Compte Rendu de l'atelier technique ODATIS du 5 Juin 2019

CR atelier technique Juin 2019

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

Numéro du délivrable	Titre court
	CR Atelier ODATIS Juin 2019

Titre long
Compte rendu de l'atelier technique ODATIS du 5 Juin 2019

Description courte

Auteur	Groupe de travail
Joël Sudre	

Dissémination	Copyright
	Pôle Odatis

Historique

Version	Auteurs	Date	Commentaires
0.1	Joël Sudre	02 Oct 2019	Version initiale
0.2	Cécile Nys	15 Oct 2019	Relecture et corrections
1	Cécile Nys	21 Oct 2019	Relecture, corrections et mise en page

Table des matières

1. Accueil et tour de table des participants

Liste des participants à l'atelier ODATIS :

- Dominique Briand (IFREMER) – DB,
- Frédéric Briol (CLS) – FB,
- Emilie Deschamps-Ostanciaux (IPGP - Formater) – EDO
- Gérald Dibarboure (CNES) – GD,
- Mark Hoebeke (Station Biologique de Roscoff) – MH,
- Dimitry Khvorostyanov (LOCEAN) – DK,
- Stevenn Lamarche (Univ. Brest) – SL,
- Laurence Lebourg (Univ. Brest) – LL,
- Maurice Libes (OSU Pytheas) – ML,
- Gilbert Maudire (IFREMER) – GM,
- Fabrice Mendes (OASU) – FM,
- Caroline Mercier (AKKA/CNES) – CM,
- Cécile Nys (IFREMER) – CN,
- Catherine Schmechtig (IMEV) – CS,
- Sabine Schmidt (Univ. Bordeaux / EPOC) – SS,
- Joel Sudre (OMP/LEGOS) – JS,

GM et JS présente l'ordre du jour (*voir : Agenda et accès aux présentations*[1]), en précisant que cet atelier est un peu différent des ateliers techniques précédents car il est uniquement sur un jour et qu'il va permettre de faire une revue technique des différents CDS du pôle ODATIS. Cet atelier étant dédié aux CDS, il devrait permettre d'avoir un échange technique des différents CDS et de permettre d'améliorer les interactions entre CDS et ODATIS et vice versa. GM rappelle le cahier des charges d'un CDS, que le modèle standard est celui du modèle OAIS et qu'il est nécessaire de mettre en place le FAIR data (trouvable, accessible, interopérable et réutilisable).

1 https://www.odatis-ocean.fr/activites/ateliers-techniques/ateliers-passes-archives/atelier-technique-juin-2019/

ODATIS propose de mettre en place deux type de CDS voire trois (ceci sera discuter au prochain atelier technique en Novembre):

- CAT (Centre d'Archivage et de Traitement)
- CATD (Centre d'Archivage, de Traitement et de Diffusion)
- CAD (Centre d'Archivage et de Diffusion)

GM présente ensuite les appels d'offre auxquels ODATIS à répondu :

- ANR Flash Copilote
- Phidias
- EOS Pillar
- Blue Cloud
- SeaDataCloud 2

Les réunions techniques de Blue Cloud et de SeaDataCloud 2 seront conjointes car l'audience est la même ainsi que les participants.

Discussion et tour de table

DK demande s'il existe un thésaurus centralisé pour le pôle.

GM répond qu'actuellement on utilise celui du NERC BODC (ceci étant en discussion aussi au niveau de IR Data Terra. En ce qui concerne les conventions, il est préférable d'utiliser celle de la convention CF, du GCMD (qui est représenté par le BODC en Europe).

CS connaît un bon interlocuteur au BODC ce qui permet de rapidement insérer de nouveaux paramètres dans les thésaurus. ODATIS recommande d'utiliser le vocabulaire P02 de Seadatnet car le P01 est trop général (voir P02 SDN parameter Discovery Vocabulary).

Le compte-rendu de l'atelier de Mars 2019 a été approuvé et mis en ligne sur le site ODATIS avec un accès publique (voir ce Compte-rendu)

2. Point d'avancement et retour d'expérience ERDDAP et Hyrax

Suite à l'atelier technique de Mars 2018, DB présente aujourd'hui un retour d'expérience sur le protocole ERDDAP implémenté à l'IFREMER et FB la pile logicielle Pangeo et Intake.

2.1. ERDDAP à l'Ifremer (Dominique Briand)

DB présente l'implémentation d'ERDDAP à l'ifremer (voir ODATIS_AT_erddap_Dbriand.pdf) après un rappel des fonctionnalités d'ERDDAP, DB informe que ERDDAP permet d'accéder à différent jeux de données à l'ifremer (voir ce lien), en particulier le jeux de données Argo, les produits Coriolis (CORA, NRTOA), les produits de SeaDataNet etc. Au total, 17 jeux sont disponibles.

Pour le jeux de données Argo, il a été nécessaire de prendre contact avec le créateur d'ERDDAP pour modifier le code afin de prendre en compte des formats multi-profils. De plus afin d'améliorer les performances d'ERDDAP, il a été nécessaire d'ignorer certaines variables, de ne pas prendre en compte les « fill_values » en fin de profils et d'adapter la gestion des « QC value ». Suite à cette introduction sur ERDDAP, DB montre différentes captures d'écran du site ERDDAP de l'Ifremer. Enfin il mentionne aussi l'existence du logiciel DIVA qui est un outil permettant le suivi des flotteurs Argo dont le code source est disponible sur github/quai20. Ce logiciel permet de faire un requétage sur les points affichés.

2.2. La pile logicielle Pangeo et Intake (Frédéric Briol)

FB présente (voir ODATIS_AT_Pangeo_Fbriol.pdf) la pile logicielle Pangeo qui est utilisable sur l'HPC du CNES : Hal). Pangeo est une communauté qui travaille au développement de logiciels et d'infrastructures pour faciliter la mise en œuvre des géosciences, dans le domaine du « big data ». Sa mission est d'entretenir et de distribuer un écosystème « open source » en python permettant avec des outils de mise à l'échelle, d'effectuer des calculs et des analyses scientifiques sur des jeux de données quelques soient leurs tailles et d'encourager une culture du développement accueillante et participative.

La pile logicielle Pangeo en python rassemble différentes librairies : Cython, NumPy, IP[y], jupyter, matplotlib, SciPy, DASK, pandas, xarray, learn, GCM, Iris, TensorFlow, cartopy, aospy.

L'architecture de Pangeo permet au travers de Jupiter d'utiliser les librairies Xarray et DASK sur un cloud ou un HPC sans trop de contrainte de programmation. La parallélisation du code se

fait très simplement par l'utilisation de ces deux piles logicielles.

Pour ce qui est d'utiliser des bibliothèques d'interpolation de grille, en général elles sont écrites en python et ne sont pas très efficaces (MetPy, Verde, pyresaple,etc.). Et il est assez déconseillé de les utiliser sur des grilles avec un nombre important de points. De plus ces bibliothèques ne gèrent pas la transition autour du méridien de Greenwich. CKDTree permet de convertir les lon, lat, altitude en coordonnées cartésiennes cependant les performances s'effondrent avec un nombre important de points (ex : 74649600 points plus de 4h pour faire une interpolation avec les plus proches voisins). Par contre les bibliothèques Boost geometry et Rtree ont des performances impressionnantes (pour le même problème que précédemment il n'a fallu que quelques secondes pour effectuer le calcul d'interpolation).

Pour gérer les axes des grilles régulières et irrégulières, il est conseillé d'utiliser « Axe » de la bibliothèque « Numpy » (voir l'exemple de l'utilisation de Axe et de Rtree dans la présentation).

La bibliothèque Intake permet la récupération de données sur un serveur distant, la gestion de données sur serveur distants, de créer des filtres de téléchargement etc.. Cette bibliothèque permet aussi d'utiliser Zarr. Cette bibliothèque devrait permettre d'utiliser efficacement les données entreposées sur le datalake du CNES.

Discussion

En ce qui concerne ERDDAP, il serait intéressant de contacter le développeur pour avoir accès au code source (→ voir politique de la NOAA à ce sujet) afin d'avoir la possibilité d'être autonome sur ERDDAP.

Il a été décidé de mettre des recommandations dans le cahier des charges des CDS pour l'utilisation d'outils. Cette action sera mise en place avant la fin de l'année 2019. Un point à discuter sur le cahier des charges est la continuité et la disponibilité des services proposés ainsi que les critères pour devenir un CDS (CAT, CATD, CAD).

Un Hackathon Pangeo est à programmer pour que les utilisateurs puissent se former à cette pile logicielle.

3. Revue technique des CDS

3.1. Introduction de la revue technique (Gilbert Maudire et Joël Sudre)

Le reste de la journée est consacrée à une revue technique de l'ensemble des CDS afin de complémenter le Bureau Exécutif qui se déroulera les 6 et 7 Juin. Le Bureau Exécutif va se concentrer sur le bilan d'activités, prospectives, les besoins budgétaire et personnel de chaque CDS.

Ici nous allons nous focaliser sur la partie technique de chaque CDS : organisation technique, infrastructure dédiée, volume de données, niveau d'interopérabilité, catalogue, besoin technique, FAIRitude, etc. Cette revue technique doit permettre de mettre à jour les points bloquants au niveau technique ainsi que les forces et faiblesses de chaque CDS par rapport au modèle OAIS (voir 07_Gilbert_Maudire_Cahier_Charges_CDS.pdf), évaluer si un CDS est un Centres d'Assemblage et de Traitement (CAT) ou un Centres d'Assemblage et de Traitement et de Diffusion (CATD).

3.2. CDS-SAT AVISO (Gérald Dibarboure)

GD présente (voir ODATIS_AT_AVISO_Gdibarboure.pdf) tout d'abord ce qu'est Aviso+ qui a été créé en 1992. Aviso est un service de diffusion, d'information et de promotion des données altimétriques. Ce CDS est constitué de deux sous-ensembles, une partie au CNES (le responsable par intérim étant T. Guinle) et une partie au LEGOS le CTOH (responsable technique F. Niño). Ce CDS est focalisé sur les produits dérivés de l'altimétrie avec quelques produits qui sont dans le périmètre de Form@ter.

Fonction « production »

Le CDS-SAT Aviso développe des produits altimétriques à valeur ajoutée comprenant des produits de niveau 2P au niveau 4 (DUACS, produit de glacio, FSLE etc.) mais aussi des produits de démonstration développés avec ou pour les laboratoires.

Le CDS-SAT Aviso fait aussi du soutien pour quelques campagnes à la mer en réponse à des demandes de laboratoires avec des données spatiales ad-hoc en temps réel et/ou retraitées en

temps différé.

La production est assurée sur les moyens informatiques de CLS et sur certains moyens internes CNES (HAL) et LEGOS/CTOH.

Fonction « entrée »

Le CDS assure l'harmonisation des données altimétriques au format NETCDF convention CF 1.6 et la création d'handbooks pour les produits développés. Le cds assure aussi le contrôle qualité en amont de la fonction « production », ce qui représente un grosse activité de Cal/Val. Il effectue aussi du contrôle qualité en aval de produits du CMEMS. Le QC est automatique avec des rapports techniques trimestriels ou hebdomadaires. Il y a des rejeux périodiques de données à problèmes qui sont remontés par les utilisateurs (réception et suivi de signalement d'anomalies).

Fonction « stockage »

Le stockage est réparti sur plusieurs moyens historiques au CNES. Les données critiques sont sur le serveur/SIPAD/SSALTO opérés par le CNES. Ce serveur est lié au système d'archivage à long terme sur bande. Les données reproductibles sont quant à elles hébergées sur plusieurs serveurs maintenus par CLS. Les données « en qualification » sont distribués par le CTOH. Enfin certaines données sont dupliquées par convenance pour être utilisées pour des protocoles simples sur des serveurs à la fois du CNES et de CLS.

La volumétrie représente environ 50 To hors données de niveau L0-L2 qui sont très volumineuses, Les données intermédiaires (de production/ suivi de QC ne sont pas prises ici aussi en compte).

Fonction « gestion des données »

Les données sont décrites dans des catalogues, des pages web sur différents produits ainsi que sur des documentations pour les utilisateurs. Les produits les plus utilisés sont décrits via des pages web et des documents dédiés.

En ce qui concerne la gestion des droits d'accès et des conditions d'accès, l'accès en écriture des données est très sécurisé, et l'accès en lecture est contrôlé pour certains produits. Certains produits sont accessibles uniquement sur demande des utilisateurs (base de données utilisateurs).

Fonction « administration »

Le CNES a un contrat d'exploitation avec CLS couvrant les items suivants :

- moyens techniques informatiques,
- maintien en conditions opérationnelles,
- maintien en configuration,
- passage en exploitation des logiciels.

Pour les produits SLA/OLA, la reprise est effectuée en 1 jour ouvré sur un incident mineur et en une semaine ouvrée sur un incident majeur. Un helpdesk est mutualisé sur plusieurs projets.

Fonction « planification de la pérennisation »

En ce qui concerne les données et produits non reproductibles, l'archivage long terme est assuré par le STAF du CNES et les données sont archivées dans deux bâtiments du CNES séparés de 300 m (fonction disaster recovery de CLS).

Les données sont toutes formatées avec un vocabulaire contrôlé, les fichiers sont en netCDF et les logiciels sont standardisés.

Fonction « accès »

- Hardware diffusion
    - Ferme web sécurisée (+redondance à chaud hors site)
    - LDAP/FTP : Vms sur ESX (+redondance froide hors site)
    - Scality RING (600 Tb utilisables, 12 serveurs, et 4 frontales dédiées)
    - Moyens additionnels internes CNES et LEGOS/CTOH

- Protocoles
  - FTP,OpenDAP, MOTU, LAS
  - popularité très dépendante du type de produits (principalement FTP et LAS)
- Contrôle d'accès
  - Accès authentifiés (web, FTP, DAP…) ou anonymes (LAS, FTP produits L2)
  - Pages personnelles pour certains produits
  - Reporting uniquement sur l'infra hébergée par CLS et au LEGOS
- Statistiques mensuelles
  - 800 utilisateurs uniques (14K inscrits)
  - 300k à 3M fichiers téléchargés
  - 1 à 3 To de téléchargement en moyenne
  - pour les produits très volumineux pas de log et de statistiques accessibles car ces données sont stockées sur l'infra du CNES.
- Helpdesk
  - environ 1500 inscriptions par an
  - environ 400 questions techniques par an
  - répartie sur 3 personnes
  - Escalation vers les équipes de production ou les experts techniques
  - Réponse en quelques heures pour les questions simples et 48h max

Besoins techniques

- Restructuration des moyens de diffusion et serveurs
- Mise en place d'archive pérenne systématique (à plus de 250 km)
- Modernisation des outils de visualisation/ communication web
- Capacité pour valoriser les produits dérivés de SWOT
    - Approche type Cloud (ne plus diffuser tout le produit)
    - Développements en cours sur le HPC interne CNES (PoC PANGEO)
    - Besoin d'une capacité de monté en charge (DIAS, EOSC)

3.3. CDS-SAT Brest (Dominique Briand)

DB présente le CDS-SAT Brest CERSAT (Centre ERS d'Archivage et de Traitement) qui est le centre de recherche et d'exploitation satellitaire de l'IFREMER fondé en 1985 pour permettre la diffusion des données de l'ESA ERS-1 et 2. Le CERSAT a ensuite mis à disposition pour la communauté scientifique les données d'Envisat, Metop, SMOS et CFOSAT en partenariat avec l'ESA, EUMETSAT, le CNES, l'Union Européenne et Météo-France. Ces principaux projets structurants se fédèrent autours des données de la température de surface, des vagues, de l'état de mer, des vents et de la salinité.

Fonction « production » et « entrées »

La collecte des données se fait au moyen d'un « downloader » en Python qui permet des recherches locales ou distantes, le mutiprotocols (ftp, http(s), opensearch) et la recherche de « pattern » (type répertoire, fichier, données etc.).

Fonction « stockage »

Les données sont ensuite post-traitées (renommage, réorganisation, checksum etc.) et sont stockées sur DATAMOR dans un espace « dataref » et un espace « datawork » (voir présentation).

Fonction « gestion des données »

La gestion des données s'effectue via un orchestrateur de chaînes qui permet de configurer les chaînes avec des spools d'entrée, des commandes d'exécution, et de faire la gestion globale des paramètres et variables. L'orchestrateur permet aussi de déployer des Crons et d'en gérer les logs.

Le catalogue des données gérées est accessible via le catalogue de métadonnées Sextant (Géonetwork), avec le profil des métadonnées respectant la norme ISO 19115-3. L'ensemble du vocabulaire utilisé est contrôlé que ce soit pour les « processings leve l », le « feature type », la convention Netcdf, etc.

Fonction « administration »

L'ensemble de l'infrastructure profite de l'infrastructure autour de Datarmor avec un suivi de production qui est accessible via un gestionnaire de logs et une console pour les opérateurs.

Le maintien en conditions opérationnelles se fait via des contrats de MCO avec le SISMER, le CERSAT et le CATDS qui sont des équipes dédiées à la gestion des données à l'Ifremer.

Fonction « planification de la pérennisation »

La pérennisation des données et des produits non reproductibles se fait sur des cartouches et une pérennisation en partenariat avec le CINES, le CNES et/ou le BRGM est envisagée et sera étudiée prochainement.

Fonction « accès »

Pour certaines données, leurs accès sont contrôlés via un enregistrement. L'ensemble des données est en accès libre une fois inscrit. Cet enregistrement permet de faire des statistiques de téléchargement. Les statistiques sur la volumétrie, la bande passante sont possibles mais cela reste à mettre en place.

L'accès aux données peut se faire via FTP, Thredds, WMS et http. Pour les données très volumineuses nécessitant un tuilage en amond le logiciel Syntool est utilisé. A noter qu'une

solution via jupiter (hub) est en cours de mise en place.

Sur le http, une temporisation des requêtes est possible et il a été noté que via le ftp il y a une surcharge de temps en temps au niveau du téléchargement. La charge de l'HPC datarmor est évaluée actuellement entre 40 et 50 % de sa charge maximale.

3.4. CDS-IS SISMER (Dominique Briand)

DB présente ensuite le CDS-IS SISMER (Système d'Information Scientifique pour la MER). Ce service a été mis en place pour répondre aux problématiques de collecte, d'archivage, de bancarisation et de diffusion des données d'observation *in situ* de l'océan qui sont par essence non reproductibles et très coûteuses à acquérir.

Il a été constaté que les données *in situ* sont à des degrés de maturité et à des stades d'élaboration très différents. Ces données *in situ* d'une même campagne sont acquises la plupart du temps séparément par différentes équipes et sont traitées de façon différente aussi. Si on veut les comparer avec d'autres données historiques, il est nécessaire de les intégrer dans des bases de données ce qui nécessite une harmonisation des données et des métadonnées associées. Ceci est souvent fait au niveau d'un réseau d'observation par exemple.

Ces données harmonisées sont dès lors, les éléments de base permettant de créer et d'alimenter de nouveaux produits. Le SISMER s'inscrit donc pour répondre à cette problématique avec une démarche qualité utilisant la norme ISO 9001. Plusieurs bases de données de l'ifremer sont concernées par cette démarche en particulier :

- Coriolis et la banque de physique/chimie marines,
- la banque de Géophysique et géologie marines,
- Quadrige[2] (qui concerne plus spécifiquement l'environnement côtier),
- Harmonie (système d'information halieutique),
- Bigood (qui est une base d'échantillons biologiques et géologiques en milieu benthique,
- la base de campagnes à la mer de la flotte française.

Fonction « entrées »

Il existe plusieurs niveaux d'automatisation de chaque observatoires. Pour les flotteurs Ago, l'ensemble de l'observatoire est automatisé, par contre pour les campagnes et les pêches l'automatisation, n'étant pas possible, se fait de manuellement. Le SISMER distribue en particulier les données des campagnes à la mer françaises acquises sur les navires de la flotte française (voir schéma de la fonction « entrées » dans la présentation). A noter que pour l'ensemble des données qui sont orphelines, il est possible de les mettre à distribution via SEANOE qui est maintenant un service du pôle ODATIS.

Fonction « stockage »

La conservation physique des données gérées est archivée sur bande et sur l'espace de données Datarmor (voir schéma dans la présentation).

Fonction « gestion des données »

La gestion des droits d'accès et des conditions d'accès est assez délicate pour les données de campagne à la mer car les scientifiques, qui en ont fait la collecte, désirent garder certaines prérogatives et en maîtriser la circulation. Pour ce faire, un embargo sur la donnée est possible pour une durée de 2 ans après sa validation.

Fonction « administration »

La fonction « administration » des moyens techniques informatiques est assurée via 3 pôles :

- le SISMER qui exploite et gère la donnée,
- le RIC qui s'occupe de l'infrastructure informatique,
- le ISI qui développe le système d'information.

Le maintien en conditions opérationnelles (MCO) est effectué via des contrats.

Fonction « planification de la pérennisation »

Afin de pérenniser les données, les métadonnées, la description des formats, du vocabulaire et des logiciels, l'ensemble est contrôlé. Les catalogues de métadonnées ISO sont insérés dans l'outil sextant (voir sextant.ifremer.fr). Cette API est disponible et elle permet d'importer le catalogue ODATIS dans un site dédié. Le serveur de données est un serveur QGIS avec les services web OCG et non normalisés. La base est une base utilisant Eleasticsearch / Cassandra.

3.5. CDS-IS CORIOLIS (Thierry Carval)

Le CDS - IS Coriolis est présenté par TC. Ce centre de données utilise la même infrastructure que le CDS-IS SISMER et les mêmes ressources informatiques (IFREMER Datamor etc.).

La base Coriolis est alimentée en continu par un flux de données *in situ* qui permet de créer le produit CORA disponible en téléchargement. Le catalogue sur le site web de Coriolis permet d'accéder à l'ensemble des produits disponibles (CORA, ANDRO, GOI, etc.).

La base Coriolis est actuellement une base Oracle qui va certainement être migrée en base Cassandra NoSQL pour être plus homogène avec les autres bases gérées par les équipes de l'Ifremer.

3.6. CDS-IS OASU (Fabrice Mendes)

FM présente le CEDONA (Centre de Données Nouvelle Aquitaine) qui héberge le CDS-IS OASU. Le CEDONA ayant déjà été présenté lors de l'atelier précédent (voir 03_Pascal_Calvat_CEDONA_OASU.pdf), FM ne présente que la partie qui concerne le CDS-IS du pôle.

Fonction « production »

Les bases inclues dans le CEDONA pour le pôle ODATIS sont :

- SNO SOMLIT (Service d'Observation en Milieu LIToral) : Importation manuel de fichier CSV avec un contrôle par analyse statistique automatique, les résultats sont décrits par des graphiques sur le site web, possibilité de faire un export en CSV, ODV (via SDC),
- SO MAGEST (surveillance de la qualité de l'eau de l'estuaire de la Gironde) : Importation automatique des fichiers de sonde avec une correction et une analyse manuelle avec une exportation des résultats sous forme de fichier CSV,
- SO Benthos Resomar :Importation manuel de fichier CSV avec une expertise en amont pour déterminer les taxons, export des résultats en CSV,
- SNO MEMO (Mammifères Echantillonneurs de Mileu Oc éanique),
- SO Molluscan Eye (valvométrie haute fréquence).

Fonction « stockage »

Toutes les données sont sur disques durs avec plusieurs copies physiques, la plupart du temps sous forme de base de données relationnelles et archivées sous format dump. La mise à jour des copies se fait soit de façon hebdomadaire en incrémentiel et tous les trois mois en intégral ou de manière quotidienne avec copie intégrale toute les semaines.

Fonction « gestion des données»

Les conditions d'accès pour SOMLIT, MAGEST et Benthos nécessite la signature d'une charte via le portail web. Après signature de la charte, l'utilisateur à accès à l'ensemble des bases. La description des données est disponible en ligne.

Fonction « planification de la pérennisation»

En ce qui concerne les données et produits non reproductibles, les fichiers de données sont sauvegardés mais pas pour tous les projets. Il en est de même pour les résultats d'analyse.

Les métadonnées ont divers niveaux de maturité selon les projets (SDC, fiches PDF, etc.).

Pour les services d'observation, il existe une documentation pour décrire les formats et le vocabulaire employés. A noter que pour les nouveaux projets, la mise en place de fiche de description sur les formats est toujours effectuée.

Les logiciels sont tous développés en locale et placés sur une forge Gitlab avec leur documentation et un framework reconnus.

Fonction « administration»

En terme de moyen technique, le CEDONA possède une salle serveur avec une dizaine de machines physiques en 3 clusteurs et un espace de stockage de ~100To. Le service informatique de OASU repose sur 7 ITA avec un % ETP sur les projets qui est défini par appel d'offre semestriel.

Le maintien en condition opérationnelles est effectué par une supervision permanente avec des interventions rapides aux heures de bureau et en « best effort » le reste du temps. L'OASU possède un clusteur en Spare et fait une sauvegarde des VM ainsi que des données. La haute disponibilité n'a pas encore été mise en place mais les données critiques sont sauvegardées quotidiennement au LEGOS à Toulouse.

Il est à noter que l'OASU dépend du réseau universitaire et que cette dépendance pose plusieurs problèmes techniques et des contraintes (problème d'accès).

Le maintien en configuration est effectué sous GIT avec un portage sur Ansible prochainement.

Fonction « accès»

Pour le SNO SOMLIT, le formulaire de demande est obligatoire mais la validation est en cours de suppression. Un accès par SDN est en cours de création.

Pour le SO MAGEST et SO Benthos, une convention est signée qui permet d'obtenir un login/password pour accéder aux données .

Fonction « utilisateurs»

il n'y a pas de base commune entre les différents services d'observation car la fonction utilisateur est faite par projet avec des technologies propres.

Besoins techniques

Pour le matériel, le besoin le plus urgent est de trouver sur un espace distant (>250 km) 10 To pour faire une duplication des sauvegardes.

Le Cedona souhaite avoir des instructions précises de la part d'ODATIS au niveau organisationnel en particulier :

- Guidelines pour les outils communs ODATIS (Ex : ERDDAP) ou les normes (DMP),
- Formation sur les outils (bilan pour liaison avec les outils locaux),
- Pourquoi pas une image figée et commune des outils Odatis (type VM ou docker).

3.7. CDS-IS Sorbonne Université LEFE/CYBER (Catherine Schmechtig)

CS présente le CDS-IS LEFE/CYBER et la base de données LEFE-CYBER associée.

CYBER (Cycles Biogéochimiques environnement et ressources) est une des composantes du programme LEFE (Les Envelloppes Fluides de l'Environnement). La base CYBER est un support pour les projets du programme LEFE-CYBER de l'INSU qui a pour but de collecter les données, les archiver et les rendre accessibles pour la communauté scientifique. Les projets sont principalement des campagnes hauturières mais il existe également des mesures au point fixe, des sorties de modèles, etc.

Fonction « production »

Du fait de la grande hétérogénéité des formats et des données qui sont envoyés par les scientifiques pour être intégrés dans la base, il n'est pas possible actuellement de créer une chaîne de traitement, et d'avoir un contrôle de cohérence des résultats. Toutes ces opérations se font de manière manuelle au cas par cas.

Fonction « entrées »

L'harmonisation des données se fait donc aussi de manière manuelle. Ces opérations sont effectuées par WD (en CDD) et elles consistent à intégrer le vocabulaire (SDN P01/P02), à harmoniser l'empreinte spatio temporelle, à formater les fichiers etc. Il n'est pas possible actuellement de faire un contrôle qualité des données car il doit être fait en amont par les scientifiques.Fonction « stockage »

Les données sont conservées sur le poste fixe de CS à Villefranche-sur-mer et sur des disques durs USB, sur le serveur LEFE-CYBER et sauvegardées par le service informatique de l'IMEV. Les données sont aussi copiées (~ tous les 6 mois) sur le poste fixe de CS à Paris.

Fonction « gestion des données »

Les données gérées et les résultats des chaînes de traitement sont décrites sur le site de la base LEFE-CYBER. La gestion des droits d'accès ainsi que des conditions d'accès sont aussi décrites sur le site web de la base.

Fonction « administration »

La base est hébergée sur un serveur autonome à Villefranche-sur-Mer (serveur LEFE-CYBER) et elle est accessible via une interface Geonetwork et ERDDAP sur Vmware (tomcat, postgreSQL)

Le maintien en condition opérationnelle se fait par une jouvence régulière tous les 4 à 5 ans du serveur et le maintien en configuration par des mises à jour régulières.

Fonction « planification de la pérennisation »

La pérennisation des données est obligatoire car les données et les produits sont non reproductibles. La pérennisation se fait aussi sur les métadonnées.

Fonction « accès »

Le contrôle d'accès est effectué via un formulaire de demande d'accès aux données. Si la donnée est sous l'emprise d'un embargo la demande est envoyé au « Principal Investigator »

(PI), sinon il y a envoi d'un login/pasword avec copie au PI pour accéder à la donnée. L'ensemble des adresses de messagerie ayant fait une demande d'accès sont stockées.

WD travaille actuellement pour mettre en place sur l'ensemble de la base des recherches par ordre alphabétique des campagnes et/ou par l'empreinte spatio temporelle.

Fonction « utilisateurs »

La bibliographie, les cartes, les présentations scientifiques, les posters ayant attrait aux données incluses dans la base sont accessibles sur le site LEFE-CYBER. Toutes ces informations sont classées par campagne. La hotline sur chaque campagne se fait par messagerie.

Besoins techniques

Il est urgent d' :

- arrêter des choix sur l'authentification,
- arrêter des choix sur le vocabulaire (mise à disposition des fichiers RDF de SDN),
- arrêter des choix sur les unités,
- arrêter des choix sur les formats (template csv, netcdf).

CS mentionne la nécessité de mettre en place une sauvegarde distante, d'avoir un outil d'authentification fourni par ODATIS pouvant séparer les données libres et avec embargo.

Le fait de ne pas avoir un endroit où l'on peut saisir des métadonnées avec des listes de vocabulaires contrôlés sur lesquelles on ne peut pas agir, de dépendre de la disponibilité de personne tierce pour rendre les données visibles et d'accorder toute la visibilité au système qui diffuse les données en masquant la visibilité des personnes qui produisent/analysent les données est très frustrant, décourageant et chronophage.

A noter que CS mentionne qu'il est urgent de faire une présentation ODATIS à Villefranche-sur-mer et qu'un CES pigment serai utile.

3.8. CDS-IS Sorbonne Université / Pelagos (Mark Hoebeke)

MH présente le CDS-IS SBR qui est intégré dans les services soutien de la station biologique de Roscoff au niveau du service informatique et bio-informatique.

Le périmètre des données concerne la base de données PELAGOS du RESOMAR dont les objectifs sont :

- la bancarisation des données de biodiversité de l'écosystème pélagique côtier,
- la mise en accès publique des métadonnées,
- la mise en accès communautaire des données.

Une application web est en place depuis 2014 permettant de faire l'export des suivis SOMLIT-ASTAN vers SDC et EMODNet Biology, ainsi que l'export de l'ensemble des suivis phytoplanctoniques vers la base du SNO PHYTOBS. Le SNO PHYTOBS a été nouvellement créé avec comme objectifs de développer une bdd dédiée avec un portail d'accès (couche OGC). Ce nouveau service est la fusion de la base PELAGOS avec la partie de la base Quadrige^2 s'intéressant à l'écosystème pélagique côtier. Cette base est en cours de réalisation.

A noter qu'une demande de SNO BENTHOBS avec pour objectif la bancarisation des données de biodiversité des populations benthiques a été déposée pour être labellisé. Cette demande fait suite à la nécessité de faire une fusion des données de la base existante Benthos de l'INSU et d'une partie de la base Quadrige^2 de l'Ifremer.

Enfin le CDS-IS SBR s'occupe aussi d'une base de données Inventaires dont les objectifs sont de mettre en accès les données de biodiversité (faune + flore) récoltées dans une zone allant de Portsall aux Sept îles. Elle est constituée de données primaires (observations de terrain et observations « historiques ») issues de faunes et flores de référence. Une application web a été mise en place pour accéder aux données depuis 2011 avec un export vers EMODNet Biology.

Au sein de ce CDS, il existe un entrepôt de données « HF » qui est la bancarisation de mesures physico-chimiques générées par des dispositifs d'acquisition embarqués sur différentes plateformes (bouées, sondes, ferrybox). Cet entrepôt permet d'accéder aux données brutes

pour les collaborateurs, d'accéder à des représentations graphiques pour le grand public via un accès web depuis 2012. Un export automatisé vers la base Coriolis d'ODATIS des données de la bouée ASTAN (dans le cadre du SNO COAST-HF) et des ferribox a été mis en place.

Fonction « production »

Il existe des chaînes de traitement pour:

- les données physico-chimiques (acquisitions automatisées avec un contrôle de cohérence des résultats basé sur des valeurs numériques spécifiques indiquant des défaillances matérielles. Il a aussi un contrôle a posteriori par les responsables scientifiques avec la possibilité de faire de corrélations,
- les données de biodiversité (insertions manuelles). Dans le cadre du CDS, il n'y a pas de traitement des données acquises et le contrôle de cohérence est réalisé en amont de l'insertion par les responsables des suivis. L'outil d'insertion effectue des contrôles de cohérence supplémentaires (pour les bases PELAGOS,Benthobs, Inventaires).

Fonction « entrées »

Un effort important a été porté sur l'harmonisation des données avec une mutualisation des métadonnées par type de capteur pour les données physico-chimiques, des protocoles et référentiels taxonomiques communs à l'ensemble des suivis des données de biodiversité.

Pour la base PELAGOS des ateliers d'experts sont organisés afin d'améliorer l'harmonisation de la détermination des spécimens.

Pour les données du SNO PHYTOBS, une liste exhaustive des taxons et des regroupements taxonomiques est en cours de validation par les partenaires.

Le contrôle qualité n'a à ce jour pas de procédures standardisées.

Fonction « stockage »

Deux méthodes de stockage sont utilisées pour l'ensemble des données du CDS : une base de données relationnelle et un espace de stockage dédiés pour les données annexes (images, espace de travail, tableaux de données pré-extraits pour la mise à disposition).

L'ensemble des données représentent un volume inférieur à 1 To et est stocké sur l'infrastructure de stockage sécurisé de la plateforme AbiMS (2 Po).

Fonction « gestion des données »

La description des données gérées est faite au moyen :

- d'un document intégré au SMQ qui recense les instances des bases de données hébergées,
- des descriptions sommaires figurant dans les pages de documentation des applications WEB permettant l'accès aux données.

En ce qui concerne la gestion des droits d'accès et des conditions d'accès, un référent scientifique par projet décide, octroie des droits sur les espaces de stockage, et utilise le help desk pour leur mise en place effective. La gestion d'accès aux parties restreintes des applications est gérée dans des tableaux partagés avec des responsables scientifiques.

Fonction « administration »

Les moyens techniques informatiques sont hébergés et sécurisés dans des locaux dédiés. L'administration de l'infrastructure de stockage et d'hébergement des applications est assurée par l'équipe technique locale (% ETP ITA du service informatique et bio-informatique).

Le maintien en condition opérationnelle et le maintien en configuration sont pris en charge par l'équipe d'administration de la plateforme AbiMS pour les composants relevants de cette infrastructure. Les composants tiers dédiés à l'activité du CDS sont pris en charge sur des % ETP affectés au CDS. Les développements internes sont inclus dans le processus certifié « ingénierie logicielle » de la plateforme.

Le passage en exploitation des logiciels pour les applications développées en interne se fait a posteriori d'une phase de test et de validation.

Fonction « planification de la pérennisation »

L'ensemble des données sont non reproductibles et il existe dans certains cas des copies partielles qui sont stockées localement sur des postes de travail ou sur des dispositifs d'acquisition avec un temps de rétention limité.

Les métadonnées sommaires (horodatage, description des protocoles) sont disponibles pour l'ensemble des suivis. Les métadonnées plus riches sont disponibles pour les jeux de données « publiés » dans des infrastructures/projets externes (SDC, EMODNet Biology,…) et elles utilisent les standards préconisés par les projets pour le vocabulaire et le format.

Les données sont exportées au format CSV (COAST-HF), Bio-ODV (SDN) ou Darwin Core Archive (EDMONet Biology) et sont téléchargeables au travers des applications web dédiées des différents projets.

Pour la partie logiciel, les :

- bases de données relationnelles sont maintenues sous plusieurs versions de PostgreSQL qui sont migrées vers des versions les plus récentes en fonction des besoins,
- applications web dédiées sont développées en interne pour l'insertion/consultation des données avec une gestion du code et de la documentation par GitLab,
- applications hébergées pour les besoins de projets externes sont déployées si nécessaire dans des VM avec un environnement figé.

Fonction « Accès »

La fonction « accès » est effectuée au travers des protocoles http et ftp avec un contrôle d'accès des espaces de stockage effectué par des ACL placés par projet. Un compte est dédié pour chaque bdd afin d'avoir un accès différent pour chaque base. Un annuaire LDAP local

peut être utilisé pour l'accès à certains applicatifs.

Besoins techniques

Un renforcement de la sécurisation des espaces de stockage par réplication sur des sites distants a été demandé.

3.9. CDS-IS SHOM (Valérie Cariou)

VC présente le CDS-IS SHOM incluant différentes thématiques :

- RADAF HF constitué de données de courant de surface (base RHF Iroise),
- MAREE constitué de données de hauteur d'eau HF (base RONIM, REFMAR et SONEL),
- Hydrologie constitué de données de température et de salinité.

Les diverses fonctions sont présentées pour chaque thématique.

La thématique Hydrologie

Les données sont collectées par le SHOM et la Marine Nationale.

. Fonction « entrées »

- Collecte des données, contrôle de la conformité des données collectées au regard des documentations (procédure, mode opératoire, formulaire conformément à démarche qualité Shom) ; documentations gérées dans un référentiel documentaire
- Qualification et bancarisation des métadonnées et données avec l'outil SCOOP2_EVOLOG; outil mutualisé avec Coriolis.

. Fonction « stockage »

L'archivage perenne des données brutes reçues au Shom (fichiers, cahiers de quart, rapports de campagnes) + sauvegarde en base de données Oracle.

. Fonction « gestion des données »

- Métadonnées (date, heure, type instrument, producteur (pays, institution, plateforme), degré de protection de la donnée, données météorologiques …) et qualification
- Données : qualification des données (code qualité par niveau et par paramètre)
- Utilisation de codes internationaux et vocabulaires communs (ex: utilisation de vocabulaires gérés par le BODC dans le cadre de la diffusion SeaDataNet http://seadatanet.maris2.nl/v_bodc_vocab_v2/welcome.asp) et de formats standards

https://www.seadatanet.org/Standards/Data-Transport-Formats)

. Fonction « planification de la préservation »

Cette fonction est assurée par service informatique du Shom.

. Fonction « accès »

- Diffusion des données publiques via Coriolis Ifremer
- Diffusion des données publiques sur le portail SeaDataNet http://seadatanet.maris2.nl/v_cdi_v3/search.asp

La thématique Radar HF

. Fonction « entrées »

- Exploitation, collecte et validation des données radar HF sous-traitées à Actimar,
- Mise à disposition par Actimar au SHOM en TR (sous 1 heure) des données de courant sur un serveur dédié (avec un accès donné également à IFREMER),
- Transmission trimestrielle au SHOM des données brutes + validées (bande, clé USB, cassette).

. Fonction « stockage »

- Conservation physique au SHOM des données,
- Données accessibles aux utilisateurs de Datarmor.

. Fonction « gestion des données »

Les métadonnées sont consultables sur data.shom.fr (fiche métadonnées) :

- Les données du portail sont référencées dans un catalogue standard https://services.data.shom.fr/geonetwork,
- A ce catalogue est associé un nœud de moissonnage CSW: https://services.data.shom.fr/geonetwork/srv/fre/csw-produits.

Un vocabulaire commun est utilisé (ex: vocabulaire géré par le BODC dans le cadre de la diffusion SeaDataNet http://seadatanet.maris2.nl/v_bodc_vocab_v2/welcome.asp) ainsi que des formats standards (exchange).

. Fonction « planification de la préservation »

- Copie des supports (bandes, clés, cassette) en cours,
- Pérennisation de sauvegarde des données traitées (à compter de 01/2016) assurée.

. Fonction « accès »

- data.shom.fr → Observations radar HF diffusées sous licence Opendata "Licence Ouverte" (version 1.0 d'octobre 2011), définie par la mission Etalab,
- diffusion sur le portail SeaDataNet (http://seadatanet.maris2.nl/v_cdi_v3/search.asp) en cours dans le cadre du projet H2020 SeaDataCloud),
- données sous datarmor (identifiant ID-90).

. Fonction « utilisateur »

Cette fonction se fait via le portail data.shom.fr avec un help desk accessible à l'adresse : data-support@shom.fr et une aide en ligne https://services.data.shom.fr/static/help/Aide-en-ligne_DATA-SHOM-FR.pdf.

La thématique Marée

. Fonction « entrées »

- Hauteurs d'eau collectées par les marégraphes RONIM et marégraphes de différents partenaires,
- Validation des données du réseau RONIM et ponctuellement de certains marégraphes de partenaires selon les besoins du Shom (conformément au GLOSS (Global sea Level Observing SyStem)).

. Fonction « stockage »

La bancarisation des données brutes et validées se fait dans une bdd ORACLE.

. Fonction « gestion des données »

Les données du portail sont référencées dans un catalogue standard (http://services.data.shom.fr/geonetwork). A ce catalogue est associé un nœud de moissonnage CSW : https://services.data.shom.fr/geonetwork/srv/fre/csw-produits.

. Fonction « planification de la préservation »

- Préservation assurée par le service informatique du Shom,
- Inventaire des documents marégraphiques historiques et numérisation des documents papiers (projet Datarescue).

. Fonction « accès »

L'accès aux données marégraphiques se fait via le site data.shom.fr. Le téléchargement des données est possible en mode manuel, en mode batch et en flux (voir la présentation pour la liste des données disponibles – planche 10 et 11).

. Fonction « utilisateur »

Cette fonction se fait via le portail data.shom.fr avec un help desk accessible à l'adresse : data-support@shom.fr et une aide en ligne https://services.data.shom.fr/static/help/Aide-en-ligne_DATA-SHOM-FR.pdf.

3.10. CDS-IS OMP (Joël Sudre)

JS présente le CDS-IS OMP dont la présentation a été préparée par Philippe Techiné et Gaël Alory.

Le tableau ci-dessous liste les jeux de données inclus dans ce CDS :

SNO SSS	LEGOS/OMP	SSS, global
ROSAME	LEGOS/OMP	Niveau de la mer, TAAF
SEDOO	OMP	SSS TD (Temps Différé) + produits SSS, global
SNO SONEL (CDS-IS SHOM)	LEGOS/OMP (admin.)	Niveau de la mer, données françaises
SNO PIRATA	LEGOS/OMP (admin.)	Mouillages ADCP + campagnes en mer, Atlantique Tropical

Il est à noter que certaines données sont envoyées en temps différé au CDS-IS Coriolis et Sismer pour intégrer les bases de ces CDS.

Fonction « production »

Les chaînes de traitement d'acquisition sont automatisées, avec un traitement et un contrôle qualité des données en Temps Réel (TR) mutualisée entre SSS et ROSAME.

Le logiciel TSGQC (ThermoSalinoGraphe Quality Control) de traitement, qualification et correction des données SSS Temps Différé (TD) est aussi utilisé.

Les résultats sont sous divers types de fichier :

- SSS TR : Fichier annuel par navire,
- SSS TD : Fichier par voyage et par navire + produits grillés 2D et 3D,
- ROSAME : Fichier par site de mesure.

Fonction « entrées »

L'ensemble des données est harmonisé avec une unicité et une automatisation du traitement en TR avec un logiciel unique TSQGC SSS. Le format des fichiers TR/TD est unique.
En ce qui concerne le contrôle qualité :

- TR : Contrôle qualité automatisé basé sur des tests recommandés par GOSUD + remontée d'alerte par messagerie en cas de problème, attribution de code de traitement et écart aux climatologies pour SSS,
- TR : Visualisation sur site web des courbes de mesures des capteurs + informations pour un suivi opérationnel des stations d'acquisition,
- SSS TD : Visualisation des données avec TSGQC, attribution de flags, comparaison et correction avec des données externes et colocalisées,
- ROSAME : Analyse et prévision de marée, comparaison et correction avec des données externes et colocalisées.

Fonction « stockage »

L'ensemble des données gérées est conservé physiquement avec un stockage LEGOS/OMP pour l'assemblage et le traitement

- données TR SSS et ROSAME dans une BDD Berkeley DB,
- données TD et produits SSS dans des fichiers ascii et netcdf,
- données ROSAME dans des fichiers ascii.

Le stockage pour la diffucion se fait au SEDOO/OMP :

- BDD PostgreSQL,
- Produits SSS dans des fichiers ascii et netcdf.

Fonction « gestion des données »

- SNO SSS : SSS, SST, date et position des mesures,
- ROSAME : Mesures de pression atmosphérique, pression de fond, température et conductivité de l'eau, mesures de tirant d'air, date des mesures, calcul du niveau de la mer.

En ce qui concerne les résultats des chaînes de traitement :

- SNO SSS : Code de traitement TR et écart aux climatologies, flag et correction TD,

- ROSAME : Données avec contrôle qualité et données corrigées de la dérive des capteurs.

Pour la gestion des droits d'accès et des conditions d'accès :

- Suivi SSS TR sur site web avec accès restreint, suivi ROSAME en accès libre,
- Accès aux données SSS TD via une interface web avec identification,
- Accès aux données ROSAME sur site ftp public LEGOS.

Fonction « administration »

Le matériel du LEGOS/OMP est utilisé pour l'assemblage et le traitement : Moyens de calcul et stockage du LEGOS dont « Virtual Machine » (VM) de traitement TR sous Linux + baie de stockage (Lustre) maintenus et sauvegardés par le service informatique du LEGOS. Le matériel exacte du SEDOO/OMP pour la diffusion n'est pas décrit.

Le logiciel d'assemblage et de traitement est situé au LEGOS/OMP avec une chaîne de traitement TR basée sur des modules génériques Perl, logiciel TSGQC Matlab, divers logiciels Fortran et C (analyse, prévision de marée). Le logiciel SEDOO/OMP pour la diffusion n'est pas décrit.

Le passage en exploitation des logiciels est réalisé par les acteurs du CDS-IS OMP.

Fonction « planification de la pérennisation »

L'ensemble des données est non reproductible (données *in situ* dans le cadre des SNO pérennes).

Pour les métadonnées :

- instrumentales (date de calibration, référence des capteurs, etc.),
- métadonnées définies lors de l'attribution de DOI sur les données, les produits SSS et sur les campagnes à la mer NIVMER/ROSAME.

La description des formats et du vocabulaire est définie sur des standards internationaux GOSUD (SSS) et GLOSS (ROSAME).
Le logiciel TSGQC a été mis sur un dépôt Git et un projet de DOI est en cours.

Fonction « accès »

L'identification des utilisateurs se fait sur l'interface web de distribution des données TD et des produits SSS (http://sss.sedoo.fr). Actuellement, il n'y a pas de contrôle lors de la distribution des données ROSAME qui est effectuée sur le site public du LEGOS.

Fonction « utilisateur »

L'interface web de distribution des données TD et produits SSS est mis en place via le SEDOO (http://sss.sedoo.fr) avec différentes rubriques disponibles :

- news sur la page d'accueil,
- rubrique Data access : Choix du/des navires, de la zone géographique, des gammes de température et salinité, de la période temporelle, du niveau de qualification des données. Possibilité de récupérer en une seule fois toutes les données ou seulement les données qualifiées Good/Probably Good,
- accès aux métadonnées à partir des landing pages des données et produits (DOI),
- rubriques Data policy, FAQ, Contact us.

Besoins techniques

L'impact de la restructuration des SI des laboratoires de l'OMP vers un seul SI OMP (action qui devrait débuter en septembre 2019) sur le SNO SSS et ROSAME n'est pas connu à ce jour.
Il y a un manque de personnel technique pour la mise à jour régulière des produits. Pour palier à ce manque de moyen humain les SNO SSS et ROSAME ont recours à des stagiaires.

3.11. MIO – Cytométrie (Maurice Libes)

Un état des lieux de la base cytométrie du MIO est présenté par ML en vue d'une future création d'un CDS-IS-Cytométrie à l'OSU Pytheas. Cette présentation a été créée avec la participation de Patrick Raimbault. Il est à noter que le MIO/OSU Pytheas, en plus des données de cytométrie, participe à l'alimentation de divers SNO en particulier pour le SNO SOMLIT (données envoyées à OASU), SNO MOOSE (données envoyées au SEDOO), SNO EMSO, SNO COAST HF, et SNO Phytobs.

La cytométrie en flux automatisée HF *in situ* (capteurs autonomes) et son traitement permet la résolution d'un ensemble de classes du phytoplancton (mais aussi du microphytoplancton) et une prise d'image. Une réflexion a été mise en place avec une collaboration de SeaDataCloud (SDC) pour la standardisation du vocabulaire associé au phytoplancton (nom des groupes fonctionnels du phytoplancton résolu par classe de tailles (pico-eucaryotes, nano-eucaryotes), nom de genre (synéchlorocossus, prochlorocossus), et identificateurs optiques).

Fonction « production »

Une chaîne de traitement est dédié pour chaque SNO (SOMLIT, MOOSE, PHYTOBS) afin de mettre en base les données d'acquisition de terrain, la collecte des échantillons sur les sites SOMLIT marseillais, et les résultats des analyses en labo des données.

Le contrôle de cohérence des résultats est effectué par le SNO via un contrôle qualité (SISMER, SEDOO).

La description des résultats se fait via les sites internet de SOMLIT (http://somlit-db.epoc.u-bordeaux1.fr/bdd.php) et MOOSE (http://www.moose-network.fr/DATA_MOOSE/app/#/view1).

. La fonction « production » du SO MIO :

Les chaînes de traitement sont développées et maintenues par les équipes du MIO et du services d'observation de l'OSU et elles intègrent les acquisitions capteurs de terrain, la collecte des échantillons sur site (drone marins ocarina, radar HF, station météo htmnet, et Robot MII EMSO), les analyses en laboratoire et la publication de DOI.

Il n'y a pas de contrôle systématique de cohérence des résultats (dépendant des projets scientifiques).

La description des résultats s'effectue via le portail d'accès (Findable, Accessible, Interoperable, Reusable (FAIR)) aux données environnementales de l'OSU Pytheas au moyen :

- d'un catalogue de métadonnées (geonetwork),
- d'un serveur de données cartographiques (geoserver),
- d'une plateforme d'accès graphiques et fichiers ERDDAP.

. La fonction « production » pour SeaDataCloud :

La chaîne de traitement comprend plusieurs éléments (Cytosense avec une extraction en R des fichier CSV cyto, un CQ basé sur Talend, une base de données SQL (cytobase) qui permet des extraction en Common Data Index (CDI) XML.

Le contrôle de cohérence des résultats se fait par un contrôle scientifique des données de cytométrie (tailles, propriétés optique des cellules) du Cytosense. Ce contrôle se fait aussi lors du workflow avec SDC (voir slide 12) sur les métadonnées décrivant chaque jeu de données

avec les paramètres de mesures, position, date, etc.

La description des résultats se fait via le portail web de SDC : https://www.seadatanet.org/Metadata mais aussi localement via le portail de la bdd cytobase : https://chrome.mio.univ-amu.fr/.

Fonction « entrées »

L'harmonisation des données est effectuée pour intégrer les différents standards des SNOs (SOMLIT, MOOSE). Pour le SO OSU, elle s'effectue sur les fichiers bruts et sur la mise au format NetCDF, ODV, CSV via la plateforme ERDDAP. Pour SeaDataCLoud les métadonnées sont produites au format CDI XML norme ISO 19115 et les données brutes au format ODV.

Le contrôle qualité est effectué par les SNOs et pour la cytométrie par un contrôle visuel des résultats dans les fichiers CSV par le responsable scientifique de la cytométrie.

Fonction « stockage »

Pour les SNOs, la conservation physique des données est externalisée dans le lieu d'implantation du SNO.

Pour la cytométrie les données sont conservées à l'OSU (sur une baie de disques, redondée et sauvegardée) et à SDC.

Fonction « gestion des données »

Pour voir la description très détaillée de cette fonction voir les planches 16,17 et 18 de la présentation.

Fonction « administration »

Les moyens techniques informatiques sont ceux de la salle serveur du MIO. Le maintien en condition opérationnelles est effectué par l'équipe informatique de l'OSU Pytheas et le maintien en configuration par son service observation (1.6 ETP). Le passage en exploitation des logiciels est effectué par ML.

Fonction « planification de la pérennisation »

L'ensemble des données sont non reproductibles et doivent être stockées et sauvegardées. Les métadonnées sont normalisées (ISO 19115) ainsi que les fichiers de données (utilisation des « standard_name » de la convention CF). Pour la cytométrie le CDI XML basé sur l'ISO 19115 et INSPIRE sont utilisés. Le vocabulaire contrôlé est en cours de création pour la cytométrie.
Les logiciels du service observation de l'OSU sont interopérables et basés sur les standards de OGC. La cytométrie utilise les logiciels fournis par SDC (DM, NEMO, MIKADO).

Fonction « accès »

Pour la cytométrie l'accès aux données est effectué sur demande et après validation par les PI via le portail SDC : http://seadatanet.maris2.nl/v_cdi_v3/search.asp. Pour le service observation de l'OSU l'accès est public : http://erddap.osupytheas.fr/erddap. Il est cependant possible de protéger les jeux de données par un login/password.

Besoins techniques

Les baies de stockage et les serveurs sont renouvelés régulièrement.
Il est nécessaire de continuer le développement du portail d'accès de l'OSU Pytheas et la diffusion des données de cytométrie vers SDC, d'améliorer la chaîne de traitement de la cytométrie en flux (SDC ou ODATIS?) et de recoder l'intégration des données dans la bdd SQL.
Le plus gros besoin se situe sur les moyens humains.

3.12. IUEM (Laurence Lebourg)

LL présente un état des lieux des bases hébergées à l'IUEM en vue d'une création possible d'un CDS. Une ARN « flash » a été déposée pour faire un état des lieux des données hébergées dans les différentes unités de l'IUEM afin de rechercher l'ensemble des données orphelines. Une NOEMI a été accepté pour l'administration des bdd (en septembre).
L'IUEM héberge une partie des données des SNOs Dynalit, SOMLIT, Thytobs, COAST-HF, Benthos et Argo. Le service d'observation de l'OSU effectue des prélèvements de benthos, de chimie, des traits de côte et il est en lien avec la Zabri.

Le service réuni la production qui est effectué dans deux UMR (LETG et LGO) et s'occupe de l'ensemble des fonctions (« entrées », « stockage », « gestion des données », « administration », « planification de la pérennisation », « accès » et « utilisateurs »).

Besoin technique

Il est nécessaire de lever l'ambiguïté du SNO Dynalit au sujet de son rattachement soit à ODATIS, soit à THEIA.

4. Synthèse de la revue technique des CDS ODATIS

Il ressort à la suite des présentations des différents CDS qu'il est nécessaire :

- de mettre en place des échanges de données entre CDS pour faire de la sauvegarde croisée (sites distants),
- de mettre en place un serveur de vocabulaire (type BODC) à ODATIS,
- de mieux définir les workflows des SNO vers les OSU et des OSU vers ODATIS,
- de formaliser rapidement les recommandations techniques (via les ateliers),
- d'anticiper l'arrivée de nouvelles données très volumineuses en *in situ*.
- de faire une présentation d'ODATIS à Villefranche-sur-mer,
- de faire remonter les besoins en moyen humain aux tutelles (~ 5ETP à minima)

L'investissement matériel ne semble pas préoccupant pour le moment car il est pris en compte par les laboratoires, OSU, etc.

5. Préparation de l'atelier d'octobre 2019

Le prochain atelier se fera le 8 et 9 octobre à Marseille sur deux jours. Il sera consacré à la mise en place de recommandation sur le vocabulaire et aux début de la formation des participants à l'utilisation du NetCDF et de la plateforme JupyterHub. A noter que le 7 octobre une réunion se tiendra à Marseille pour discuter de la cytométrie.