Compte Rendu de l'atelier technique ODATIS du 11 et 12 Mars 2019

CR atelier technique Mars 2019

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

Numéro du délivrable	Titre court
	CR Atelier ODATIS Mars 2019
Titre long	
Compte rendu de l'atelier technique ODATIS du 11 et 12 Mars 2019	
Description courte	
Auteur	**Groupe de travail**
Joël Sudre	
Dissémination	**Copyright**
	Pôle Odatis

Historique

Version	Auteurs	Date	Commentaires
0.1	Joël Sudre	06 Mai 2019	Version initiale
0.2	Cécile Nys	14 mai 2019	Relecture et correction

Table des matières

1. Accueil et tour de table des participants

Liste des participants à l'atelier ODATIS :

- Mohamed Adjou (Geotraces, UK) – MA,
- Frédéric Briol (CLS) – FB,
- Guillaume Brissebrat (OMP/SEDOO) – GB,
- Pascal Calvat (OASU) – PC,
- Gérald Dibarboure (CNES) – GD,
- Wendy Diruit (CDD ODATIS) – WD,
- Mark Hoebeke (Station Biologique de Roscoff) – MH,
- Dimitry Khvorostyanov (LOCEAN) – DK,
- Stevenn Lamarche (Univ. Brest) – SL,
- Maurice Libes (OSU Pytheas) – ML,
- Didier Mallarino (OSU Pytheas) – DM,
- Gilbert Maudire (IFREMER) – GM,
- Fabrice Mendes (OASU) – FM,
- Caroline Mercier (AKKA/CNES) – CM,
- Richard Moreno (CNES) – RM,
- Catherine Schmechtig (OOV) – CS,
- Sabine Schmidt (Univ. Bordeaux / EPOC) – SS,
- Joel Sudre (OMP/LEGOS) – JS,
- Isabel Tardieu (Univ. Brest) – IT,

JS présente l'ordre du jour (*voir : Agenda et accès aux présentations*), en précisant que RM sera en visioconférence uniquement le second jour pour avoir les retours et avis du Directeur Technique de l'IR Système Terre sur les discussions et réflexions autour de l'architecture globale. Le vol de GM ayant du retard, le point d'information du pôle et de l'IR ST se fera après le point d'avancement et retour d'expérience ERDDAP et HYRAX.

Tour de table des participants.

Le compte-rendu de l'atelier de Novembre 2018 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 Novembre 2018 où la partie Hackathon s'est tenu sur des solutions ERDDAP et HYRAX, FB présente aujourd'hui un retour d'expérience sur les formats de données associés au stockage.

2.1. Formats de stockage : Utilisation versus distribution (Frédéric Briol)

FB présente le service de diffusion avancé avec un accès direct à la donnée proposé par le CNES pour l'utilisation de volume de données important (> 100 Go – Big Data – voir 01_Frederic_Briol_stockage.pdf).

Le format NetCDF est un format qui permet d'échanger des données cependant ce format pose un problème lorsque l'on veut traiter dans un mode parallèle (multi-processus) ou de calcul distribué la donnée qui y est stockée. En effet les données multidimensionnelles stockées en NetCDF ou HDF5 sont difficiles d'accès sur des plateformes de calculs distribués ou sur du stockage traditionnel des clouds à cause des accès concurrentiels. Il est donc nécessaire de différentier le format de stockage et le format d'échange. La présentation 01_Frederic_Briol_stockage.pdf passe en revue le type de données (tabulaire ou multimentionnel) ainsi que les outils permettant de les manipuler.

Si l'on veut se soustraire aux accès concurrentiels, il est nécessaire de concevoir un nouveau format. Il existe actuellement des solutions clés en main (format propriétaire) tel que Rasdaman, Thredds, SciDB mais qui sont des solutions commerciales assez onéreuses ou des solutions de type SPARK, YARN, HADDOP mais qui manquent de souplesse par rapport à la diversité et la complexité des cas réels d'utilisation des données en géoscience.

Le format Zarr semble être une alternative intéressante au NetCDF4 pour le stockage interne. De plus ce format est intégré dans la pile Pangeo.

2.2. Tour de table

Deux points ont été abordés au cours de ce tour de table :

Le premier point concerne le cheminement de la donnée entre les CDS, ODATIS et l'Europe. Pour le moment ODATIS n'est pas encore une entité reconnue au niveau de l'Europe, seul certains CDS sont enregistrés au niveau européen. L'un des objectifs d'ODATIS est de devenir l'interlocuteur privilégié avec l'Europe pour porter à l'Europe un discours cohérent et homogène de tous les CDS. Actuellement nous sommes donc dans une période de transition qui va encore durer environ 2 ans.

Le second point abordé lors de ce tour de table concerne les données de profondeur

intermédiaire du SOMLIT. Actuellement dans le contrat du SOMLIT avec l'INSU seules les données de surface et les données à la plus grande profondeur sont distribuées. Cependant il existe des données à des profondeurs intermédiaires (ex : données de Villefranche sur Mer) dans le réseau SOMLIT qui ne sont pas dans ce contrat de distribution et qui ne sont donc pas distribuées par le site du SOMLIT. Ceci pose donc la question du devenir de la distribution de ces données. Cette question va être posée lors du prochain Conseil Scientifique du SOMLIT qui se tiendra mercredi 20 mars.

3. Point d'information sur le pôle ODATIS et l'IR Système Terre (Gilbert Maudire)

GM présente (voir 02_Gilbert_Maudire_IR_Système_Terre.pdf) tout d'abord les points d'information sur l'IR Système Terre et ensuite sur le Pôle ODATIS.

3.1. IR Système Terre

L'IR Système Terre recherche toujours un nom et s'oriente vers Data Terra.

Une UMS de soutien à l'IR et aux pôles a été créée le premier janvier 2019. Cette UMS porte le numéro 2013 à l'INSU et elle est composée du directeur de l'IR (F. Huynh), du directeur technique (R. Moreno), des directeurs de pôles à temps partiel. Cette UMS va permettre de mettre en place du personnel « en transversal » pour les pôles et l'IR.

Une nouvelle structure transverse a aussi été mise en place pour avoir un accès unique à l'imagerie haute résolution et pour fédérer l'ensemble des contrats et conventions avec les différentes missions. Cette nouvelle structure : Dinamis va permettre à la communauté scientifique d'avoir accès gratuitement à l'ensemble du catalogue de données images à haute résolution (ex : SPOT, Sentinel 3,...).

Deux groupe de travail (GT) ont été créés au niveau de l'IR :

- le GT technique qui va implémenter les concepts étudiés par le GT Interpôles. Ce groupe est coordonné par RM.
- le GT « Communication » coordonné par F. Hyunh qui a pour but d'homogénéiser les sites web des pôles et de développer le site web de l'IRST. Caroline Mercier et Cécile Nys (qui succède à Linn Sekund) représentent le pôle ODATIS à ce GT. Les pôles Aeris, Form@ter, et

Theia utilise l'offre de l'OMP Sedoo pour leur site web, cependant, ODATIS ayant fait une refonte de son site web en 2018 va rester sur sa solution propre.

3.2. Appels à projets

Le Pôle ODATIS avec l'IR dans son ensemble a répondu à plusieurs appels d'offres européens : EnvriFAIR (en janvier), Phidias, EOSC-Pillar. A noter que EOSC-Pillar a d'ores et déjà été accepté. Le pôle ODATIS a aussi répondu en son nom propre à Blue Cloud et à Sea Data Cloud 2.

3.3. Les Consortiums d'Expertise Scientifique (CES)

Le Conseil Scientifique d'ODATIS a lancé deux Consortiums d'Expertise Scientifique (CES) :

- CES Couleur de l'eau et données *in-situ* associées,
- CES Oxygène.

Un CES Salinité sera lancé ultérieurement car il n'est actuellement pas raisonnable au niveau financier de constituer un troisième CES.

3.4. Agenda Odatis

Les prochaines échéances du pôle ODATIS sont :

- Réunion du CES Couleur de l'eau le 28 et 29 mai 2019,
- Réunion BE plénier le 6 et 7 juin,
- Réunion du CES Oxygène le 2 et 3 juillet.
- La date du Comité Directeur n'a pas encore été décidé (fin juin, début juillet ou en septembre)

A noter que SS et JS animeront une table ronde sur le cycle de vie de la donnée au AEI 2019 qui se dérouleront du 9 au 11 juillet 2019 à Lille.

En 2019 certains points prioritaires sont à étudier, en particulier les relations et interfaces avec l'IR d'observation côtière ILICO et les relations avec le « Système d'Information sur le Milieu Marin (SIMM) » (MTES) avec la demande de participation de l'Agence Française pour la Biodiversité.

4. Présentation du Centre de Données et Services OASU

L'atelier technique ODATIS étant organisé à Bordeaux pour cette session, il a été demandé à l'OASU de présenter son CDS et son fonctionnement interne.

4.1. CeDONA, Centre de Données pour l'Observation en Nouvelle-Aquitaine (Pascal Calvat)

PC présente le centre de données pour l'observation en Nouvelle-Aquitaine (voir 03_Pascal_Calvat_CEDONA_OASU.pdf) qui a pour mission de renforcer la visibilité nationale et internationale des projets portés par les équipes des laboratoire de l'OASU, d'assurer la cohésion des services dans un contexte de recherche multidisciplinaire et de répondre aux standards internationaux de mise à disposition des données. Le CeDONA procure aussi des services à valeur ajoutée sur les données issues des équipes de recherche de l'OASU.

Depuis 2003, L'OASU est constitué de 3 UMR (LAB, EPOC, LIENSs), de 2 UR (ETBX, EABX) de l'IRSTEA (Institut national de Recherche en Science et Technologie pour l'Environnement et l'Agriculture) et de l'UMS POREA. L'OASU possède aussi des Services d'Observation labellisés par l'INSU en Astronomie – Astrophysique (IVS, ALMA, JUICE, SKA, GAIA, KIDA), en environnement (SOMLIT, DynaLIT, Photons), le SOERE Trait de Côte (ALLENVI) ainsi que des SO qui ne sont pas encore labellisés INSU (SOLAQUI, MAGEST, MOLUSCAN eye).

Le CeDONA propose une ou deux fois par an des appels à projets pour l'ensemble des unités de l'OASU afin d'offrir ses services. Cet appel à projets est évalué par le CS de l'OASU qui attribue un pourcentage d'ETP (1 développeur permanent + 1 prestataire) suivant les spécifications fonctionnelles (obligatoire) pour réaliser le projet. Le CeDONA propose des services informatiques tel que :

- développement et déploiement de site web,
- aide à l'utilisation de services nationaux de diffusion de données,
- backup et/ou réplication de données,
- versionning de codes (gitlab),
- calcul en local et sur le meso-centre MCIA,
- accompagnement pour la rédaction de PGD (Plan de Gestion de Données).

4.2. Les projets « données » à l'OASU et choix techniques (Fabrice Mendes)

Suite à la présentation du CeDONA par PC, FM présente les projets « données » à l'OASU qui sont associés au SNO (voir 04_Fabrice_Mendes_CDS_OASU.pdf).

Le CeDONA gère plusieurs développements spécifiques pour des projets de l'OASU en rapport avec les SNO d'Astronomie /Astrophysique et les projets d'Océanographie (SNO Côtier, Trait de côte et sédimentologie, SNO SONEL, SNO PAMELI, BDD Benthos). Ces développements sont de plusieurs types : développement web, de BDD, de gestion de données et de métadonnées, amélioration des sorties graphiques sur le web (haute fréquence), etc.

A noter, qu'il y a eu récemment la mise en place d'un backup des bases avec le LEGOS/OMP. En terme de perspective à moyen et court terme, le CeDONA va travailler sur la mise en place de DOI, de DMP et sur « FAIRisation » de ses données ainsi que sur l'amélioration des codes par des tests. Le CeDONA va aussi mettre en place un GeoNetwork pour la découverte de ses données et un outil spécifique du type ERDDAP à intégrer dans son environnement pour être moissonnable par le pôle ODATIS.

5. PHYTOBS : Portail d'accès aux données et interopérabilité avec la base PELAGOS (Mark Hoebeke)

MH présente le nouvelle SNO Phytobs qui est un réseau élémentaire de l'IR ILICO dédié à l'observation du phytoplancton (voir 05_Mark_Hoebeke_Phytobs.pdf). Ce SNO fédère des activités qui étaient menées auparavant dans le cadre du REPHY et de SOMLIT. Les principaux acteurs sont le CNRS, les Universités Marines et l'IFREMER en partenariat avec les Agences de l'eau. Maud Lemoine (IFREMER) et Pascal Claquin (Université de CAEN, UMR BOREA) sont les coordinateurs du SNO Phytobs qui vient d'être labellisé. Le SNO s'est doté d'un comité de pilotage constitué des coordinateurs, des référents des Bdd, d'experts scientifiques et des représentants des tutelles.

Dans un contexte de l'anthropisation généralisée des eaux côtières, les objectifs du SNO Phytobs sont entre autres:

- Analyse des niches écologiques et des habitats,
- Détection des variations de phénologie,
- Caractérisation des traits et des groupes fonctionnels,

- Étude des relations biodiversité/productivité.

La méthodologie, les données associées ainsi que la description du portail d'accès du SNO Phytobs sont décrites dans la présentation.

6. DMP ODATIS avec use case mouillage de Clipperton (Joël Sudre)

Lors de cet atelier une première approche du DMP (Data Management Plan) ou PGD (Plan de Gestion de la Donnée) est présentée par JS (voir 06_Joel_Sudre_DMP_Clipperton.pdf).

Le DMP permet de formaliser le cycle de vie de la donnée (scientifique) en répondant aux questions suivantes :

- Qui fait quoi ?;
- À quel moment ?;
- Où ?;
- Comment ?;
- Avec quels moyens ? (financier, RH, matériel).

Cela permet d'avoir un document unique pour décrire comment les données sont obtenues, documentées, analysées, disséminées, et utilisées. Il est à noter que ce type de document devient de plus en plus obligatoire dans les appels d'offre européens (H2020, ERC, …) et que l'ANR va le rendre obligatoire prochainement pour s'assurer que la donnée ne soit pas perdue, non disséminée et/ou non pérennisée.

Un DMP n'est pas un document figé, il doit être descriptif, prospectif et évolutif dans le temps afin de permettre de générer un cycle efficace, complet mais aussi d'améliorer l'accès à la donnée (FAIR).

Le DMP va donc concerner tous producteurs et/ou distributeurs de donnée voulant décrire en détail leurs données qui seront exploitées ou rendues accessibles.

Il existe plusieurs modèles de DMP suivant qu'il soit adressé aux financeurs, aux établissements, etc. (voir OPIDor de l'Inist qui regroupe plusieurs modèles : H2020 FAIR, CIRAD, ERC, INRA,…). En 2017, IR ILICO a demandé par le biais d'un questionnaire un DMP à tous les SNO entrant dans son cadre.

Discussion suite à la présentation de JS :

Une discussion s'est engagée à la suite de la présentation de JS portant sur le périmètre et la granularité à donner aux DMP. En particulier de nombreuses questions restent en suspens sur plusieurs points pour définir le périmètre de IR ILICO et du pôle ODATIS (et de IR Système Terre).

Le DMP abordant le cycle de vie entier de la donnée, il est nécessaire de faire la distinction entre les attentes de IR ILICO qui s'intéresse plus à la partie conception et réalisation de la donnée et le pôle ODATIS qui se focalise plus sur la partie publication, conservation, préservation, diffusion et réutilisation de la donnée. Il en ressort qu'il est nécessaire que IR ILICO et le pôle ODATIS doivent échanger à ce sujet afin de créer un DMP intégrant mieux les différents rôles de chaque entité.

Les échanges ayant été très nombreux à la suite de cette partie de la présentation, la partie exemple d'un DMP sur le mouillage de Clipperton n'a pas été abordée faute de temps.

7. Compte-rendu et synthèse sur le cahier des charges des CDS (Gilbert Maudire)

GM présente le compte-rendu et la synthèse sur le cahier des charges des CDS (voir 07_Gilbert_Maudire_Cahier_Charges_CDS.pdf).

Le cahier des charges d'un CDS va permettre de définir au sein du Pôle ODATIS les tâches à accomplir par le CDS et son interfaçage avec le pôle. Le CDS a toute liberté de la manière de les mettre en œuvre. Ce document va permettre, pour les responsables de CDS, de se projeter sur les fonctions à mettre en place et de justifier ses besoins matériels et humains. Ce document va aussi servir à mieux planifier les coûts complets, à mieux planifier la disponibilité de nouveaux services et la mise à jour du catalogue du Pôle.

Les fonctions d'un CDS (Fig 1) sont calquées sur le modèle de référence pour un système ouvert d'archivage d'information dit modèle OAIS (voir modèle OAIS), avec des adaptations car un CDS ne fait pas que de l'archivage.

Figure 1 : Les 6 fonctions du modèle OAIS (source CINES)

La **fonction « entrées »** reçoit, contrôle et valide les données à gérer, soit :

- Une harmonisation des données du même type provenant de différentes équipes et/ou de différents systèmes d'observation : utilisation de référentiels communs, d'une même codification de variables et paramètres, d'une structuration homogène.
- Un contrôle qualité des données reçues, quand cela est préconisé par les programmes d'observation et de gestion de données pour ce type … A minima, un contrôle de complétude.

La **fonction « stockage »** assure la conservation physique des données gérées. Il s'agit au sens du Pôle de la copie primaire, dite de référence, des données. Elle s'appuie sur la fonction de pérennisation pour en assurer la préservation et la sécurité.

La **fonction « planification de la pérennisation »**, assure la pérennisation physique des informations (réalisation de copies multiples, renouvellement des supports anciens, transition vers de nouvelles génération de supports/lecteurs informatiques, etc.) soit en mettant en œuvre ses moyens propres, soit en délégant, conventionnant et supervisant cette tâche auprès d'un service externe.

La **fonction « gestion des données »** permet la tenue à jour des informations nécessaires à la gestion et à l'utilisation ultérieure des données:

- La description des données gérées, en support aux fournisseurs de données,

conformément à la définition du catalogue du pôle (métadonnées à la norme ISO 19115 par exemple, utilisation des vocabulaires communs définis). Ces descriptions doivent notamment comprendre : l'origine de la donnée (fournisseur individuel ou système d'observation), les conditions techniques et organisationnelles d'observation, la liste des variables observées, les conditions et limites d'utilisation des données (licences d'utilisation, domaines d'applicabilité, etc.) ;

- La description des résultats des chaînes de traitement conformément à la définition du catalogue du pôle ;
- La gestion des droits d'accès et des conditions d'accès aux données, conformément à la réglementation en vigueur et en accord avec les producteurs de données : période d'exclusivité à l'équipe d'observation, etc.

La **fonction « administration »** assure la coordination générale du système. Elle veille à la qualité globale du service rendu et à son amélioration. Elle rend compte au management et au Pôle : évolution des données gérées, interruptions de services éventuelles, … :

- Mise en œuvre des évolutions demandées par le Pôle ;
- Les moyens techniques informatiques, soit en propre, soit par délégation, conventionnement ou contractualisation, conformément aux exigences de service ;
- Le maintien en conditions opérationnelles, le maintien en configuration (tests, gestion des versions, …), les passages en exploitation des logiciels utilisés par le CDS : outil de formatage et de contrôle des données, chaines de traitement, etc. ;
- Production des statistiques d'activité (production, utilisation) qui seront fournies aux instances du Pôle.

La **fonction « accès »** regroupe tous les services qui sont en interface avec le Pôle soit directement avec les utilisateurs, soit par délégation auprès d'une infrastructure externe partagée. Outre les fonctions de contrôle d'accès, il s'agit d'opérer en continu, avec un taux de disponibilité opérationnel, les services d'accès et/ou de téléchargement des données gérées.

Les conditions d'exercice de ces fonctions doivent respecter un contexte général prenant en compte les principes FAIR de la donnée mais aussi tous les aspects réglementaires (droits des données, Open Data, EtatLab, Creative Commons, etc.

Suivant les données gérées, certaines tâches sont à mener en continu, en routine et selon une périodicité appropriée. Pour assurer ces tâches un contrat d'« engagement de service » dérivées de celles édictées par la certification de la RDA peut être mis en place.

La disponibilité des services en ligne est un point critique dont l'objectif est d'assurer une disponibilité 7j/7 et 24h/24 ! Cette objectif étant extrêmement ambitieux et quasiment impossible à obtenir pour un CDS (car extrêmement coûteux), l'idée est de soulager les CDS des tâches opérationnelles en fédérant les données (au moyen d'un cache technique par exemple) en amont des services de diffusion qui seront mis en commun. Pour ce faire, deux types de CDS sont envisagés :

- Les Centres d'Assemblage et de Traitement (CDS-CAT), consacrés aux tâches de gestion de données et de génération de produits, mais délégant la diffusion des données ;
- Les Centres d'Assemblage, de Traitement et de Diffusion (CDS-CATD), qui outre les tâches de gestion de données, proposent également des services opérationnels de diffusion. Les Centres d'Assemblage gérant des données de volumes importants, pouvant difficilement être répliquées par le réseau internet, doivent se conformer à ce niveau de service.

L'interfaces CDS/ Pôles pour l'«accès aux données» se fera donc de deux manières différentes suivant le type de CDS :

- Pour CDS-CAT, une réplication des données vers le cache technique avec un niveau de disponibilité de type « Best Effort », (ex : Moissonnage des métadonnées + un service de type ERDDAP, interrogé une fois par jour ouvrable),
- Pour les CDS-CATD, une mise en œuvre de services en ligne avec un taux de disponibilité élevé (95 – 98%) pour les services de téléchargement, subsetting, visualisation etc.

8. Uses case ODATIS et IR Système Terre

Au sein du pôle ODATIS, l'Ifremer et le CNES sont des CDS ayant de gros volumes de données satellitales et aussi pour l'Ifremer des données *in-situ*. Cette diversité des données permet de mettre en place des cas d'étude pour tester au niveau infrastructure, calcul, logiciel, des solutions qui pourraient être mises en place au niveau nationale à l'IR Système Terre, dans les centres de

calcul régionaux, dans les laboratoires et OSU hébergeant des CDS ayant des volumes de données moins importants.

L'idée ici, est de tester des solutions qui permettent à un scientifique de travailler sur des gros volumes de donnée à distance car il n'est pas possible au vu des volumes de pouvoir recopier localement ces données dans leur intégralité. Il est possible, par contre, de faire un sous-ensemble local où les algorithmes peuvent être testés et ensuite lorsque les algorithmes et la chaîne de traitement locale associée sont opérationnels de les porter sur des centres de calcul ayant la même suite logiciel et pouvant les déployer sur des volumes de donnée très importants voire de façon synoptique.

8.1. Thredds, ERDDAP, Hyrax à l'Ifremer (Gilbert Maudire)

Au niveau des services déployés à l'Ifremer (voir 08_Gilbert_Maudire_Use_Case_Ifremer.pdf), Thredds est un service opérationnel depuis 5 ans avec 150 jeux de données dynamiques, 100 jeux de données statiques et une sortie en WMS utilisée pour la visualisation via le catalogue Sextant.

ERDDAP a été mis en place depuis 18 mois afin de diffuser les agrégations d'observations. Cette solution est jugée très satisfaisante et convient parfaitement au jeux de données hébergés à l'Ifremer. Actuellement, un petit groupe de l'Ifremer travaille sur une solution Hyrax. Ce groupe ne préconise pas de passer d'ERDDAP à Hyrax pour le moment. Il n'est donc pas prévu de tester plus en avant une solution Hyrax, car les tests révèlent le même type de problème qu'avec ERDDAP (souci avec un grand nombre de fichiers, obligation de travailler sur des agrégats pour limiter le nombre de fichiers, matrices creuses, etc.), mais plutôt de permettre à d'autres utilisateurs potentiels de leur fournir un support technique pour l'installation, la configuration et l'utilisation de ces deux solutions : Threeds et ERDDAP (exemple : déploiement pour les flotteurs Argo).

Il est à noter que la problématique des matrices creuses, et de l'optimisation des fichiers sont des sujets qui sont abordés depuis 2 ans via Sea Data Net 2.

Autour de la structuration des données volumineuses, un test a été effectué sur Datarmor avec des données synthétiques SWOT avec une conversion des données au format Zarr. Ce test a saturé les I/O du calculateur, ce qui indique que pour le moment cette solution n'est pas optimale pour le HPC datamor dans sa configuration actuelle.

Actuellement un cas d'étude de « machine learning » est testé sur les donnée Argo, les premiers résultats font état que le NetCDF natif est « très lent » pour les ouvertures/fermetures des fichiers. La solution la plus satisfaisante en terme d'accès et d'utilisation des données est de les placer dans

la base noSQL Cassandra.

8.2. Use cases démonstrateur CNES pour Odatis et SWOT (Gérald Dibarboure)

GD présente en préambule la mission SWOT et pourquoi cette mission est un démonstrateur idéal pour tester de nouvelles solutions (voir 09_Gerald_Dibarboure_Use_Case_Cnes.pdf). Un changement de paradigme est nécessaire au vu des volumes importants de données que va engendrer le lancement de cette mission dans deux ans (entre 1 à 10 To/jour pour les données océaniques). Il y a donc deux besoins :

- Pour ODATIS de faire une diffusion centralisée avec des fonctions avancées,
- Pour SWOT de permettre le calcul déporté à proximité de la donnée.

Dans ce cadre, le Logiciel SEASCOPE développé par Ocean Data Lab en open source permet d'explorer et de visualiser des données sous toutes les plateformes (windows, linux, mac). Ce logiciel est compatible avec la pile PANGEO et est capable de traiter des données spatiales et *in-situ* volumineuses (Seascope).

L'objectif du démonstrateur est de faire des tests pratiques et pragmatiques, de tester des technologies et des logiciels, de prototyper les fonctionnalités principales sur quelques exemples et dès que possible de proposer des services concrets à des scientifiques désirant essayer ses nouvelles fonctionnalités.

Pour cela, une première étape va permettre de déployer pour les données satellitales cette solution sur l'HPC du CNES (Hal) associé à un Cloud hébergeant ces données. Ensuite, dans une seconde étape, de déployer cette solution sur un second HPC (Datarmor et/ou CINES) distant hébergeant les données *in-situ*, ce qui permettra de tester la capacité de bascule sur un autre centre de données. Ce cas d'étude nécessite de faire aussi des tests sur des volumes réduits de données provenant d'autres CDS avec leurs contraintes de formats et de types de données, de développer des mécanismes de synchronisation avec ces CDS, d'installer et de configurer des solutions de type (DAP+WFS, WMS+WCS) sur les HPC.

Pour la pile logiciel PANGEO associée à SWOT, l'objectif est de générer et de cataloguer un produit simulé SWOT en attendant le lancement du satellite, d'activer un algorithme sur un an de données et d'utiliser le logiciel Seascope pour visualiser les résultats.

Pour la pile logiciel PANGEO associée à ODATIS, l'objectif est de prendre une série de données *in-*

situ dans un CDS (à déterminer), de co-localiser dynamiquement ces données avec des données altimétriques HF (20Hz) qui représentent un volume de données de 100 To. Il est aussi prévu de faire de la co-localisation avec des données provenant de Sentinel-1 (PEPS) et d'un diffusiomètre dont les données sont hébergées au CERSAT. In fine de visualiser l'ensemble sur le logiciel Seascope.

Discussion et tour de table

Au vu des deux présentations de GM et GD, il est nécessaire de poursuivre les tests sur les deux HPC avant de pouvoir identifier de nouveaux besoins et de nouvelles demandes car il est encore trop tôt pour pouvoir insérer de nouveaux cas d'étude. L'ouverture vers de nouveaux cas d'étude se fera plutôt en 2020.

Actuellement, l'évolution des nouvelles technologies permettant de faire du calcul à distance est très rapide et n'est pas encore mature et stable pour un déploiement général. Il est donc nécessaire de continuer à tester différentes solutions avant de décider de la meilleure option.

9. Discussion et réflexion autour de l'architecture globale

9.1. Architecture du Pôle ODATIS (Gérald Dibarboure)

Suite aux précédentes présentations, il est demandé à l'ensemble des participants de l'atelier ODATIS de faire des retours sur les différents démonstrateurs.

Le choix de l'architecture « finale » n'est pas encore décidé et doit encore être évalué. De nombreuses questions n'ont pas encore trouvé de réponses :

- Comment arriver à avoir un système qui permet une disponibilité supérieure à 98 % pour l'ensemble des CDS ? (Objectif à atteindre en 2021)
- Comment mettre en place une authentification avec deux niveaux (faible de type réseaux sociaux et forte du type Renater ou ORCID) ? Ceci nécessite de mettre en place des cercles de confiance sur les données. Comment se coordonner avec les SSO (Sigle Sign-On, authentification unique) des autres pôles Theia, Aeris. Va-t-on utiliser des fournisseurs d'identité (ORCID, Shibboleth, BE2, WSO2, etc.) ?
- Doit-on utiliser un catalogue local pour chaque CDS ou un catalogue web global ? Est-ce que l'on va continuer avec le catalogue Sextant pour le pôle ODATIS ?
- Pour les données, les tests montrent la nécessité d'avoir à minima deux copies, une en

format natif pour la diffusion, une pour un accès à des services de calcul ? Doit-on doubler le volume ?

- Comment utiliser des données tierces de très gros volumes (ex : données satellitales de la NASA) ? Où vont-elles être stockées ? Quels sont leur volume ? Où va-t-on mettre les services qui vont leur être adossées ? Quelle est la position du MESRI à ce sujet ?
- Où va-t-on implémenter les outils qui vont permettre de faire des statistiques d'usage sur les données ?
- Pour des fonctionnalités avancées permettant de croiser des données des différents pôles, est-ce que l'IR système Terre va se reposer sur l'infrastructure des pôles ou sur une infrastructure centralisée ?
- Comment mettre un helpdesk centralisé pour l'ensemble des pôles et qui permet de faire redescendre la demande d'information vers les pôles et les CDS concernés ?
- Est-ce que l'on met en place un système avec un centre de données national ? Est-ce que l'on se repose sur les centres régionaux, ce qui implique un réseau bien dimensionné en terme de flux ?
- Les pôles ayant depuis 2016 pour mandat l'archivage pérenne des données, qui va s'occuper de l'archivage pérenne sachant que seuls les CDS sont capables de dire si un jeu de données doit-être sauvegardé ou pas et qu'un format de données est obsolète? Doit-on sauvegarder la donnée, et/ou la chaîne de traitement qui a permis de générer la donnée ?

L'ensemble de ces questions va devoir être abordé aux niveaux des CDS, du pôle et de l'IR Système Terre tout en tenant compte des contraintes provenant de l'international !

9.2. Architecture de l'IR Système Terre (Richard Moreno)

RM présente l'IR Système Terre et son rôle pour fédérer les pôles de données (voir 10_Richard_Moreno_Data_IR_Systeme_Terre.pdf). L'infrastructure de recherche regroupe 4 pôles de donnée (bientôt 5 avec le PNDB), plusieurs initiatives transverses (Dinamis, le comité technique Interpôles, un groupe de travail « Europe »), 34 partenaires dont 8 composent son bureau exécutif (CNRS, CNES, IFREMER, IGN, IRD, IRSTEA, Météo-France, MESRI). En janvier 2019, une unité mixte de services (UMS 2013 – Data Terra) a été créée pour lui donner une structure opérationnelle. Plusieurs groupes de travail y ont été créés :

- GT Europe,
- GT technique,
- GT communication.

Ces groupes permettent de prendre des décisions de manière collégiale avec l'ensemble des partenaires et des CDS.

L'IR ST participe activement à différentes initiatives européennes (H2020-EOSC-PILLAR, FP9, ESFRI, PHIDIAS, ENVRIFair, Copernicus,…) et internationales (GEO/GEOSS, ONU-Env, GO FAIR). Les données que va devoir prendre en compte l'IR ST sont très diverses suivant les pôles. Elles n'ont pas le même niveau de « FAIRitude », peuvent aussi provenir de l'Europe mais aussi de l'internationale (NASA, NOAA, USGS, JAXA,…). Ces données sont à la fois des données satellitales, *in-situ* et aussi synthétiques provenant de modèles. Elles sont distribuées de diverses manières par des centres de donnée distribués sur l'ensemble de la planète. De plus à la différence des pôles, l'IR ST est mandaté pour fournir le secteur aval en plus du secteur recherche.

Afin de mettre en place une telle structure nationale, il est possible de s'inspirer de modèles développés hors de nos frontières comme NASA EOSDIS, le HUB Pangeo, GeoDAB, EuroGEOSS, AmeriGEOSS, NextGEOSS, ESA DCB. Il est aussi nécessaire de prendre en compte que le volume des données est en incessante augmentation voire en augmentation exponentielle avec l'apparition de données haute résolution, haute fréquence et la multiplication des plateformes supportant les capteurs (drones, réseau d'observation via des logiciels sur smartphone, etc.).

Le pôle ODATIS a plusieurs actions qu'il mène en partenariat avec IRST :

- Identification des tâches et des ressources en lien avec IRST,
- L'initiative SWOT,
- Eosc Pillar,
- Phidias,
- Implication dans EnvriFAIR,
- Appel d'offre ANR-Flash.

Pour 2019, les objectifs principaux du pôle vont être :

- de faire un état des lieux « technique » de chaque CDS,
- de poursuivre l'enrichissement du catalogue,
- de mettre en place dans chaque CDS un service web et un service de découverte des données (voire de visualisation de celles-ci) qui peut être moissonnable par le site web du pôle ODATIS (visualisation de la position de chaque donnée, visualisation de la donnée, téléchargement, etc.),
- de réfléchir sur la mise en place de l'authentification permettant par exemple un moratoire d'exclusivité de certaines données, la mise à disposition de données « sensibles » ou confidentielles pour un groupe de personnes, etc.,
- de faire un prototypage de l'architecture future du pôle aux moyens de cas d'étude,
- de mettre en place la charte entre le pôle et chaque CDS.

10. Préparation du second atelier 2019

Contrairement à ce qui avait été décidé lors de cet atelier et afin de prendre en compte le BE ODATIS qui se tiendra les 6 et 7 Juin à Paris, le prochain atelier se fera le 5 Juin à Paris sur un jour. Il sera consacré à une revue technique de ses CDS.