Compte Rendu de l'atelier technique ODATIS du 20 et 21 Novembre 2018

CR atelier technique Nov 2018

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

Numéro du délivrable	Titre court
	CR Atelier ODATIS Nov 2018
Titre long	
Compte rendu de l'atelier technique ODATIS du 20 et 21 Novembre 2018	
Description courte	
Auteur	**Groupe de travail**
Joël Sudre	
Dissémination	**Copyright**
	Pôle Odatis

Historique

Version	Auteurs	Date	Commentaires
0.1	Joël Sudre	08 Décembre 2018	Version initiale
0.2	Jérôme LLido	11 Décembre 2018	Relecture et correction
0.3	Maurice Libes	12 Décembre 2018	Relecture et précisions
0.4	Frédéric Briol	12 Décembre 2018	Relecture et précisions

Table des matières

1. Accueil et tour de table des participants

Liste des participants à l'atelier ODATIS :

- Karim Bernardet (CNRS- DT INSU) – KB,
- Aurélie Briand (IFREMER) – AB,
- Frédéric Briol (CLS) – FB,
- Guillaume Brissebrat (SEDOO) – GB,
- Pascal Calvat (OASU) – PC,
- Gérald Dibarboure (CNES) – GD,
- Mark Hoebeke (CNRS – SBR) – MH,
- Dimitry Khvorostyanov (LOCEAN) – DK,
- Maurice Libes (OSU Pytheas) – ML,
- Didier Mallarino (OSU Pytheas) – DM,
- Gilbert Maudire (Ifremer) – GM,
- Fabrice Mendes (CNRS – OASU) – FM,
- Catherine Schmechtig (CNRS – OOV) – CS,
- Joël Sudre (CNRS -LEGOS) – JS,

JS présente l'ordre du jour (*voir : 00_Joel_Sudre_agenda_atelier_odatis_nov_2018.pdf*), en précisant que la partie Hackathon sur les solutions ERDDAP se fera entièrement le premier jour pour consacrer la deuxième journée à l'architecture globale du pôle ODATIS.

Tour de table des participants.

Le compte rendu de l'atelier de Octobre 2018 a été approuvé et mis en ligne sur le site ODATIS avec un accès publique.

2. Hackathon sur les solutions DAP pour les CDSs

Suite à l'atelier technique d'Octobre 2018 où la partie Hackathon n'avait pu être effectuée faute d'animateurs présents (et excusés), l'hackathon s'est tenu lors de cet atelier. Des solutions alternatives ayant étaient évoquées lors de l'atelier précédent, une machine virtuelle a été installée pour tester le logiciel Hyrax en alternative à ERDDAP. Les sous-parties suivantes rassemblent le

contenu de l'hackathon sans suivre le déroulement chronologique de cette journée dédiée.

2.1. Présentation ERDDAP (Maurice Libes et Didier Mallarino)

ML présente une démonstration sur la machine virtuelle configurée avec le logiciel ERDDAP. ERDDAP propose un accès à la donnée brute, des recherches avancées sur un jeu de données avec la possibilité d'effectuer des requêtes permettant :

- d'accéder aux données,
- de créer des graphes simples avec le jeu de données.
- d'extraire tout ou partie des données d'un jeu de données (via une sélection sur les paramètres et attributs inclus dans le jeu).
- de générer un flux WMS.

ERDDAP est donc une solution qui offre la possibilité d'accéder à la donnée de façon simple et qui permet de rendre FAIR un jeu de données.

2.2. Retour d'expérience ERDDAP à l'Ifremer (Aurélie Briand)

AB présente un retour d'expérience de la solution ERDDAP à l'Ifremer (voir 01_Aurelie_Briand_erddap_ifremer.pdf). Ce logiciel a été mis en œuvre pour les jeux de données Argo tels que le jeu Argo T et S ou les Argos Bio. Les produits Coriolis (CORA, NRTOA), les produits SeaDataNet (Climatologies T et S), ainsi que le jeu de données de Gliders ont aussi été mis à disposition via ERDDAP. Le lien suivant permet d'accéder à ces jeux de données :

http://www.ifremer.fr/erddap/info/index.html?page=1&itemsPerPage=1000

ERDDAP étant une application RESTFUL (REpresentational State Transfer), elle autorise de mettre en place une URL qui permet de récupérer de la donnée, des graphes ainsi que des informations sur le jeu de donnée sélectionné.

Le jeu de donnée ARGO représentant environ ~12000 fichiers « multi-profiles » distribués dans 11 centres Argo différents des échanges avec Bob Simons (BS) de la NOAA (National Oceanic and Admospheric Administration), créateur d'ERDDAP, se sont engagés pour mettre en place une

nouvelle façon d'agréger les fichiers. Il est à noter que BS est très réactif et fait évoluer ERDDAP en fonction des demandes des utilisateurs. Suite à ces échanges, BS a modifié son logiciel afin de prendre en compte les spécificités du jeu Argo (voir les transparents 5,6 et 7 de la présentation pour voir en détail les améliorations apportées à ERDDAP dans ce cas d'étude - amélioration de la vitesse de chargement de la page web, prise en compte des « missing_value », « _FillValue », « direction », « data_mode » et « *_qc variables »).

Ce lien permet de visualiser un exemple d'ERDDAP sur les données Argo.

En ce qui concerne les jeux de données des produits Coriolis CORA, NRTOA ainsi que la climatologie SeaDataNet, ils ont été configurés sans souci mis à part un message d'erreur non bloquant (qui a été traité par la suite) sur le dernier jeu cité.

Suite au retour d'expérience sur ERDDAP, AB présente deux outils de visualisation :

- l'outil DIVAA (voir présentation 02_Aurelie_Briand_divaa_ifremer.pdf) qui est une application web développée par Kevin Balem (IFREMER/LOPS - voir map.argo.france.fr pour avoir un aperçu de cet outil). DIVAA a été développé en utilisant le flux geojson d'ERDDAP. Il a été écrit en Java et est un logiciel opensource téléchargeable en suivant ce lien sur github : https://github.com/quai20/DIVAA,
- l'outil Argo visualisation 3D qui est aussi une application web permettant la visualisation 3D des profils Argo en environnement profond. Cette application a été développée par Anthonin Lize du JCOMMOPS. Elle utilise le flux json ERDDAP Argo.

Discussion suite à la présentation de AB :

Une discussion s'est engagée à la suite des présentations d'AB sur la nécessité d'adosser un module d'authentification dans ERDDAP pour pouvoir permettre d'identifier et de distribuer des données qui nécessitent une distribution restreinte (données avec embargo, données en cours de qualification, etc.). GM rappelle qu'un groupe de travail au niveau de l'inter-pôles a travaillé sur ce sujet et qu'actuellement il est recommandé d'utiliser la solution B2ACCESS de l'EUDAT pour faire de l'authentification au niveau des pôles de données et de leurs CDSs associés. Cette solution permet à la fois d'utiliser les identifications institutionnelles, mais aussi des identifiants génériques comme celui du compte Google, et il permet aussi de créer un identifiant EUDAT si une personne n'a aucun des identifiants cités précédemment. L'utilisation de ce type d'authentification permettra à terme de définir des cercles de confiance plus ou moins restreints en fonction d'un jeu de

donnée (ex : cercle de confiance pour la liste des participants à une campagne à la mer sur un jeu de données biogéochimiques en cours de calibration et de qualification).

2.3. Présentation et Hackathon HYRAX (Frédéric Briol)

FB présente une machine virtuelle contenant le logiciel HYRAX et possédant le même jeu de données que la machine virtuel contenant la solution ERDDAP afin de comparer les deux solutions proposées. En introduction de sa présentation, FB fait un tour d'horizon des solutions OpeNDAP pouvant être utilisées (voir 03_Frederic_Briol_serveursOPeNDAP_Hyrax.pdf) :

- Dapper (fortement déconseillé car ce logiciel n'est plus maintenu),
- GDS (comme Dapper cette solution est fortement déconseillé par elle n'est plus maintenue),
- PyDAP est un client écrit en Python avec une partie serveur. Cette librairie écrite en python est bien pour développer des clients qui vont interroger un serveur distant mais elle n'est pas très utilisée comme solution côté serveur,
- TDS (Thredds), cette offre logicielle est portée par UCAR/unidata, avec de nombreux contributeurs et des « commits » très nombreux et récents, ce qui permet d'envisager une bonne évolution et pérennité de cette offre. TDS est écrit en JAVA,
- ERDDAP, cette solution est porté par Bob Simons de la NOAA et un autre contributeur,
- Hyrax est une solution écrite en C++ et Java. Elle est portée par la NOAA, la NASA, le NSF, l'australian bureau of meteorology, ce qui permet de la classer dans les solutions pérennes qui ont un avenir certain. De nombreux contributeurs (répartis en groupe de travail) continuent à l'améliorer.

Hyrax est un serveur OpeNDAP qui a l'avantage d'être puissant et rapide. Son architecture originale, séparant le serveur de diffusion du frontal du serveur web, lui permet une mise à l'échelle et le cloisonnement des données distribuées du trafic internet. Comme tous serveurs OpeNDAP, il permet l'interrogation, l'extraction et le sous-échantillonnage d'un jeu de données. De plus Hyrax permet nativement de faire de l'agrégation de grands jeux de données (contenant plusieurs centaines de fichiers). Pour avoir plus de détails sur Hyrax, le lecteur est encouragé à lire la présentation 03_Frederic_Briol_serveursOPeNDAP_Hyrax.pdf. Hyrax est un logiciel jeune (moins

de 5 ans) mais qui est arrivé à maturité, faisant de lui une solution alternative pouvant-être déployée dans les CDSs.

2.4. Conclusions et recommandations

À la suite des présentations et des hackathons ERDDAP et HYRAX, il s'avère que ces deux logiciels sont de bonnes solutions pouvant être déployées dans les CDSs afin de mettre à disposition les données ainsi que des flux permettant de faire de la visualisation, de l'interrogation, de l'extraction, etc. Chaque solution ayant ces avantages et inconvénients, le tableau 1 ci-dessous permet de comparer ces deux solutions.

Hyrax est un serveur OpeNDAP. L'utilisateur peut accéder au serveur par une interface WEB pour en découvrir son catalogue, visualiser les données diffusées par le protocole WMS et découvrir certaines de ces fonctionnalités. Mais il ne propose pas un client à proprement parler, comme ERDDAP. Pour l'interroger, il faut passer par un client OPeNDAP (NetCDF, Panoply, PyDAP, par exemple). Il possède en outre une API web d'agrégation.
ERDDAP propose des fonctionnalités supplémentaires d'accès aux données sous forme de graphes et de requêtes, mais sa fonction serveur est moins efficace sur de gros volumes.

	ERDDAP	***HYRAX***
Caractéristiques de la donnée	In-situ (profiles, série temporelles, trajectoires), données sur grille régulière	In-situ profiles, série temporelles, trajectoires), données sur grille régulière et maillage non structuré
Formats d'entrée	NetCDF (SDN, ODV, DIVA), database, WMS, SOS, opendap, etc…	NetCDF(SDN, ODV, DIVA), database (SQL/ODBC), GDAL (GRIB, GeoTIFF), CVS, FreeForm
Services de sortie	WMS, OpeNDAP, etc.	WMS, Thredds catalog, WCS, W10n
Formats de sortie	Un grand nombre de formats de sortie : NetCDF, ODV txt, matlab, R, json,csv …	NetCDF, DAP, ASCII, JSON
Fiabilité	Bonne	Bonne
Maintenance et évolutions	Dépendante de 2 personnes	OPeNDAP
Conformité INSPIRE	À vérifier	IOOS (US)
Produits agrégés	Oui (si éclaté en plusieurs fichiers NetCDF)	Oui
Authentification sur dataset	Oui par compte interne ou compte google	Oui

Tableau 1 : Comparatif entre le logiciel ERDDAP et Hyrax

Au vue du tableau 1, la solution ERDDAP semble bien adapté aux CDSs In-Situ et Hyrax aux CDSs Satéllitales.

3. Architecture globale du pôle ODATIS (Gilbert Maudire)

La seconde journée de l'atelier technique de Novembre est dédiée à la présentation de l'architecture actuelle du pôle ODATIS, et à son évolution avec un focus sur la feuille de route du pôle pour l'année 2019.

GM présente (voir 04_Gilbert_Maudire_architecture_odatis.pdf) tout d'abord les services de traitement qui devraient être proposés à terme aux travers du portail web du pôle, l'objectif étant de fournir une capacité de traitement à la demande « là où les données sont disponibles ». Ces services incluent des suites logicielles permettant :

- d'accéder aux seules données utiles via la possibilité de faire du sous-échantillonnage,
- de visualiser conjointement des données gérées par divers CDSs,
- de proposer de l'accès direct à la donnée (sans la nécessité de téléchargement local) suivant des protocoles variés (cloud, OGC, OpeNDAP, etc.).

Certaines solutions permettent déjà d'effectuer ces tâches, on peut citer par exemple :

- des outils de programmation avec des Notebooks comme Jupyter (Python) ou Rstudio (R),
- la suite logicielle Pangeo (écosystème Python pour les sciences de l'environnement),
- la suite logicielle OceanWorks qui permet la découverte et l'analyse de jeux de données,
- les logiciels : DIVAA, Ocean Data View Online (ODV)
- l'environnement Galaxy. Cet environnement est dédié aux biologistes et bioanalystes pour des projets -omique ou méta-omique. Cet environnement est destiné aux utilisateurs biologistes/bioanalystes peu familiers avec la ligne de commande mais souhaitant exécuter des workflows sur des données de type -omique. Il est basé sur une interface Web qui permet, au niveau le plus élémentaire, d'exécuter des workflows sur des données préalablement téléversées. La sélection des paramètres d'entrée se fait au travers de formulaires, et l'outil permet de suivre la progression de l'exécution de chaque tâche

élémentaire, puis de visualiser leurs résultats. Galaxy permet également, toujours au travers de l'interface Web, de construire de nouveaux workflows en chaînant les outils qui vont le composer. Chacun de ces outils correspond à un outil en ligne de commande. Dans un installation classique de Galaxy, l'exécution de ces outils est déportée sur une infrastructure de type cluster. Le partage de workflows entre utilisateurs d'une même instance de Galaxy est également prévu. L'intégration de nouveaux outils dans une instance Galaxy (i.e. leur mise à disposition au travers de l'interface Web) se fait en composant un wrapper qui décrit de manière normalisée la structure des paramètres en entrée et en sortie.

L'infrastructure du pôle ODATIS est extrêmement distribuée avec des CDSs In-situ (CDS-IS) et Satellites (CDS-SAT) (voir 04_Gilbert_Maudire_architecture_odatis.pdf planches 3 et 4 pour la distribution régionale et la répartition des CDSs IS et SAT). Cette distribution est aussi bien géographique que thématique.

Concernant les CDS-SAT, ils sont au nombre de 3 avec les CDS-SAT de Brest et de Toulouse stables et pérennes. Le 3ème CDS-SAT-Couleur est actuellement en cours de réorganisation avec des objectifs et des partenariats à revoir à la suite de la fin du GIS COOC.

En ce qui concerne les CDS-IS, l'architecture actuelle en compte 6 clairement identifiés (Coriolis, Shom, OMP, SISMER, Marseille, OASU) et 1 (ou 3) correspondant au CDS-IS-UPMC qui comprend 3 sites distincts (Villefranche/Paris avec la base LEFE/CYBER, Roscoff avec la base Pelagos, et enfin Banyuls).

Les données distribuées par ODATIS sont donc de deux catégories différentes qui sont soit des données in-situ, soit des données satellitales. Il est à noter que certains produits synthétiques (i.e. : sorties de modèles numériques océaniques) sont aussi distribués via le portail ODATIS.

Les données satellitales :

Les données satellitales (et de modélisation) possèdent une forte volumétrie qui se compte en PetaOctets et qui nécessite une importante capacité :

- de stockage,
- d'archivage pérenne,

- de traitement (en particulier pour les études sur l'ensemble de l'océan global).

Ces données possèdent des métadonnées limitées (au format ISO 19115/19139-version 2 ou 3), qui se limitent quasiment à la description de la mission et du producteur à l'exception près des données provenant d'imagerie à la demande qui peuvent avoir des métadonnées beaucoup plus volumineuses. Cependant les données d'imagerie à la demande ne seront pas directement distribuées par ODATIS mais via un pôle transverse qui est en train de se construire et qui se nomme DINAMIS.

Au niveau d'Odatis, les données satellitales représentent actuellement un volume global d'environ 2500 To avec une augmentation annuelle de 500 To/an. Cette volumétrie représente les données des 3 CDSs existants et comprend aussi les données de niveau L1. Par contre elle ne prend pas en ligne de compte les nouveaux satellites (Sentinel 3, SWOT, etc.). La plus grande partie de ces données sont au format NetCDF (3 ou 4) avec quelques jeux de données avec des formats images (png, jpg, etc.).

L'accès à la donnée satellitale se fait essentiellement via des protocoles de téléchargement tels que FTP et OpeNDAP. Ces données nécessitent une grande capacité de calcul proche de la donnée de type HPC ou cluster.

Les données in-situ :

Les données in-situ, d'un volume existant actuellement de 300 To en tenant compte des données d'imagerie optique (vidéo, cytométrie, etc.) et acoustique (sonar, sondeur, sismique), ont une augmentation annuelle d'environ 70 To/an. Leurs traitements nécessitent peu de puissance de calcul et leurs mises en ligne requièrent des serveurs ou des machines virtuelles (Linux).

Ces données connaissent une très forte hétérogénéité en terme de structure de données. En effet elles peuvent être :

- sous forme de bases de données relationnelles (chimie, biologie, métadonnées) de type Postgre/Postgis, Oracle, MySQL,
- aux formats textuels « à colonnes » (type CSV),
- aux formats NetCDF (3 ou 4), en particulier les données physiques et de bathymétrie,
- aux formats SEG pour les données de sismique,

- aux formats images et vidéos.

La métadonnée quant à elle, lorsqu'elle est structurée se retrouve aussi sous divers format : Dublin Core, ISO 19115/19139 ou des fichiers d'index (Thredds).

À noter qu'il reste un effort important à apporter sur les données in-situ afin de les pérenniser, ainsi que de les rendre interopérables (métadonnées incomplètes, absence de valeur pour la données manquantes, qualification de la donnée très hétérogène, etc.).

L'accès à la donnée se fait par divers protocoles comme FTP, OpeNDAP (donnée physique), OGC (données du domaine littoral et côtier) ou au moyen de protocoles spécifiques (OBIS, biogéographie, répartition d'espèce).

L'augmentation « naturelle » de ces données, si on se place à capteurs constants, est de 10 à 20 % par an. Cependant il est important de noter qu'il y a actuellement une explosion du volume de données in-situ (mosaïques benthiques, suivie espèces invasives, etc.) de part l'apparition de l'imagerie optique, acoustique, des levées caméra ainsi que des vidéos provenant de nouveaux instruments (ROV, caméra embarquée, caméra fixe, système automatisé d'acquisition de la donnée in-situ, etc.). En ce qui concerne les données génomiques, elles représentent actuellement un volume faible (10 à 100Go) qui risque d'évoluer très rapidement dans les prochaines années. De plus ce nouveau type de donnée qui est en phase d'observation doit impérativement être stocké à la fois sous son format brut et sous son format de distribution car les algorithmes de traitement ne sont pas encore standardisés et sont encore en cours d'évolution. D'autres données apparaissent aussi comme les données de radar-HF (~10To/an/radar), la cytométrie, etc.

À cette augmentation, il faut rajouter que la pérennisation nécessite une copie des données brutes pour un retraitement a posteriori.

Il y a souvent de la duplication des données in-situ (et satellitales) pour les organiser de façon différente pour répondre à une utilisation particulière.

Il est donc nécessaire de s'attendre à une augmentation quasi exponentielle de la donnée in-situ dans les années à venir.

<u>L'évolution du stockage et de la puissance de calcul :</u>

Si on additionne le traitement des données satellitales avec celui des données in-situ, la capacité

de faire des traitements conjoints, l'apparition d'algorithme d'apprentissage (classification, reconnaissance de forme ou de groupe), l'apparition de l'analyse géostatistique, qui sont des traitements très gourmands en puissance de calcul, il va aussi être nécessaire de faire rapidement évoluer l'accès à des puissances de calcul plus importantes et plus proches de la donnée.

De la même façon, les services d'accès et de visualisation vont aussi nécessiter que la donnée soit accessible immédiatement et à tout moment. Ceci est difficile à mettre en adéquation avec une infrastructure technique très distribuée où la disponibilité de la ressource de calcul et de la donnée dans chaque CDS ne sont pas identiques. Il va donc être nécessaire d'avoir au niveau du pôle, une capacité de calcul et de traitement suffisamment proche de la donnée pour rendre des services via le web sans délai d'attente trop important.

Les services aux producteurs (services amonts) :

En 2018, le pôle ODATIS a fait le constat qu'il y avait une demande croissance chez les producteurs d'obtenir un DOI sur leurs jeux de données et d'avoir un service d'archivage pérenne pour ceux-ci afin de pouvoir publier leurs données. De plus, il a été évalué qu'environ 60 % des données historiques ont été perdues. Une organisation a donc été mise en place afin de permettre aux producteurs d'obtenir facilement un DOI et de placer leurs données dans un centre qui les pérennisent. Il est donc possible actuellement pour tous producteurs de données d'obtenir facilement un DOI en utilisant différentes solutions. En particulier, il est possible d'obtenir un DOI via SeaNoe ou au Sedoo avec la possibilité de faire distribuer la donnée via ces deux centres.

Une demande grandissante est aussi apparue pour avoir la possibilité d'associer le pôle ODATIS en temps que centre certifié pour la pérennisation des données lors de l'écriture de projet demandant un Data Management Plan (DMP). Le pôle ODATIS est donc actuellement en train de mettre en place un template de DMP pour en faciliter l'écriture. Il est à noter qu'une collection de DMP sera certainement nécessaire à mettre en place en fonction de divers paramètres (types de demandes, nature des données, etc.). En parallèle de cela, une action est en cours pour obtenir une certification CoreTrustSeal de l'architecture globale du pôle.

Les services aux utilisateurs (services avals):

Le pôle ODATIS doit répondre à plusieurs catégories d'« utilisateur/chercheur » :

- le producteur de données (qui génère des données de niveau plus élevé),

- l'utilisateur primaire qui va utiliser la donnée pour sa thématique de recherche,
- l'utilisateur secondaire qui va vouloir reproduire un résultat publié dans un journal,
- l'intégrateur qui désire développer un produit de type climatologie par exemple,
- le modélisateur qui va soit assimiler la donnée, soit l'utiliser pour calibrer et/ou valider son modèle numérique,
- etc.

À ces catégories d'utilisateur/chercheur, le pôle doit aussi répondre à d'autres communautés hors recherche (appui aux politiques publiques, DSCMM, …). Il y a donc plusieurs niveaux d'utilisation de la donnée très distincts les uns des autres qui amènent des demandes et des services à mettre en place très divers.

Actuellement, le pôle ODATIS cherche à mettre en place de façon systématique pour toutes les données qu'il distribue, des services aux utilisateurs :

- téléchargement du jeu de données complet, via le DOI (http) ou via un téléchargement direct (http, ftp),
- téléchargement via un portail d'un sous-ensemble d'un jeu de données au moyen d'un outil de sélection,
- prévisualisation des données (et des métadonnées) par la localisation des positions d'observation, et/ou prévisualisation de la donnée elle-même,
- service d'accès direct à la donnée :
  - par l'utilisation d'un HPC et de fichiers locaux, mais aussi par un système de virtualisation de fichier (Irods),
  - via des protocoles OpeNDAP ,
  - via des protocoles OGC (WFS, WCS, WMS) avec éventuellement O&M SOS.

Depuis peu, on voit apparaître une forte demande des utilisateurs pour effectuer des traitements à la demande sur des jeux de donnée globaux en utilisant des notebooks (Python,R, outil Galaxy) afin de traiter ce jeu de donnée avec des algorithmes développés sur un sous-ensemble (passage à l'échelle d'application développée localement). Il existe aussi une demande pour que le pôle fournisse une VRE (Virtual Research Environment – suites logicielles permettant l'analyse et le

traitement).

Contexte national et international :

Le pôle ODATIS doit s'insérer dans un contexte à la fois national et international.

Il faut donc prendre en compte la politique nationale du MESRI qui désire une rationalisation des centres de données via InfraNum (3 centres de niveau « Européen » : Hydris, IN2P3, CINES et des centres régionaux : un MésoCentre par région).

Au niveau européen, il est nécessaire de s'insérer dans les démarches de l'European Open Science Cloud (EOSC) pour les données in-situ de la recherche et dans les Data Information and Access Services (DIAS) pour les données satellitales et synthétiques (pour l'océan avec Eumersat, ECMWF, Mercator, etc.).

De plus le Pôle Odatis au même titre que les autres pôles doit répondre via une demande groupée au niveau de l'IR « système terre » au Programme des Investissements d'Avenir (PIA3). Cette démarche groupée et coordonnée est un critère fort de recevabilité des propositions au PIA3.

Pour répondre à ce contexte national et international le pôle ODATIS est en train de mettre en place des services hiérarchisés dont l'imbrication est résumée par la figure 1.

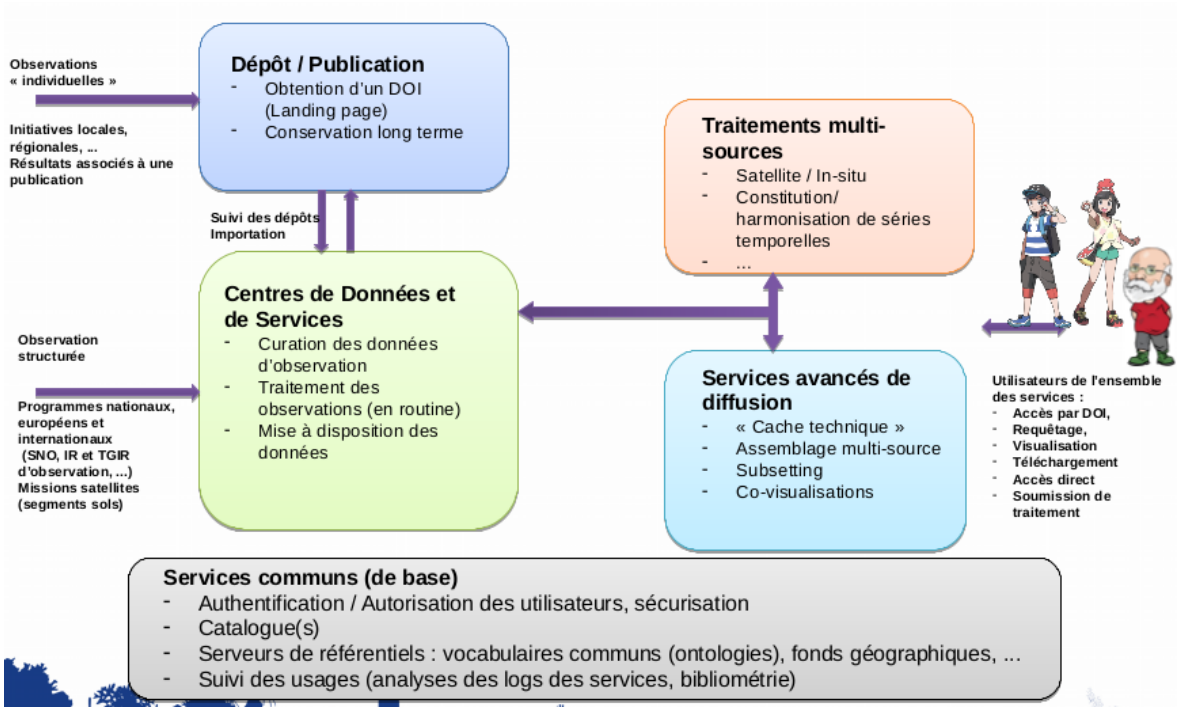

Figure 1 : Services hiérarchisés du pôle ODATIS

Le cœur de l'architecture proposée se situe au niveau des CDSs dont l'objectif principal est de fournir un service thématique de préservation et de production, tout en étant garant:

- de la qualité des données et de leur « harmonisation » dans le temps,
- de la qualité des métadonnées (précises, définition de référentiels communs, contrôle qualité, …),
- de la production régulière de données dérivées selon des méthodes établies par les pôles (Conseil Scientifique, Consortium d'Expertise Scientifiques, …),
- de l'archivage long terme (avec leurs moyens ou par délégation, en s'appuyant par exemple, sur les services de dépôt),
- de la mise à disposition des données pour leur communauté,
- de l'interopérabilité technique et sémantique.

Il est important de noter que le nombre de CDSs doit rester restreint, qu'un CDS doit être approuvé par le bureau exécutif (BE) d'ODATIS ainsi que par son comité directeur (CODIR) et qu'un CDS n'est pas institutionnel mais thématique (ex : un OSU ou un institut n'est pas forcement un CDS, un CDS n'appartient pas forcément à un OSU).

Le BE et le CODIR d'ODATIS auront une attention particulière pour toute nouvelle demande de création d'un CDS sur :

- les compétences thématiques du CDS et la position géographique du CDS avec la proximité thématique (i.e. : proche des compétences techniques et scientifiques du thème abordé par le CDS),
- la proximité avec les programmes d'observation (en particulier pour les mesures in-situ),
- le matériel informatique, connexion réseau, etc. qui sont disponibles dans le CDS pour son bon fonctionnement (les moyens humains seront aussi pris en compte).

Par contre l'organisation interne des données dans le CDS ne sera pas un critère d'évaluation lors de la création car il doit répondre au besoin de la thématique (organisation en fichiers, en base de données,…), seuls les moyens de mise à disposition de la donnée et l'interopérabilité avec le pôle ODATIS seront évalués. Le CDS devra donc être conforme avec les exigences techniques du pôle pour la mise à disponibilité des données (charte entre le CDS et le pôle) :

- niveau de service minimal,

- adoption de formats d'échange standardisés (NetCDF convention CF, CSV/ODV Spreadsheet)
- adoption de métadonnées standardisées (ISO19115 ou équivalent, CSW (ou OAI/PMH),
- référentiels commun pour les échanges (harmonisation sémantique),
- interopérabilité,
- synchronisation avec les services de diffusion (WMS pour la prévisualisation, OpeNDAP/ERDDAP,...),
- ...

Le service de dépôt et de publication :

Afin de répondre à une demande croissante des producteurs de données, le pôle ODATIS s'est doté d'un service permettant de déposer un jeu de données et d'obtenir un DOI sur ce jeu. Ce service de dépôt permet de préserver « en l'état » un jeu de données plutôt que voir ce jeu disparaître à moyen terme (actuellement ~60 % des données in-situ ont été perdues). Bien qu'il soit fortement conseillé de mettre les données dans un des deux formats préconisés par le pôle, ce site permet de déposer les données sous n'importe quel format. Il est à noter que ce service permet aussi de répondre à la demande grandissante des éditeurs de journaux scientifiques qui imposent le dépôt des données utilisées dans la publication dans un site certifié (certification de type RDA – CoreTrustSeal Data repository ou ISO 16363) avec un DOI sur le jeu de données. Ce type de dépôt est surtout à envisager pour des données « orphelines » de CDS.

Si un jeu de données se situe dans la même thématique qu'un CDS existant, le pôle Odatis conseille le rapprochement vers ce CDS plutôt qu'un dépôt de type « données orphelines », afin de filtrer des dépôts redondants ou indésirables.

À côté de ce type de dépôt, les CDSs doivent ingérer, traiter leurs données, créer des jeux de données « stabilisés » et ensuite mettre un DOI sur leurs jeux de données, ce qui permet de les valoriser.

Plusieurs centres permettent déjà de placer un DOI sur un jeu de données : SeaNoe, le SEDOO, Eudat (CINES) mais aussi via la plateforme Zenodo (projets européens) ou via le logiciel DataVerse.

Vers des services de diffusion avancés :

L'objectif du pôle est de fournir des services d'accès performants aux utilisateurs pour l'ensemble des données géré par le pôle. Adossé à ce services d'accès, le pôle veut mettre en place un service de visualisation permettant :

- de combiner, lors d'une même requête des données de différentes origines (permettre l'assemblage des données),
- de faire du sous-échantillonnage d'un jeu de données,
- de visualiser conjointement des données provenant de plusieurs CDSs,
- de proposer de l'accès direct à la donnée sans avoir à la télécharger, suivant divers protocoles (cloud, OGC, OpeNDAP,…).

Afin de mettre en place ce type de service, il est nécessaire d'avoir des données qui soient accessibles immédiatement et sans rupture de continuité de service de distribution. Il est aussi nécessaire de ne pas avoir de latence au niveau des requêtes. Pour cela, une architecture distribuée comme le pôle ODATIS va se heurter à de nombreux problèmes avec ses CDSs car il faut absolument avoir des temps de réponses très courts et une disponibilité quasi immédiate de la donnée. Afin de palier aux principaux problèmes de latence, le pôle réfléchie à la mise en place d'un « cache technique », invisible aux utilisateurs qui permettrait de mettre l'ensemble des données proche du serveur qui aura la charge de visualiser ou de calculer sur ces jeux de données. L'idée du « cache technique » est de mettre en place une distribution et un accès aux données extrêmement rapide et fiable sans pour autant mettre le CDS au second plan.

Cette structuration va aussi permettre de réorganiser les données dans le « cache technique » pour avoir une orientation « utilisation » plutôt qu'une orientation « observation et téléchargement » des données. Il va donc y avoir une duplication de la donnée. Ceci sera possible pour les données in-situ et les produits élaborées uniquement. Il ne va pas être possible de mettre dans ce « cache technique », l'ensemble des données satellitales (ceci ne veut pas dire que les données satellitales ne seront pas accessibles via le site mais qu'elles ne seront pas dupliquées vue leurs volumes).

Certains CDSs auront la possibilité d'héberger des services en ligne du pôle. Ces CDSs devront satisfaire à des contraintes beaucoup plus strictes en ce qui concerne :

- la disponibilité des données (avec de la redondance dans leurs infrastructures informatiques permettant une distribution 7j/7 et 24h/24),
- la sécurité (système d'authentification identique à celui mis en place par le pôle),

- l'absorption des pics d'audience (serveur permettant plusieurs traitements concurrents) et les pics de charge CPU/RAM/mémoire partagée,...,
- la connectivité au réseau (connectivité permettant de transférer des flux importants de données via le réseau).

Chaque CDS aura la possibilité soit de déléguer au pôle ODATIS les tâches informatiques et logistiques pour maintenir un service opérationnel 24h/24 et 7j/7 (« hébergement délégué » des services en ligne), soit de maintenir lui-même ses services. À noter que ce n'est pas le CDS qui va décider de sa propre initiative mais que chaque CDS va faire l'objet d'un état des lieux en 2019 et que suivant cet état des lieux, le CDS sera classé soit sous un « hébergement délégué », soit sous un « hébergement mandaté » par le pôle.

Il va donc être nécessaire de mettre en place dans chaque CDS une synchronisation des données avec le pôle ODATIS (rsync, OpeNDAP, etc.).

4. Conclusions et objectifs du prochain atelier

Lors de cet atelier, l'accent a été mis sur l'étude de deux outils pouvant être mis en place dans chaque CDS (ERDDAP et HYRAX).

Le second jour a été consacré à l'architecture existante du pôle ODATIS et à son évolution.
Les objectifs qui avaient été fixés en début d'année ont été atteints en particulier :

- la mise en place de la version 2 du site web avec homogénéisation de la mise en page avec les autres pôles,
- la mise en place d'une charte graphique pour le pôle ODATIS,
- l'enrichissement du catalogue,
- la mise en place de l'atelier technique du pôle ODATIS (3 ateliers ont été effectués et un atelier RESOMAR/ODATIS),
- la sélection de format d'échange (NetCDF et ODV/CSV Spreadsheet).

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 (voir 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.

Pour le dernier point ci-dessus, le CNES propose de mettre à disposition des ressources de son HPC afin de tester les possibilités de travail à distance via les outils mis en place au CNES. De la même façon, l'HPC de l'ifremer (datarmor) va aussi être une plateforme de test. Afin de faire des cas d'études réalistes permettant de mettre à contribution des données à la fois satellitales et in-situ, le CNES propose de faire un miroir des données in-situ pour les rendre disponible sur son HPC.

Le prochain atelier se fera vers la fin Mars (la date et le lieu étant encore à déterminer à ce jour).
Au cours de cet atelier une journée sera consacré aux outils à déployer au niveau du « système central » du pôle ODATIS et à l'organisation des données. La seconde journée sera dédiée aux CDSs (état des lieux, avancement, etc.).