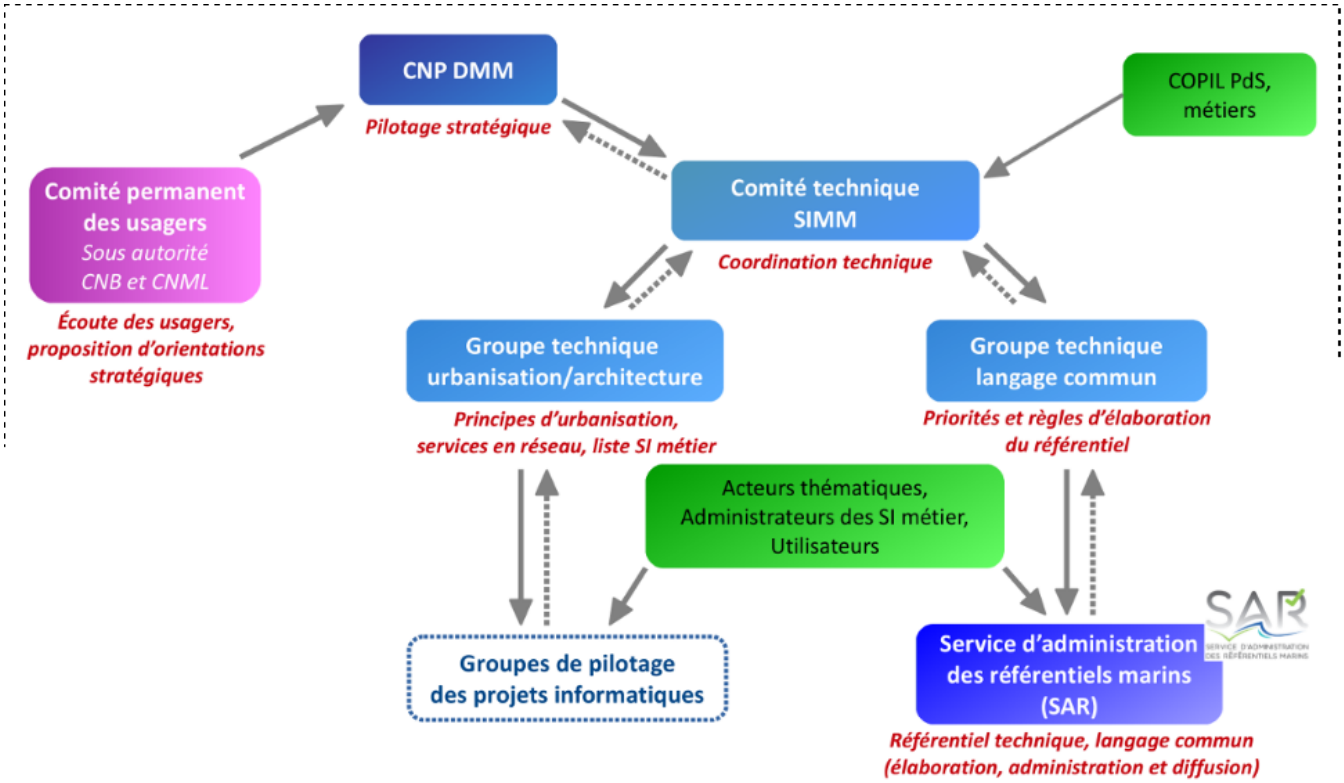

Figure 3. Instances de gouvernance du SIMM

Les principes généraux du SAR sont d'analyser les référentiels existants et de les retenir en état s'ils conviennent ou de les modifier s'ils ne sont pas directement adaptés et de les créer si rien n'existe. La procédure de mise en place d'un nouveau référentiel est résumé dans la Figure 4. Ces référentiels doivent être en conformité avec le cadre commun d'architecture des référentiels de données (publié par le Secrétariat général pour la modernisation de l'action publique le 18 décembre 2013), des normes en vigueurs (ISO, INSPIRE, OCG, etc.) ainsi qu'avec les référentiels européens et internationaux.

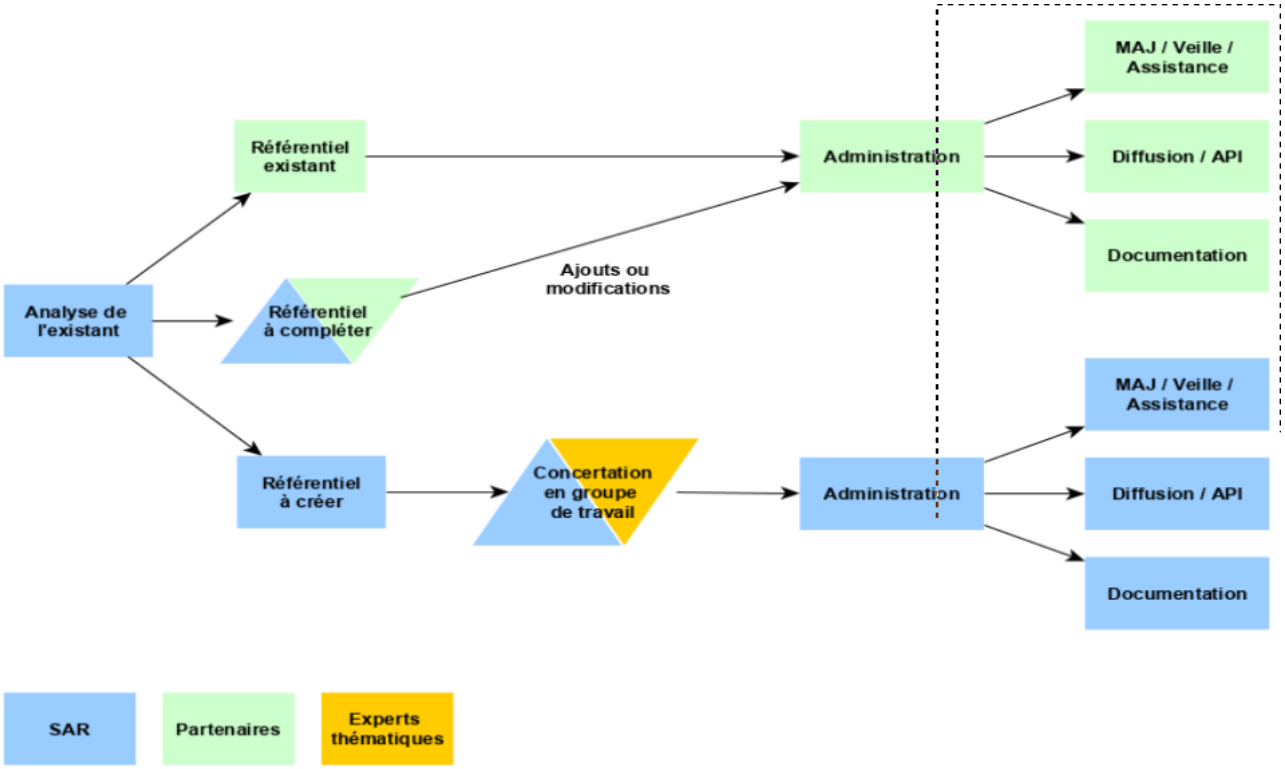

Figure 4. Procédure de mise en place d'un nouveau référentiel

Le secrétariat du SAR a été confié à l'Ifremer. Cette équipe est en charge de rédiger les modèles, de tenir à jour les données de référence et de diffuser l'ensemble aux politiques publiques. La coordination du SAR est assurée par l'Ifremer et l'AFB et s'appuie sur l'ensemble des acteurs du Milieu Marin.

Le Groupe de Pilotage du Langage Commun a approuvé les modèles de données « déchets » et d'appellation taxonomique du Service d'Administration Nationale des Données et Référentiels sur l'Eau (SANDRE). Ces 2 référentiels sont accessibles via la rubrique Diffusion du portail du SAR qui est un lien vers le portail du SANDRE). Plusieurs référentiels prioritaires sont en cours de création comme le référentiel « Interlocuteurs/Organismes », « Qualité de la donnée », « Paramètre » (traduction en français du P01 du BODC), « Ports », « Ouvrages », etc.

D'autres référentiels vont aussi être à traiter comme le référentiel « Habitats », « Activités », « Engins de pêche », « instruments de mesure »,…

Le SAR est en train de mettre en place plusieurs outils :

- un outil de modélisation UML (Modelio/ Enterprise Architect),
- un outil de gestion de registres (référentiels alphanumériques – Ukgovld 2.0),
- un outil de gestion des référentiels cartographiques (catalogue Sextant),
- un outil de gestion des demandes des utilisateurs (Mantis).

Le portail du SAR va permettre la diffusion des référentiels, de s'y abonner et d'avoir accès a un formulaire de demande afin d'enrichir les référentiels (mise en place prévue en 2020).

Le pôle ODATIS et le SIMM doivent travailler ensemble en particulier sur les référentiels « paramètres », « réseaux de mesure », « organismes », « géographiques » afin d'utiliser les mêmes référentiels.

7. Recommandations ODATIS pour les CDS

7.1. Format des métadonnées [Obligatoire]:

Deux types de formats de métadonnées sont à recommander :

- ASCII TSV (ODV speadsheet normalisé SDN)
- NetCDF convention Climate and Forecast (CF) (v1.6 ou supérieure)

Il est recommandé d'utiliser à minima le format ASCII TSV (ODV speadsheet normalisé SDN) et si possible du NetCDF convention CF.

Le NetCDF version 4 (sans les groupes) est le format NetCDF à privilégier (Les anciens fichiers en NetCDF 3 sont à migrer en NetCDF 4 dès que cela est possible). Il est recommandé d'utiliser la version de convention CF 1.6 à minima ou une version supérieure.

Les données in situ sont a minima à mettre dans un fichier ASCII ODV spreadsheet avec l'ensemble des attributs globaux en en-tête de fichier.

7.2. Attributs de paramètres [Fortement recommandée]

Afin d'avoir la liste des attributs globaux et des attributs de paramètres qui est nécessaire a minima de mettre dans un fichier NetCDF, le lecteur peut se référer aux recommandations :

- de la conventions CF (http://cfconventions.org/) et en particulier le chapitre 2 du document http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/cf-conventions.html
- du site SeaDataNet qui donne accès au document décrivant les formats (NetCDF et ODV) : https://archimer.ifremer.fr/doc/00454/56547/ (DOI 10.13155/56547). Sur le site SeaDataNet, il y a aussi des exemples qu'on peut télécharger à la page : https://www.seadatanet.org/Standards/Data-Transport-Formats

7.2.1. Quelle convention pour quelle discipline ? [Obligatoire]:

La convention **« Climate and Forecast » (CF)** ainsi que la convention **SeaDataNet (SDN)** sont les deux conventions à utiliser pour les données océanographiques. Les attributs des deux conventions sont à insérer dans les fichiers NetCDF et ASCII TSV :

- pour les paramètres physiques où la convention CF est largement répandue, l'utilisation de la convention SDN est facultative mais il est fortement recommandé de les insérer en attribut supplémentaire décrivant le paramètre, le « long name » correspondant au paramètre décrit en convention SDN (liste P01 de SDN),
- pour les paramètres de biogéochimie, de chimie, de plastique et micro-plastique, etc. (tous les paramètres non physiques), il est obligatoire d'utiliser la convention SDN (liste P01 de

SDN) et d'insérer en attribut supplémentaire le « long name » en convention CF correspondant au nom du paramètre en convention SDN (liste P07),

- pour les paramètres de biologie, qui sont des cas particuliers, où actuellement deux voire trois standards co-existent (BioODV, Darwin Core, EML). Il est possible d'utiliser la convention:
    - BioODV : promu par SDN, format tabulaire avec fichier unique. Ce format ne permet pas actuellement de décrire avec une finesse identique tous les paramètres d'acquisition ou tous les traits de vie liés aux spécimens observés. De plus, il n'est pas utilisé dans les projets majeurs de publication de données de biodiversité (EMODNet Biology alimente OBIS et donc GBIF),
    - Darwin Core : promu par EMODNet Biology, format tabulaire comprenant 3 fichiers : un pour les caractéristiques de base des événements d'échantillonnage (lieu, date), un pour les caractéristiques de base des spécimens observés (identification du spécimen, nombre d'occurrences), un pour les descripteurs additionnels qu'ils soient liés à l'échantillonnage (caractéristiques des engins de prélèvement) ou aux occurrences (caractéristiques des spécimens mesurés : taille, poids, sexe, stade de développement etc...)
    - EML (environmental Modeling Language) : promu par le PNDB (à vérifier au prochain atelier).

A noter qu'à ce jour :

- il n'existe pas de passerelle automatique permettant l'ingestion des données de biodiversité publiées dans SDN dans EMODNet Biology
- il est plus facile de convertir le DarwinCore en SDN que l'inverse.

7.2.2. Convention paramètres et données de SeaDataNet (SDN) [Obligatoire]:

Les **unités des paramètres** (physique ou autres) sont à prendre dans la **liste P06 de SDN** et la **discipline** dans la **liste P08 de SDN**.

Pour les fiches de métadonnées, les **« Essential Variable » (EV) sont obligatoires**. Il existe plusieurs types d'EV : EOV, EBV, ECV (liste A05 de SDN). Ces EV peuvent être complétées dans les fiches de métadonnées par un paramètre de découverte (Liste P02) si besoin.

Pour la nomenclature d'un paramètre il est **fortement conseillé** de mettre la nomenclature la plus précise possible (**liste P01 de SDN**) plutôt que la nomenclature générique de la liste de découverte de paramètre (liste P02 de SDN). Dans le cas d'un paramètre précis manquant dans la liste P01, il est conseillé de se rapprocher de SDN pour le faire insérer dans la liste.

7.2.3. Convention paramètres et données du NetCDF-CF [Fortement recommandée]

Bien que le NetCDF version 4 apporte 5 nouveaux types de donnée utilisateur (**« UserDefinedType » : « Enum », « Opaque », « Compound », « VariableLength »**), il est **fortement déconseillé** de les utiliser.

Le NetCDF 4 apporte un niveau de hiérarchisation supplémentaire avec la notion de groupe dans son modèle (afin être conforme avec le HDF-5), il est **fortement déconseillé** de l'utiliser. En effet,

l'utilisation de plusieurs groupes dans un fichier NetCDF complexifie grandement l'utilisation de ce fichier.

En ce qui concerne les paramètres **date et heure** dans les fichiers NetCDF, il est **fortement recommandé** de les insérer sous forme d'entier (type « long ») avec un offset (optionnel) et un scale-factor (obligatoire). L'échelle de temps a adopter est **obligatoirement l'UTC** (Universel Temps Coordonné).

Lorsque des **axes** sont nécessaires dans un fichier NetCDF, il est **fortement conseillé** de bien définir les **axes et l'orientation**. La profondeur est souvent insérer en positif dans les fichiers NetCDF (ex 2000m), ce qui pose des problèmes lorsque l'on va utiliser ces fichiers avec des fichiers ayant des paramètres atmosphériques qui eux vont être aussi en altitude positive (2000m)

7.3. Paramètres de positionnement [Obligatoire] :

Il est nécessaire de préciser dans les attributs globaux, le **système de coordonnées géodésique** qui est utilisé comme référentiel pour les paramètres de positionnement. Le code **EPSG (European Petroleum Survey Group)** est à insérer dans les attributs globaux. A noter que ce standard est aussi utilisé par OGC (Open Geospatial Consortium).