Compte rendu de l'atelier technique des 22 et 23 septembre 2020

CR atelier technique septembre 2020

Titre court
CR_AT_sept2020
Titre long
Compte rendu de l'atelier technique du 22 et 23 septembre 2020
Auteur
Joël Sudre

Dissémination	Copyright
	Pôle ODATIS

Historique

Version	Auteurs	Date	Commentaires
0.1	Joël Sudre	29 septembre 2020	Version initiale
0.2	Cécile Nys	3 novembre 2020	Relecture et corrections
0.3	Sabine Schmidt	7 novembre 2020	Relecture et corrections
0.4	Valérie Harscoat	9 novembre 2020	Relecture et corrections
0.5	Cécile Nys & Joël Sudre	12 novembre 2020	Acceptations corrections
0.6	Valérie Harscoat	17 novembre 2020	Relecture et corrections
0.7	Cécile Nys & Joël Sudre	18 novembre 2020	Amendements, harmonisation avant envoie pour relecture par les participants
0.8	Clémence Rabevolo	18 novembre 2020	Révision SIMM – Clémence
0.9	Gemma Gimenez-Papiol	18 novembre 2020	Révision EMBRC
0.10	Mark Hoebeke	18 novembre 2020	Révision SBR/SDN
0.11	Catherine Borremans	23 novembre 2020	Révision imagerie
0.12	Cécile Nys	26 novembre 2020	Relecture, acceptation corrections
0.13	Joël Sudre & Cécile Nys	13 janvier 2021	Relevé synthétique
0.14	Arnaud Rouilly	14 janvier 2021	Révision Quadrige
0.15	Cécile Nys	14 janvier 2021	Relecture, acceptation corrections
1.0	Joël Sudre & Cécile Nys	14 janvier 2021	Version diffusable

Table des matières

Table des illustrations

1. Accueil et tour de table des participants

Tableau 1. Liste des participants à l'atelier ODATIS #7 des 22 et 23 septembre 2020

Liste des participants à l'Atelier Technique #7	
Armelle Rouyer *(IFREMER, SIMM) - AR*	Julia Uitz *(IMEV) – JU*
Arnaud Rouilly *(IFREMER, Quadrige) - AR*	Kévin Rodriguez *(IFREMER) – KR*
Brendan Hennebaut *(IFREMER) - BH*	Laurence Lebourg *(IUEM) – LL*
Caroline Mercier - *CM*	Laurent Coppola *(Sorbonne Université / IMEV) – LC*
Catherine Borremans *(IFREMER, Environnement Profond) - CB*	Marc Sourisseau *(IFREMER) – MS*
Catherine Schmechtig *(Sorbonne Université / CNRS) - CS*	Mark Hoebeke *(Sorbonne Université / CNRS) – MH*
Cécile Nys *(IFREMER) - CN*	Maurice Libes *(MIO) – ML*
Cédric Chauvel *(IUEM) - CC*	Michèle Fichaut *(IFREMER) – MF*
Clémence Rabévolo *(IFREMER, SIMM) - CR*	Nicolas Desroy *(IFREMER) – ND*
Fabrice Mendes *(CNRS/EPOC) - FM*	Raffaela Cattaneo *(CNRS - EMBRC - IMEV) – RC*
Felipe Artigas *(CNRS/ULCO) - FA*	Raphaëlle Sauzède *(CNRS/IMEV) – RS*
Florence Conquet *(IFREMER) - FC*	Sabine Schmidt *(Univ. Bordeaux, OASU - EPOC) – SS*
Gemma Gimenez-Papiol *(EMBRC - IMEV) - GGP*	Sophie Pamerlon *(MNHN, GBIF) – SP*
Gérald Gregori *(MIO) - GG*	Sylvie Van Iseghem *(IFREMER) – SVI*
Gilbert Maudire *(IFREMER) - GM*	Thierry Carval (IFREMER) – TC
Jean-Olivier Irisson *(Sorbonne Université / IMEV) - JOI*	Valérie Harscoat *(IFREMER) – VH*
Jérôme Detoc *(IFREMER) - JD*	Yolanda Del Amo *(OASU) - YDA*
Joel Sudre *(UMS CPST/CNRS) - JS*	Yvan Le Bras *(MNHN, PNDB) - YLB*

JS présente l'ordre du jour (*voir : Agenda et accès aux présentations*). Le dernier Atelier Technique (AT #6 – Compte-rendu[1]) précédent ayant fait remonter les besoins des CDS, il a été noté une demande forte des CDS pour que l'atelier technique d'ODATIS apporte des solutions techniques, des mises en pratique (avec une prise en main des solutions) et des recommandations. Cet atelier technique se présente sous un nouveau format en prenant en compte ces besoins et est dédié à la **gestion des données biologiques marines.**

[1]https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_201910/AtelierTechnique_Odatis_CR_201910.pdf

2. Présentation du Pôle de données Océan – (Gilbert Maudire)

GM présente un résumé des événements qui ont marqué le pôle ODATIS et l'IR Data Terra depuis l'atelier technique d'octobre 2019 afin d'informer l'ensemble des CDS. GM rappelle que l'IR Data Terra est composée de 4 pôles thématiques (AERIS, FORM@TER, ODATIS, THEIA) ainsi que différentes structures transverses telles que :

- DINAMIS ;
- le groupe Interpôles ;
- le Groupe de Travail (GT) Tech ;
- le GT Science ;
- le GT Europe & International.

A moyen terme, le pôle ODATIS proposera d'effectuer des traitements plus évolués, comme la co-localisation de données entre données in-situ et satellite, sur de longue série de données, avec des ressources informatiques importantes en mode VRE/VAP pour les utilisateurs. Cette évolution passe nécessairement par la mise en place de services plus évoluées partagés entre les pôles au niveau de l'IR Data Terra. Cette ambition, portée par l'ensemble des pôles, a permis d'écrire et de déposer le projet GAIA DATA (GlobAl IntegrAted DATA and services research infrastructure for Earth system, biodiversity and environment observation, modelling and understanding) en réponse à l'appel à projet PIA3/Equipex+. GAIA Data est porté par trois IRs : Data Terra, Climera et PNDB. Ce projet, si fructueux, permettra de mettre à la disposition des utilisateurs deux types de plateforme :

- VRE : des logiciels existants sont rendus disponibles sur une plateforme d'analyse de données, par exemple Galaxy dans le domaine biologie,
- VAP : environnement destiné à un profil développeur pour les besoins de faire tourner/tester à l'échelle des traitements développés sur des moyens plus légers.

Ces VRE et VAP seront mis à disposition sur des moyens de calcul HPC ou HPDA qui seront associées à des espaces techniques (*data lake*) offrant un accès rapide et efficace aux données multi-sources référencées dans les catalogues de l'IR qui comprennent toutes les données des différents pôles et structures transverses.

Pour atteindre ces objectifs, il y a donc une forte nécessité de mettre en place des recommandations sur les données, les métadonnées, les référentiels, ainsi que sur les normes à appliquer. Les ateliers précédents ont permis d'émettre des recommandations sur les données physiques et biogéochimiques. Il est maintenant nécessaire de débuter la même démarche pour les paramètres biologiques.

GM précise que le présent atelier est consacré aux données biologiques sur l'ensemble du domaine Océan, du littoral au grand large, dont il faut faire avancer la gestion dans de bonnes conditions dans ODATIS, que ce soit des données in-situ, y compris acquises ponctuellement, et des données satellitaires. GM explique que l'objectif est, entre autres, d'harmoniser les jeux de données acquises et faire un état des lieux : quelles données, quels logiciels, quels besoins d'analyse de données et à

localiser sur quels types d'infrastructure ; et voir comment collaborer ensemble puisque le périmètre de ce type de donnée dépasse celui du pôle ODATIS. En effet, il concerne aussi le PNDB, le SAR/SIMM, le GBIF, l'OFB…

3. Présentation des initiatives dans la gestion des données biologiques marines

3.1. PNDB, de la gestion de données à son analyse, présentation des outils et services – (Yvan Le Bras)

YLB présente le Pôle National de Données de Biodiversité (PNDB) ainsi que les outils et services de ce pôle. Le PNDB est une e-infrastructure nationale de recherche, tout comme l'IR Data Terra, elle dépend du Ministère de l'Enseignement Supérieur de la Recherche et de l'Innovation (MESRI). Le PNDB est un consortium de 11 partenaires institutionnels (FRB, BRGM, CIRAD, CNRS, IFREMER, INERIS, INRAE, IRD, MNHM, Univ. Montpellier et l'OFB). Les objectifs, la stratégie ainsi que le cahier des charges du PNDB étant clairement indiqués sur le support de la présentation, ils ne seront pas détaillés dans ce compte rendu mais ils sont disponibles sur le site ODATIS (voir 202009_ODATIS_Atelier_YLeBras_PNDB.pdf[2]).

Le site web du PNDB est accessible en suivant ce lien : https://www.pndb.fr/

Le portail de données et de métadonnées du PNDB est accessible en suivant ce lien : https://data.pndb.fr/

YLB présente ensuite l'outil metaShARK (https://metashark.test.pndb.fr/) qui est un outil de saisie de métadonnées (API rshiny) permettant de rendre accessible l'Ecological Metadata Language (EML). Cet outil permet aussi de créer des « datapackages » qui sont des ensembles de données et de métadonnées et autres objets de recherche comme des scripts ou des protocoles.

Le PNDB s'est aussi doté d'une plateforme d'analyse de données Galaxy (https://ecology.usegalaxy.eu/) pour l'écologie qui permet de mettre à disposition des briques élémentaires facilement utilisables en workflow par un scientifique ne sachant pas coder en langage R ou autres. A noter qu'il est nécessaire de transcoder les scripts en R en brique élémentaire avant de les incorporer à l'environnement Galaxy et que l'utilisation de la plateforme est bienvenue, tout comme les retours sur son utilisation. Différents exemples de workflows sont présents sur la plateforme, entre autres, une reprise d'indicateurs de biodiversité qui étaient présents dans l'application pampa de l'IFREMER (API obsolescente) qui a été transcodée dans Galaxy. C'est la communauté du PNDB qui assure que les workflows fonctionnent. Afin d'intégrer des briques élémentaires dans Galaxy, il est nécessaire d'avoir en amont toute une infrastructure et des codes « propres » avec des dépendances bien identifiées et containerisés (Github, Conda, containers, jupyter Notebook, Rstudio). La communauté Galaxy étant très active, elle effectue une révision systématique des codes avant de les mettre à disposition (voir diapo 13,14). Certaines

[2]https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202009/202009_ODATIS_Atelier_YLeBras_PNDB.pdf

démonstrations des outils du PNDB sont accessibles sur la chaîne Youtube (voir : https://youtu.be/OVViSMzRGtw).

3.2. EMBRC (France) Données biologiques marines dans le Centre National de Ressources Biologiques Marines – (Gemma GIMENEZ-PAPIOL)

GGP présente la composante française de l'European Marine Biological Resource Centre (EMBRC-France) qui est devenu depuis 2018 un European Research Infrastructure Consortium (ERIC). Cette branche française de l'EMBRC se nomme Centre National de Ressources Biologiques Marine.. La mission de l'EMBRC-France est de soutenir la recherche sur les organismes et les écosystèmes marins ainsi que de favoriser leur exploration et leur exploitation durables. L'EMBRC au niveau européen est constitué actuellement de 9 nœuds nationaux (Belgique, France, Grèce, Norvège, Royaume-Uni, Italie, Islande, Espagne, Portugal), d'un siège central se trouvant à Paris et de 45 sites mis à disposition. Les services offerts par l'EMBRC sont actuellement en révision, cependant certains services sont toujours accessibles aux utilisateurs comme l'accès :

- aux écosystèmes ;
- aux ressources biologiques ;
- aux installations expérimentales ;
- aux plateformes technologiques ;
- au matériel éducatif et aux centres de documentation ;
- au logement et à la restauration de chaque site.

Des e-services sont aussi accessibles comme les outils et les logiciels d'analyse de données ainsi que des jeux de données de chaque site.

La stratégie scientifique, pour la période 2020-2023, de cet ERIC est constituée de 8 chantiers principaux :

- Surveillance, caractérisation et étude taxonomique de la biodiversité ;
- Domestication des espèces marines ;
- Développer des outils post-génomiques pour les organismes marins ;
- Développer des systèmes expérimentaux pour étudier un monde en évolution ;
- Mis en conformité ABS des sites EMBRC et des bioressources ;
- Formation et éducation ;
- FAIRisation des données et des supports ;
- Promouvoir les séries de données à long terme via l'EMBRC.

L'EMBRC - France a deux tutelles : l'université de la Sorbonne et le CNRS et a des financements provenant de l'ANR depuis 2012. L'EMBRC est impliqué dans la feuille de route française dans les infrastructures en biologie et santé mais aussi en écologie, environnement et océanologie.

Actuellement en France trois stations marines sont impliquées dans l'EMBRC : la Station Biologique de Roscoff (SBR), l'Institut de la Mer de Villefranche (IMEV), l'Observatoire Océanologique de Banyuls-sur-Mer (OOB) et le centre opérationnel se trouve à l'Université de la Sorbonne. L'EMBRC-France a accès aux données et aux services français comme Ecotaxa, les données de biodiversité, …

GGP présente ensuite plusieurs situations montrant le parcours de la donnée de l'acquisition jusqu'à la mise à disposition pour les partenaires, et la situation d'EMBRC-France par rapport à ce parcours et autres Infrastructures, afin de montrer la coordination nécessaire pour répondre aux besoins des utilisateurs.

Suite à la présentation GM fait les remarques suivantes :

EMBRC-France peut venir chercher les données environnementales sur le portail ODATIS. ODATIS serait intéressé par un accès aux données de caractérisation de la biomasse qui pourraient être confrontées aux données environnementales. Il y a aussi de nombreux défis et sujets communs entre EMBRC-France et ODATIS.

3.3. Imagerie – « Gestion de données biologiques marines » (Catherine BORREMANS)

CB est ingénieure biologiste dans le domaine de l'imagerie à l'IFREMER au Laboratoire Environnement Profond (LEP) et coordonne avec Dominique Pelletier le groupe IFREMER Imagerie. CB présente les activités de ce groupe transverse à l'IFREMER en se focalisant sur la gestion des données d'imagerie ayant rapport avec la biologie marine (voir 202009_ODATIS_Atelier_CBorremans_Imagerie.pdf[3]).

L'imagerie est utilisée depuis environ 50 ans pour les sciences marines mais connait une augmentation exponentielle du fait des progrès en termes de technologie et de coût mais aussi en raison de son caractère non destructif pour l'écosystème observé. L'imagerie permet aussi de faire du suivi et de la surveillance. L'imagerie est utilisée dans tous les domaines marins, du côtier à l'océan hauturier profond et pour tous les compartiments biologiques (du benthos jusqu'au plancton pélagique). Il existe plusieurs types d'imagerie avec différents systèmes d'acquisition, différents types d'images et différents formats. Il y a de l'imagerie 2D ou 3D fixe, rotative ou embarquée sur différentes plateformes allant du sous-marin aux satellites en passant par l'imagerie de cameras aériennes ou de laboratoires. Il existe aussi toute l'imagerie provenant de la microscopie, du zooscan, … Il y a donc un nombre très important de systèmes d'observation permettant l'acquisition d'image.

La présentation se focalise ici uniquement sur l'imagerie in-situ benthique. Il y a cependant une description générale des flux de données sur la diapo 5 et les type de données à gérer sur la diapo 6 de la présentation 202009_ODATIS_Atelier_CBorremans_Imagerie.pdf. L'objectif est d'extraire à partir des images les données sur les taxons, leur positionnement dans l'image, la taille des organismes, le type de substrat, etc. et d'en déduire des variables telles que la densité, l'abondance,

[3]https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202009/202009_ODATIS_Atelier_CBorremans_Imagerie.pdf

etc. Il y a donc une association nécessaire de ces données (i) aux métadonnées relatives à l'acquisition des images et (ii) à des référentiels.

Les données brutes sont archivées et un portail y donne accès pour l'imagerie acquise par les engins sous-marins, en revanche il n'y a pas de lien entre les données brutes et les données extraites. De nombreux logiciels permettent d'utiliser, d'explorer et d'analyser les images (voir diapos de 8 à 13) tels que 'Adelie Vidéo' (analyse de vidéo), Biigle (plateforme d'annotation en ligne d'image fixe et vidéo avec un module de « machine learning »), Matisse (reconstruction 3D), DeepSeaSpy (plateforme d'annotation en science participative avec une BDD archivée et bancarisée à l'Ifremer).

Afin d'appliquer des méthodes de « machine learning » sur les données issues des images, il est nécessaire de standardiser les protocoles et les workflows, et donc les données (exemple de workflow/format : PAMPA, pour les données issues des caméras STAVIRO et MICADO). Au niveau international, le groupe MBON travaille aussi sur la standardisation des protocoles et workflows afin de produire à partir d'imagerie les EBV/EOV (voir aussi les travaux de Schoening sur les workflows des données d'image provenant d'AUV). Le groupe local ALLOHa[4] (automatisation de l'analyse des images marines optiques) a permis de mettre en place des procédures d'annotation pour le développement d'outils d'annotation automatique d'image sous-marines (machine learning). A noter que le MBARI développe un jeu de données images pour entraîner les modèles de « machine learning » afin de faire de la classification. Il existe différents catalogues de taxons pour l'imagerie (voir Althaus et al. (2015), doi: 10.1371/journal.pone.0141039[5]; Howell et al. (2019), doi : 10.1371/journal.pone.0218904[6] et diapo 20 pour les communautés actives sur le « machine learning »).

Plusieurs projets sont en cours à l'IFREMER sur ces sujets, en particulier un projet issu de la AMII Océan 2100 s'intéressant à la standardisation et au format des données (Voir MH). Le projet BLUEREVOLUTION s'intéresse aux images 3D et à la classification automatique de microorganismes (méiofaune). Il existe aussi une équipe autour du développement du logiciel BIIGLE à l'IFREMER.

A noter que le Marine Imaging Workshop est organisé à Brest en 2022.

3.4. ECOTAXA & Zooplancton – (Jean-Olivier IRISSON)

Présentation accessible via la page suivante et UNIQUEMENT sur authentification : Atelier Technique - septembre 2020[7].

JOI présente les observations de zooplancton et le flux de données associé. La structuration de la communauté s'est faite lors de deux ateliers RESOMAR-ILICO (l'un à Villefranche en 2016 et l'autre à Toulon en 2019) coordonnés par L. Mousseau, S. Gasparini et J-L Jamet. L'objectif était d'harmoniser le protocole de prélèvement, les méthodes de comptage et de définir le flux de données.

[4] https://www.campusmer.fr/ALLOHa-3558-0-0-0.html
[5] https://doi.org/10.1371/journal.pone.0141039
[6] https://doi.org/10.1371/journal.pone.0218904
[7] https://www.odatis-ocean.fr/?id=434

Le protocole de prélèvement a été harmonisé pour toutes les stations côtières à un prélèvement tous les 15j avec un Filet WP2, des préconisations en termes de profondeur, de vitesse, de mesure de métadonnées, une fixation au formol (4%) tamponné au borax, un contrôle du pH des bocaux 1 mois plus tard et stockage dans des flacons avec double opercule.

Le protocole de comptage se déroule comme suit :

1. Nettoyage au formol tamis 200µm ;
2. Fractionnement avec boîte de Motoda, Folsom ou pipette ;
3. Tri selon deux méthodes de comptages (deux types de données) :
    a. Tri binoculaire : taxonomie fine (cuve Dofus ou Bogorov, identification de 300 individus minimum sur la base d'une liste taxonomique commune matchée sur WoRMS), (deux listes communes : une liste « Méditerranée » et une liste « Manche + Atlantique »). Il y a donc deux niveaux d'identification : un niveau au plus fin des capacités de chaque station et un niveau dégradé provenant de la liste commune (protocole calqué sur PhytOBS). En sortie, un fichier tableur Excel est créé avec les métadonnées et des flags de contrôle qualité, et l'identification taxonomique ;
    b. Tri ZooScan (EcoTaxa) : abondance relative aux grands groupes avec identification de la taille des individus pour avoir une idée du biovolume, séparation en deux fractions (>1000 ou >500µm selon les régions), scan de 1000 - 1500 individus minimum sur grand cadre avec identification via EcoTaxa sur la liste taxonomique commune.

EcoTaxa permet de classifier par projet l'ensemble des espèces zoo-planctoniques et de les classifier par objet (ex : copépodes). Lorsque la catégorie d'objet est suffisamment importante pour faire de l'apprentissage (méthode de « deep » et de « machine learning » combinées), il est possible de faire de la prédiction sur la classification des objets restants avec un bon pourcentage de confiance. On peut aussi trier les objets en fonction de ce pourcentage. EcoTaxa gère aussi les métadonnées associées à l'image. EcoTaxa est un service d'EMBRC-France avec quasi 120 millions d'images dont 53 millions revues par un opérateur humain. Actuellement il y a 943 utilisateurs provenant de 310 institutions. Le modèle de donnée utilisé dans EcoTaxa est présenté sur la Figure 1.

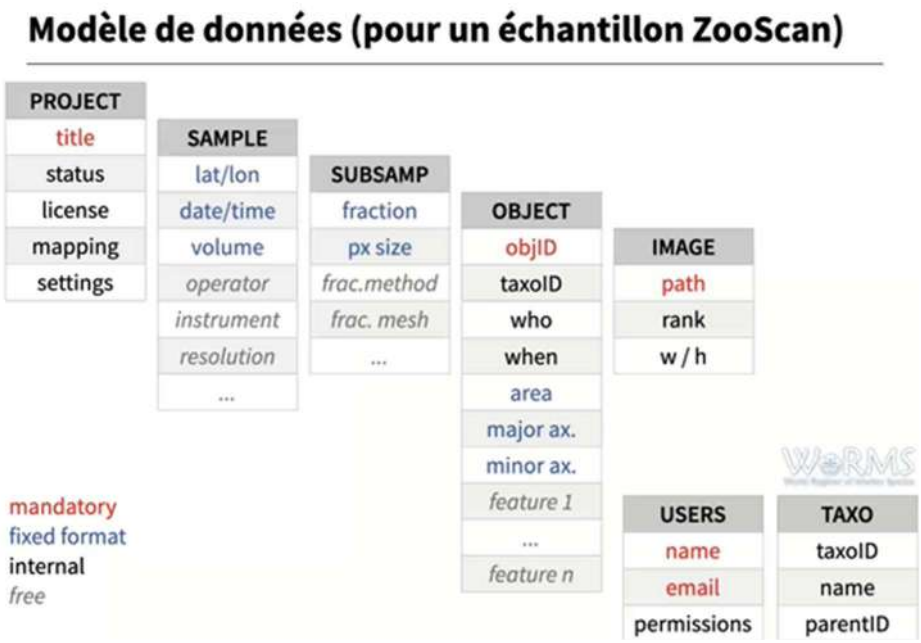

Figure 1: Modèle de données utilisé pour un échantillon ZooScan inséré dans EcoTaxa

EcoTaxa peut faire différents types d'export de données : type tableur (.tsv) ou format DarwinCore (standard international). EcoTaxa possède aussi une API permettant d'automatiser l'export dans certaines bases internationales (ex : EMODnet Biologie).

Le flux de données est en train d'être FAIRisé et est décrit dans un DMP. Pour EcoTaxa, la FAIRisation est bien avancée (envoyée à EMODnet Biology et donc OBIS), par contre pour les tris binoculaires, cela est beaucoup moins le cas car il n'y a aucune base définie (mais il faudra à l'avenir que ces données arrivent dans EMODnet Biology qui est la base de référence mondiale). Le standard pour rendre interopérable les données utilisées est celui de Darwin Core. L'aspect « Reusable » est assuré par la mise en place de licence CC (CC0, CC-BY, CC-BY-NC seule licence reconnue par EMODnet, la CC-BY a été retenue pour Villefranche). Le choix de la licence revient au propriétaire de la donnée.

A noter qu'une instance d'EcoTaxa est implémentée à l'IFREMER de Brest.

En terme de gestion de données, les 3 niveaux sont :

- Données brutes acquises (très volumineux environ 10-100TB sur l'ensemble des laboratoires),
- Le résultat du traitement (imagettes et les données d'identification) dans EcoTaxa,
- Les résumés (agrégation de données) transmis à EMODnet bio,

et il est nécessaire de garder l'ensemble des données et d'en faire une sauvegarde pérenne.

Certaines questions restent encore en suspens comme :

- Où agréger les données de comptage binoculaires (Pelagos ?), est-ce qu'un service de dépôt de fiches de comptage doit être mis en place ?
- Comment et qui détermine les identifiants uniques ?
- Où distribuer les données dans des bases de fichiers (type SEANOE) ?
- Comment transformer les tableaux de comptage au format DarwinCore ?
- Quel hébergement et modèle de service à long terme pour EcoTaxa, car Villefranche est en train d'atteindre ses limites en capacité d'accueil ?
- Quelle licence déposée pour les données du réseau ?
- Comment intégrer à court, moyen et long terme, l'ajout de données génomiques (metaB puis metaG) ?

3.5. SeaDataCloud/EMODNET BIO/PhytOBS : Mise en place des chaînes de traitement à Roscoff pour la publication des données des suivis à long terme – (Mark HOEBEKE)

MH présente la mise en place des chaînes de traitement à Roscoff pour la publication des données de suivi à long terme (voir 202009_ODATIS_Atelier_MHoebeke_SBR.pdf[8]). Ces chaînes de traitement incluent l'ensemble des étapes, de la collecte des données jusqu'à la mise en place de leur accès.

En tant qu'OSU, les missions de la station Biologique de Roscoff (SBR), incluent des activités d'observations, et notamment le suivi à long terme de la biodiversité du plancton (phytoplancton et zooplancton), des communautés benthiques, ainsi que des paramètres physico-chimiques les environnant. SBR est intégrée à des réseaux nationaux d'observation comme le RESOMAR/Pelagos et REBENT et est responsable de points d'échantillonnage pour les SNO PhytOBS, BenthOBS (en incubation) et SOMLIT. L'objectif du SNO Phytobs par exemple est de caractériser la biodiversité des communautés planctoniques aux points de collecte par un échantillonnage bimensuel depuis l'an 2000 (échantillons d'eau de mer en surface et au fond - environ 60m pour SBR). L'analyse des échantillons se fait après fixation (entre 15j et 1 à 2 ans après échantillonnage) avec une détermination taxonomique et une abondance en microscopie optique pour le micro-phytoplancton et une quantification et classification en groupes fonctionnels basée sur la cytométrie en flux pour le nano- et le pico-phytoplancton. L'ensemble de ces données est bancarisé dans la BDD PEGAGOS du RESOMAR. Cette base a été développée en 2010 et est en fonction depuis 2012. Elle a ensuite subi différentes améliorations en particulier en 2012, 2014 et 2019. Le public n'a accès qu'aux métadonnées, l'accès aux données (données brutes, abondances, taxonomie, groupement fonctionnels) étant restreint aux utilisateurs ayant signé la charte et dont la demande

[8] https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202009/202009_ODATIS_Atelier_MHoebeke_SBR.pdf

a été approuvée par le comité de pilotage. La base de données PELAGOS représente 18 jeux de données, environ 5 000 échantillons soit environ 1 200 taxons.

En ce qui concerne le micro-phytoplancton sont bancarisées dans la BDD PELAGOS :

- la caractérisation de l'échantillonnage (nom du suivi, lieu et heure de collecte,…) ;
- la caractérisation de l'analyse (intervenants, date, …) ;
- la caractérisation du spécimen (dénomination originelle, identifiant, dénomination normalisée, …).

La bancarisation se fait au moyen de gabarits (Excel) et chaque utilisateur authentifié dispose d'un espace de travail pour téléverser et réaliser des corrections assistées afin d'insérer ceux-ci dans la base. Le référentiel taxonomique utilisé pour les identifiants est WoRMS et, pour les exports, les vocabulaires contrôlés proviennent de SeaDataCloud/SeaDataNet (P01 - P06) sauf pour les termes taxonomiques désignant les taxons qui sont remplacés par un terme générique permettant de les stocker dans une seule colonne à l'aide des identifiants WoRMS. Il y a donc deux colonnes : l'une avec l'identifiant WoRMS du taxon et une autre avec l'abondance associée.

Le diagramme de la Figure 2 montre les différentes étapes d'annotation et de curation nécessaires pour intégrer les données de la base PELAGOS dans les différentes bases internationales. Les fichiers étant générés à l'aide de routine Python/Notebook ou Java, ils sont extraits des données de la base et mis en forme en divers formats (bioODV ou DwCA – voir diapos de 10 à 18 pour la structuration interne de ces formats de fichier et leurs exports vers d'autres bdd).

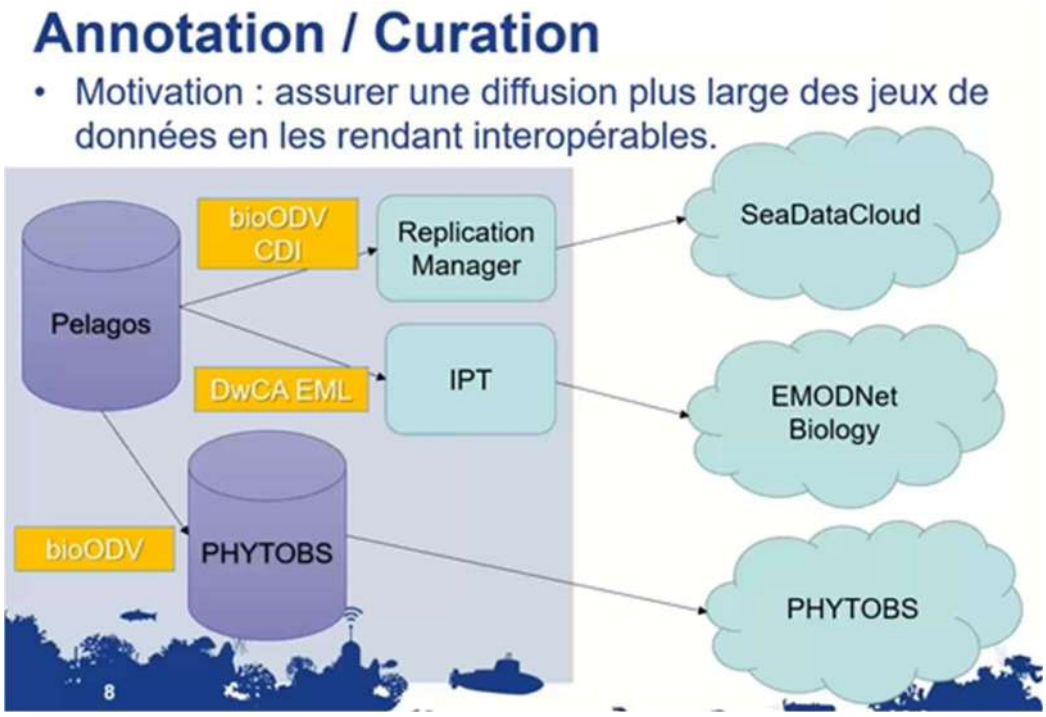

Figure 2: interopérabilité des données de la base PELAGOS

A noter que pour l'export vers SeaDataCloud, une base SQLite est créée avec l'ensemble des métadonnées nécessaires. Cette base étant utilisée par l'outil MIKADO pour générer le fichier de métadonnées.

Pour l'export vers EMODnet Biology, une ressource est créée dans l'instance locale de l'IPT (Integrated Publishing Toolkit), composée de 3 fichierse, puis contrôlée via l'outil de contrôle qualité en ligne (http://rshiny.lifewatch.be/BioCheck/). L'IPT communique avec les différents portails européens (EurOBIS, GBIF, EMODnet, OBIS). L'export des données se fait uniquement après accord du propriétaire de la donnée.

Pour le SNO PhytOBS, un export (1 à 2 fois par an), sous forme de BioODV, dans une base de données indépendante (BDD PhytoOBS) est effectué à partir des BDD Pelagos et Quadrige2.

En termes de retour d'expérience, le point le plus sensible est de convaincre les experts de s'approprier et utiliser les référentiels ainsi que les vocabulaires contrôlés et d'harmoniser les données avant leur mise en accès.

A noter que le support des deux infrastructures (SeaDataCloud/SeaDataNet et EMODnet Bio) est très réactif, qu'il organise des sessions de formation et a créé une abondante documentation.

Le modèle d'architecture du SNO PhytOBS est en cours de transposition pour le futur SNO BenthOBS.

3.6. SNO-LIKE BENTHOBS – (Vincent BOUCHET / Nicolas DESROY)

ND présente BenthOBS (voir 202009_ODATIS_Atelier_NDesroy_BENTHOBS.pdf[9]), qui est en cours d'incubation en vue d'une labélisation SNO. Cette présentation est axée sur la genèse de BenthOBS car la BDD BenthOBS sera calquée sur celle de PhytOBS (voir présentation précédente de MH). L'idée de la création d'un SNO est apparue en mai 2018 lors de l'atelier taxonomique RESOMAR à Brest, sous l'impulsion de la direction d'ILICO. Après consultation de la communauté, un dossier a été déposé en Avril 2019 à la campagne de labélisation OA. La demande a été rejetée mais la CSOA a accepté que cette demande soit réévaluée en 2021.

En ce qui concerne le contexte scientifique, la biodiversité benthique est une des sources premières des services rendus par les écosystèmes marins et les changements de l'environnement constituent une trame de pressions pour la biodiversité, qui modifie la composition des cortèges faunistiques. La macrofaune benthique des substrats meubles représente l'une des composantes fonctionnelles les plus importantes des écosystèmes côtiers. La diversité des communautés vivantes et les processus biologiques, chimiques, et physiques qui leur sont associés, contrôlent les processus biogéochimiques globaux au sein des écosystèmes.

L'objectif scientifique de BenthOBS est de qualifier l'état écologique de ce milieu et de surveiller l'impact anthropique qui est en augmentation et qui devient un sujet essentiel pour la gestion des mers européennes. Dans ce contexte, il est essentiel de disposer de séries temporelles capables de

[9]https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202009/202009_ODATIS_Atelier_NDesroy_BENTHOBS.pdf

mettre en évidence et de comprendre les changements en cours et leurs conséquences sur le fonctionnement des écosystèmes marins.

Les données existantes se trouvent, soit dans la BDD Benthos ILICO (anciennement nommée BDD RESOMAR), soit dans la base Quadrige pour les données relatives à la DCE, au REBENT ou à la DCSMM. Cependant le maillage actuel des données existantes est lâche et la fréquence d'échantillonnage est annuelle ou pluriannuelle, ce qui ne permet pas un suivi efficace.

Le projet de SNO BenthOBS vise à intensifier la fréquence de prélèvement (bi-annuelle) et d'obtenir une vision à l'échelle locale et nationale de la variabilité intra- et inter-annuelle et, à long terme des communautés macro-benthiques. Plusieurs objectifs scientifiques ont donc été fixés pour répondre à ces questionnements en particulier, il est nécessaire de :

- **comprendre / comparer** les trajectoires temporelles des communautés de la macrofaune benthique, (action coordonnée : « Traitement de la donnée » - resp. : O. Gauthier),
- **caractériser** la dynamique fonctionnelle de ces communautés (avec lien entre diversité taxonomique et fonctionnelle),
- **prédire** la trajectoire des communautés,
- **coupler** les données BenthOBS avec les données existantes sur le macro-zoobenthos (BDD Benthos ILICO, Quadrige-DCE/DCSMM) mais surtout avec les données issues d'autres SNO (Coast-HF, SOMLIT, PhytOBS), (action coordonnée : couplage entre les modèles physiques et les données BENTHOBS avec les données issues de SOMLIT, Coast-HF et PhytOBS - resp. : F. Orvain),
- plus généralement, dans un scénario d'érosion établie de la biodiversité, **acquérir** des connaissances naturalistes sur les espèces marines côtières de métropole, leur nombre, leur nature et leur dynamique, (action coordonnée : Recensement et suivi des espèces introduites et invasives - resp. : P.G. Sauriau).

Le futur SNO sera sous la responsabilité de VB et de ND, et rattaché à l'UMR8187 (LOG). La BDD est hébergée à Roscoff. Le SNO comprendra 9 partenaires (voir présentation diapo 6) et la description de sa structuration et de son comité exécutif se trouve sur la diapo 7. La diapo 8 permet d'avoir une vue générale de la structure de BenthOBS. Le réseau d'observation se base sur 18 stations avec des écosystèmes différents (côtiers, grands et petits estuaires, lagunes). Pour chaque station, 5 réplicats sont effectués pour la macrofaune et 1 réplicat lié au sédiment, ce qui représente environ 200 réplicats/an.

3.7. Quadrige – (Arnaud ROUILLY)

AR présente la BDD Quadrige de l'IFREMER (voir 202009_ODATIS_Atelier_ARouilly_QUADRIGE.pdf[10]) Quadrige est une BDD qui est née en 1996 suite au besoin de sécurisation et de partage des données entre le Réseau National d'Observation pour la chimie (RNO), le REPHY, le suivi des Infrastructures et Grands Aménagements (IGA) les installations nucléaires et la microbiologie avec le REMI. Depuis, le projet a évolué pour prendre en compte le besoin de la Directive Cadre Eau (DCE) puis la Directive Cadre Stratégie pour le Milieu

[10] https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202009/202009_ODATIS_Atelier_ARouilly_QUADRIGE.pdf

Marin (DCSMM). Actuellement cette BDD cumule jusqu'à 40ans de données de suivis et environ 25 ans d'évolution du modèle de données. Ce modèle de données démontre sa souplesse par la large gamme de thématiques bancarisées (voir diapo 4).

La version actuelle de Quadrige se nomme Quadrige2 (2010) et comprend des outils permettant intégrer des données et métadonnées dans la BDD Quadrige centralisée à Brest. D'autres outils, interfaces de saisies, comme BD Récif ou DALI permettent d'ingérer des données dans la base en mode on ou offline. Des outils ont aussi été mis en place pour extraire et diffuser des données (ex Surval). L'import des fichiers peut également se faire au moyen de fichier 'Excel normée par le référentiel du SANDRE'. Via Quadrige plusieurs référentiels sont gérés et permettent de faire du transcodage d'un référentiel à un autre (un référentiel propre à Quadrige et ceux du SANDRE, WoRMS, TAXREF, SAR/SIMM).

Les données sont de différents types :

- Mesures ponctuelles in-situ ou en laboratoire ;
- Observations in-situ faune et flore (présence/absence) ;
- Observations surfaciques (couches cartographiques) ;
- Fichiers issus d'analyseurs ;
- Photos.

Il y a environ 13 millions de mesures présentent dans la base qui compte quasi 550 utilisateurs dont la moitié sont extérieurs à l'IFREMER.

En ce qui concerne le cycle de vie de la donnée, la base permet des imports automatiques, une diffusion ouverte ou restreinte suivant l'existence de moratoire. Certaines communautés peuvent avoir des accès restreints. L'ensemble des données qui ne sont pas sous moratoire sera, d'ici la fin de l'année, basculé en « open data », dès lors qu'il s'agit de données acquises sur des financements publics. La partie contrôle est effectuée par l'opérateur qui a saisi les données. Une fois la donnée validée, elle ne fera plus l'objet de modification et ensuite les données sont qualifiées assez finement avec des commentaires de qualification. La particularité de Quadrige vient aussi de sa gouvernance (550 utilisateurs), il y a donc une comitologie avec des groupes utilisateurs et des comités de projet pour identifier et prioriser des évolutions nécessaires. Le projet migre vers Quadrige3 qui fera l'objet d'une attention particulière pour les interfaces de saisies qui seront adaptées à la grande diversité thématique des fournisseurs de données. L'accès pour les utilisateurs sera simplifié et beaucoup plus performant. Enfin, le projet Quadrige dispose d'un comité de pilotage transverse auquel participent les financeurs et les coordinateurs.

3.8. APA & SI MORSE : Le projet MORSE, vers un nouveau portail de suivi des échantillons biologiques – (Sylvie VAN ISEGHEM)

SVI présente le nouveau Système d'Information (SI) interne de référence pour le suivi des ressources et organismes marins (Marine Organisms and Ressources Storage systEM – MORSE - voir 202009_ODATIS_Atelier_SVanIseghem_MORSE.pdf[11]). MORSE est un nouveau portail qui va permettre de faire un suivi des échantillons biologiques avec un objectif à la fois réglementaire et scientifique. Le contenu de ce portail sera une carte d'identité de chaque échantillon biologique qui impliquera que les utilisateurs seront les scientifiques, les laboratoires et les services juridiques pour la conformité de l'utilisation de ces échantillons.

Au départ, le projet se nommait le SI APA (Accès aux ressources génétiques et le Partage juste et équitable des Avantages découlant de leur utilisation), projet établi en 1992 par la convention sur la diversité biologique et précisé en 2010 par le protocole de Nagoya. Cependant l'usage des échantillons biologique est impacté par d'autres réglementations comme la CITES. Le périmètre de MORSE s'est donc rapidement élargi.

L'APA a instauré une réglementation entre le pays fournisseur de la ressource génétique (RG) et le pays utilisateur. Entre autres, deux documents le Prior Informed Consent (PIC) et le Mutual Agreed Terms (MAT) sont obligatoires.

Le PIC est le permis d'accès à la ressource (consentement préalable donné en connaissance de cause par le pays souverain sur les RG). Il autorise l'accès à la RG et son utilisation.

Le MAT décrit les conditions convenues d'un commun accord selon lesquelles s'effectue le partage des avantages avec le pays souverain sur les RG.

Le projet MORSE est devenu un projet structurant à l'IFREMER avec du temps dédié et des instances de gouvernance (comité directeur, comité de pilotage, équipe projet et un groupe de travail sur la traçabilité avec des correspondant locaux). Le système MORSE assurera la centralisation des informations relatives aux échantillons biologiques d'Ifremer incluant : un système de référencement unique des échantillons biologiques et des informations liées aux obligations réglementaires. Ce portail ira chercher tant que possible les informations dans les autres SI de l'IFREMER (Quadrige, SISMER, LabCollector des laboratoires, …). Il permettra de faire des exports (avec moratoire possible). Le contenu de MORSE va représenter toutes les ressources biologiques et génétiques. Les référentiels utilisés seront ceux de SDN et pour les référentiels taxonomiques : WORMS, TAXRef, … Cependant le choix définitif des référentiels n'est pas encore fixé et il sera pris en concertation avec les différents partenaires. Le portail sera ouvert fin 2021 et l'exhaustivité des échantillons se fera en 2022.

La diapo 10 présente les champs identifiés pour assurer la traçabilité de l'échantillon (5 sections : origine, documents liés, filiation, RG, valorisation).

[11] https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202009/202009_ODATIS_Atelier_SVanIseghem_MORSE.pdf

A noter qu'il y a une volonté de positionner des DOI sur les échantillons comme les IGSN sur les carottes.

Le prototype du portail est consultable en suivant le lien : http://morse.ifremer.fr:8080/.

3.9. SAR/SIMM : Référentiels utilisés dans le cadre du Système d'Information sur le Milieu Marin – (Clémence RABEVOLO)

CR présente les référentiels utilisés dans le cadre du Système d'Information sur le Milieu Marin (SIMM – voir 202009_ODATIS_Atelier_CRabevolo_SAR_SIMM.pdf[12]). Le SIMM et le SAR ayant été présentés à l'atelier technique précédent, les auteurs de ce rapport encouragent les lecteurs à lire le compte rendu de l'atelier précédent (AtelierTechnique_Odatis_CR_201910.pdf[13]). Le rôle du SAR (https://sar.milieumarinfrance.fr/) est d'assurer l'interopérabilité des données du SIMM.

Les référentiels préconisés par le SIMM et le SAR :

- Référentiel taxonomique du SIMM :

  Le SAR a décidé de retenir le référentiel du SANDRE pour l'appellation taxonomique. A noter qu'un POC a été fait sur un outil de diffusion des référentiels SANDRE, TaxRef, WoRMS basé sur des technologies du web sémantique permettant le transcodage d'un référentiel à un autre Les résultats de ce POC sont encourageant.

- Référentiel des interlocuteurs du SIMM (interlocuteurs = producteurs de données) : le choix s'est aussi porté sur le référentiel du SANDRE qui est le plus proche des besoins du SIMM. Cependant pour le rendre opérationnel, il y a plusieurs opérations en cours (intégration des codes EDMO, création d'une liste « SIMM » des interlocuteurs SANDRE, création d'un outil de diffusion). La première utilisation de l'outil sera effectuée dans le cadre de Quadrige version 3.

- Référentiel PSFMU du SIMM (Paramètre, Support, Fraction, Méthode, Unité) : deux référentiels ont été étudiés : celui du SANDRE et celui du P01/P06 du BODC (SDN). Le SIMM a choisi de retenir le référentiel P01/P06 du BODC. Un transcodage entre le SANDRE et BODC est en cours avec une traduction en français des termes du BODC, mais aussi une intégration des traductions dans l'outils « NVS SKOS vocabularies » du BODC, la création des nouveaux P01 et enfin le développement des services d'accès aux traductions françaises des P01.

- Autres travaux en cours, référentiels géographiques :
  - Typologie des ouvrages maritimes
  - Référentiel des ports français (maritime et fluviaux),
  - Qualification des données du SIMM.

[12] https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202009/202009_ODATIS_Atelier_CRabevolo_SAR_SIMM.pdf
[13] https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_201910/AtelierTechnique_Odatis_CR_201910.pdf

Il existe aussi des référentiels géographiques sur le portail du SIMM (https://sar.milieumarinfrance.fr/Nos-rubriques/Referentiels-geographiques#/search?from=1&to=30). Une lettre d'information du SAR trimestriel est disponible et un abonnement est possible (https://sar.milieumarinfrance.fr/A-propos/Lettre-d-information).

3.10. Cytométrie – (Gérald GREGORI, Felipe ARTIGAS, Maurice LIBES, Marc Sourisseau, Melilotus Thysen)

3.10.1. Qu'est-ce que la cytométrie en flux (Diapos 1-10)

GG présente la première partie de la présentation sur la cytométrie (voir 202009_ODATIS_Atelier_cytometrie.pdf[14]). La cytométrie en flux provient d'une technique médicale destinée au dénombrement et à l'analyse des cellules sanguines. Cette méthode a été adaptée (vers 1970) pour la détection et le dénombrement des microorganismes planctoniques sur des volumes plus importants. C'est une technologie qui utilise les propriétés optiques des cellules. Chaque cellule est analysée séparément. Elle est placée dans un ou plusieurs faisceaux laser dont on analyse la diffusion des photons ce qui permet d'avoir des informations sur la taille et la forme des cellules. Le faisceau peut aussi être absorbé par la cellule et la cellule peut réémettre sa propre lumière, ce qui permet d'étudier la bioluminescence des cellules (et la photosynthèse pour le phytoplancton). La cytométrie permet d'avoir la classe de tailles des organismes, ce qui identifie uniquement le groupe de l'espèce et non l'espèce elle-même la plupart du temps. En revanche, elle permet de travailler sur une grande plage de taille des organismes de 0.3 µm jusqu'à des chaînes pouvant atteindre plusieurs centaines de micromètre (ce qui représente environ 3 à 4 ordre de grandeur).

La cytométrie a permis certaines découvertes notables comme les Prochlorococcus et l'Ostreococcus tauri.

Les applications de la cytométrie permettent de faire des suivis sur des stations fixes de différentes espèces (par ex., voir les résultats sur le réseau SOMLIT – Station SOLEMIO) mais aussi sur des échantillons d'eau de mer de campagnes océanographiques (abondance et structure des microorganismes marins).

Il existe aussi des instruments de cytométrie automatisés in-situ (FlowCytobot, Cytosense flow, Sea Flow cytometer) qui peuvent être déployés de plusieurs jours à plusieurs mois directement sur le terrain via un navire, une bouée ou sur un submersible et faire des cartographies de zone. Ces instruments deviennent de plus en plus populaires car leurs coûts diminuent, leurs capacités ont augmentées (analyse de plusieurs milliers de cellules/s) et les données renvoyées sont numériques.

La nécessité d'harmoniser et de fédérer la communauté cytométrie en flux a fait émerger le besoin de créer un CES cytométrie à ODATIS dont l'objectif est de bancariser de façon homogène et harmonisée l'ensemble des données de cytométrie récoltées en France (voir https://www.odatis-ocean.fr/activites/consortium-dexpertise-scientifique/ces-cytometrie-en-flux).

[14] https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202009/202009_ODATIS_Atelier_cytometrie.pdf

3.10.2. Cytométrie en flux automatisée et accessibilité des données pico-nano-microplancton (Diapos 11-30)

FA présente la partie cytométrie automatisée ainsi que la partie accessibilité et vocabulaire.

Il existe actuellement des instruments permettant d'utiliser la cytométrie de manière in-situ ou in-vivo par le pompage d'eau de mer à partir de navire, de station fixe ou de bouée et qui est directement analysée.

La diversité fonctionnelle (taille, structure, dynamique, abondance, mobilité) joue un rôle essentiel dans la productivité et la façon qu'a le carbone organique de circuler dans le réseau trophique et l'exportation via la pompe biologique. Sa variabilité peut être liée à des changements de conditions environnementales au niveau global, régional ou local. La cytométrie en flux permet d'accéder à ce niveau de diversité fonctionnelle et son automatisation, permet d'approcher la haute fréquence et la haute résolution spatiale. Mais un grand nombre de données reste souvent inaccessible à la communauté parce qu'il n'existe pas :

- De vocabulaire commun ;
- De système interopérable ;
- De bases de données nationales /internationales.

Il est donc nécessaire de se diriger vers une interopérabilité des mesures automatisée au niveau des observatoires côtiers à l'international. Pour ce faire, plusieurs projets internationaux ont vu le jour cette dernière décennie :

- Projets d'inter-calibration et inter-comparaison des résultats de mesures de différents appareils avec différentes configurations sur des cultures et lors de campagnes en mer internationales (projets INTERREG IV A "2 Mers" DYMAPHY 2010- 2014 et H2020 JERICO NEXT 2015-2019),
- Travaux sur un vocabulaire interopérable pour permettre aux données issues de la cytométrie en flux d'intégrer des bases de données marines (projet SeaDataCloud : accessibilité via les portails de gestion de données marines – 2016-2020),
- Vers une meilleure interopérabilité de l'utilisation de la technique de cytométrie en flux automatisée (CES Cytométrie ODATIS, JERICO S3 et autres projets), avec comme objectif :
    - La définition de procédures opérationnelles communes aux utilisateurs de la technique au niveau de la communauté scientifique,
    - L'automatisation des analyses des signaux (supervisée, semi-supervisée et non supervisée) en développant des outils et en explorant d'autres. Proposition d'accès virtuels à certains outils,
    - L'avancement dans la définition des formats et qualification des données pour la mise en place de flux (workflows) alimentant les bases de données.

Suite à l'étude du MIO (Projet Chrome 2015-2016 – S. Lahbib et SDC), sur le vocabulaire, une liste des paramètres à exporter a pu être identifiée :

- Abundance (cell.cm-3);
- Functional group names;
- Average Red Fluorescence;

- Standard deviation Red Fluorescence;
- Average Orange Fluorescence;
- Standard deviation Orange Fluorescence;
- Average Side Ward Scatter (Area, length);
- Standard deviation Side Ward Scatter (Area, length);
- Average Forward Scatter (Area, length);
- Standard deviation Forward Scatter (Area, length);
- Other.

Cependant il est encore nécessaire de finaliser le vocabulaire ainsi que de mettre en place le workflow, les métadonnées, la gestion des données campagnes/institutionnelles. Le travail à effectuer dans JERICO S3 doit aussi intégrer des bonnes pratiques et du contrôle qualité. L'ensemble de ces tâches fait partie des sujets à traiter au CES Cytométrie d'ODATIS.

3.10.3. Avancement chaîne de traitement de cytométrie en flux (Diapos 31-45)

ML présente les avancés de la chaîne de traitement de cytométrie en flux développée au MIO. Dans un premier temps, il a été nécessaire de faire une étude comparative des worflows existants sur la cytométrie en flux en particulier ceux du MIO et ceux du Laboratoire d'Ecologie Pélagique de l'IFREMER (Marc Sourisseau). Cette étude a permis d'établir des recommandations pour le worflow à soumettre au CES Odatis en vue de leur utilisation dans les divers centres qui utilisent la cytométrie en flux. Cette étude est accessible en suivant ce lien :

https://docs.google.com/document/d/1MrRDDRzuA4thz3evZ0kyqsG4RKAjpOFk14EAXiqaM0M/edit?usp=sharing

A la suite de cela, la chaîne de traitement de cytométrie a complètement été réécrite en Python avec un code maitrisé de bout en bout, l'insertion d'un contrôle qualité sur tous les paramètres et la prise en compte des cytomètres : Cytoclus3 et Cytoclus4. La Fig. 3 représente un résumé de ce workflow.

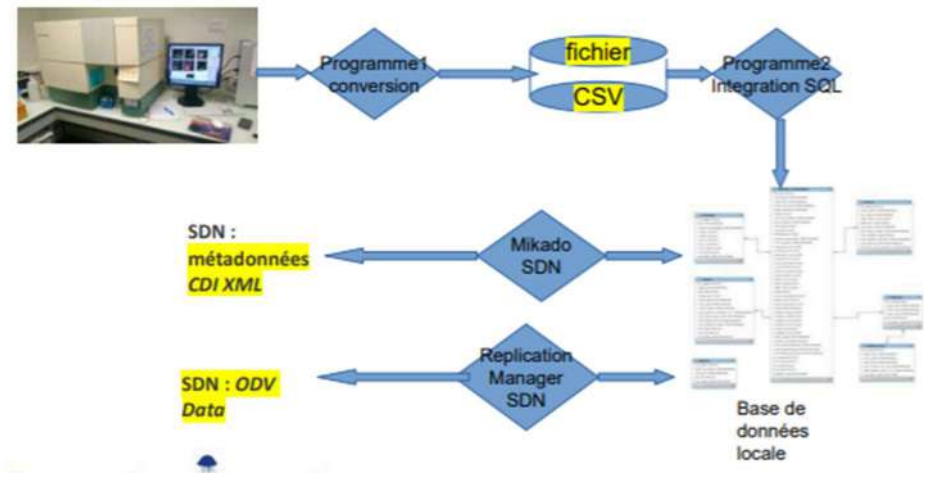

Figure 3: Résumé du workflow avec les différentes étapes communes importantes0

L'étude s'est aussi portée sur la création d'un fichier CSV standardisé en sortie de cytomètre avec un ensemble de métadonnées et de paramètres communs (voir diapos 33, 34, 35 pour les noms de paramètres). A noter que le vocabulaire interopérable pour la cytométrie est en cours de publication

et issue d'un consensus entre 32 experts internationaux. Après ingestion dans une BDD de ces fichiers, l'export doit se faire avec des fichiers dont le format interopérable (ODV ou NetCDF) permettra de les envoyer à SDN ainsi qu'à d'autres bases internationales. Pour cela, il est aussi nécessaire de créer des fichiers d'export des métadonnées au format XML. La table de vocabulaire recommandée est celle de SDN.

MS présente la partie réalisée dans le cadre de DYNECO/PEL pour la période allant de juillet 2020 à mars 2021 sur la chaîne de traitement de l'IFREMER et sa BDD associée. Plusieurs chantiers sont prévus en particulier :

- Intégration des images associées à l'imagerie en flux dans la BD locale (fini) ;
- Intégration des données de deux suivis temporels (Octobre) ;
- Mise en place de procédure d'annotation (Novembre) ;
- Mise en place de procédure d'export de ces images vers des plateformes d'imagerie (Janvier).

La Fig. 4 représente le workflow mis en place à l'IFREMER.

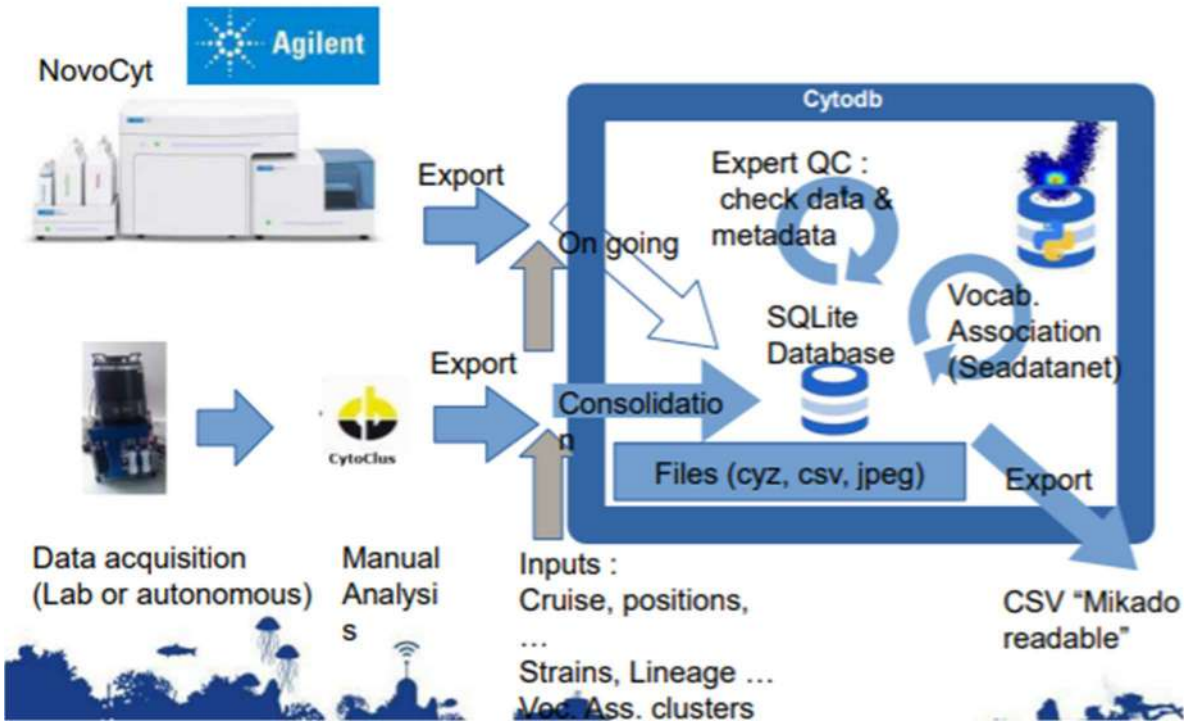

Figure 4: Workflow mis en place à l'IFREMER

4. Atelier de discussions

La seconde journée de l'atelier ODATIS est consacrée à initier des discussions autour de certains thèmes afin d'arriver, *in fine*, à avoir des recommandations sur la gestion des données de biologie marine. L'ambition du pôle, au travers de ces ateliers de discussions, est d'apporter des solutions, quand il y a besoin, et surtout de ne pas rajouter de la complexité ou une surcouche inutile. Il ne faut surtout pas « réinventer la roue », mais au contraire trouver des solutions pour faciliter le travail des producteurs de données. Les quatre thèmes abordés sont les référentiels, les formats de données, les outils et enfin l'imagerie. Il a été consacré environ 1h pour chaque thème. Les sections ci-dessous représentent la synthèse des échanges de cette journée.

4.1. Référentiels

Afin de rendre FAIR et en particulier interopérable les données de biologie marine, au niveau national et international, il est nécessaire d'utiliser un langage et un vocabulaire communs en s'appuyant sur les référentiels internationaux qui sont incontournables. Les référentiels sont donc dans le cas présent un moyen pour une ou plusieurs communautés de définir avec précisions les termes qu'elle emploie. Cependant, dans le cas de la biologie marine il faut tenir compte de l'évolution très rapide de la taxonomie et de la phylogénie. Il est donc nécessaire pour le référentiel de pouvoir évoluer rapidement en prenant en compte l'arrivée de ces nouvelles informations. Au-delà du référentiel lui-même, c'est toute la donnée (et la métadonnée associée) qui doit être révisée pour prendre en compte ces évolutions. Il y a donc un besoin d'évolution et d'adaptation systématique des bases de données mais aussi de leur contenu, ce qui rend extrêmement complexe leur mise en place et leur alimentation, mais aussi leur export vers des bases internationales. Ajouter à cela qu'il existe différents référentiels pour la classification en biologie marine que l'on peut ranger en deux grands types :

- Les référentiels Taxonomiques « scientifiques » comme WoRMS et Taxref qui sont une vision exhaustive de la connaissance scientifique actuelle. Ils référencent les espèces ayant fait l'objet d'une publication scientifique et mettent à jour leur référentiel au fur et à mesure des avancées scientifiques.
- les référentiels « opérationnels » comme celui des « appellations de taxons » SANDRE retenu par le SAR, qui référencent seulement les espèces présentent dans les données et donc demandées par les utilisateurs. Les référentiels opérationnels reposent sur des référentiels scientifiques pour être à jour des connaissances scientifiques, mais permettent aussi aux utilisateurs de créer des espèces qui n'auraient pas encore fait l'objet d'une publication (nouvelle espèce découverte par exemple). Cela permet de bancariser des données sans créer de codes transitoires. Les référentiels opérationnels permettent donc aux utilisateurs d'obtenir un code rapidement, tandis ce que les référentiels scientifiques doivent attendre la publication d'un article. A côté de cela, le Sandre propose aussi un référentiel des « groupes d'appellations de taxons » qui créer des groupes d'espèces sans réalité phylogénétiques mais permettant de bancariser les données de suivis n'ayant permis d'aller jusqu'au niveau de l'espèce pour des espèces très proches morphologiquement (groupe de plusieurs espèces d'ulves par exemple).

Pour plus d'informations sur la différence entre les référentiels scientifiques et opérationnels, vous pouvez consulter le document de conclusion sur les référentiels taxonomiques du SIMM : https://sar.milieumarinfrance.fr/A-propos/Publications-et-supports-de-presentation/Conclusion-du-travail-sur-le-referentiel-taxonomique-du-SIMM

A noter que le GBIF a créé son propre référentiel « GBIF Backbone Taxonomy », inspiré de « Catalogue of Life » avec plusieurs checklists dont WoRMS et que ce nouveau référentiel apporte des nouveautés par rapport à WoRMS. Même si cette classification n'a pas de réalité taxonomique, elle est souvent nécessaire et importante pour certaines communautés. On peut citer pour exemple la classification par groupe pigmentaires (classification plutôt chimique) à partir desquels on peut déterminer des taxons correspondant à des données biologiques.

Il est donc nécessaire, si on veut être le plus exhaustif possible, de tenir compte de l'ensemble de ces référentiels et d'arriver à transcoder (au mieux !) les uns dans les autres, ce qui est loin d'être simple et trivial ! De plus, **Il est donc nécessaire de déterminer non pas un référentiel, mais toute une liste de référentiels pour arriver à gérer les données de biologie marine**. La discussion débouche sur une ébauche de liste des référentiels nécessaires a minima :

- Référentiel taxonomique de type WoRMS ;
- Référentiel taxonomique ad hoc, avec des groupes morphologiques de type SANDRE ;
- Référentiel sur les instruments ;
- Référentiel sur le niveau d'expertise ;
- Référentiel sur les méthodes d'observations, d'échantillonnages et procédures ;
- Référentiel sur les limites de détection, de saturation de l'appareil, des limites de détection ;
- Référentiel de calibration des instruments (n° de série, …) ;
- Référentiel sur les habitats ;
- Référentiels sur les traits de vie ;
- Référentiels des paramètres (taille, poids, …).

Enfin il est nécessaire, de conserver dans les bases, une traçabilité, un historique et une gestion de versions qui permettront de faire évoluer la classification d'une donnée. En effet un expert peut se prononcer sur le rattachement de l'observation qu'il a faite avec tel élément du référentiel. A l'occasion d'une révision du référentiel ou par rapport à d'autres éléments nouveaux de contexte, il peut se prononcer de manière différente et notamment dans les liens entre taxons et substrat, ce qui est complexe à gérer. Il faut donc pouvoir revenir sur des associations qui ont pu être faites à un moment donné. **Il est donc impératif de garder la traçabilité et la gestion de version dans les bases de données**.

Si cette première discussion sur les référentiels n'a pas permis d'émettre des recommandations, il est tout de même possible de consulter le document de conclusion sur le référentiel taxonomique du SIMM (Conclusion_du_travail_sur_le_referentiel_taxonomique_du_SIMM.v2.2.pdf[15]). En revanche cette discussion a permis de soulever les interrogations et pistes listées ci-dessous :

- Comment faire évoluer certains référentiels et comment prendre en compte ces évolutions dans les bases de données les utilisant ?
- Comment gère-t-on la classification taxinomique des données ? deux modes seraient envisageables : « expert » utilisant WoRMS en l'associant au niveau d'expertise de l'observateur, et « groupe » dans le cas d'observation plus automatisée utilisant les référentiels SANDRE ou vocabulaires BODC, le contenu des listes de « groupe » pouvant différées d'une équipe d'observation à l'autre.
- Comment gère-t-on les correspondances entre les référentiels purement taxonomiques (comme WoRMS) et les référentiels un peu plus ad hoc ou liés à un système d'observation ? il y a un besoin technique important d'avoir des « mapping » élaborées entre ces référentiels.
- Est-ce que par exemple, ODATIS peut intervenir ?
- Est-ce que les référentiels/outils mis en place par le SANDRE peuvent être adoptés par ODATIS ?
- Prise en compte de la génomique ?

En conclusion, il reste énormément de travail à faire sur ces référentiels avant d'avoir des recommandations à destination des producteurs de données. Il est nécessaire d'identifier des correspondants afin de faire du « mapping » entre les référentiels. Trouver un lieu pour discuter de ce thème semble primordial car il ne semble pas qu'au niveau européen des groupes travaillent sur ce sujet.

Dictionnaire de donnée c'est un document XML ou pdf qui décrit le modèle de données, les champs. Implementation de la norme ISO19110

[15] https://sar.milieumarinfrance.fr/content/download/4603/file/Conclusion%20du%20travail%20sur%20le%20r%C3%A9f%C3%A9rentiel%20taxonomique%20du%20SIMM.v2.2.pdf

4.2. Outils

Il existe de nombreux outils que l'on peut dissocier en deux grandes catégories en suivant le cycle de vie de la donnée :

- Les outils utilisés en amont de la création du fichier de données, qui permettent la saisie de la donnée, la mise en format, ou remplaçant le cahier de paillasse, qui permet de prendre des notes. Il existe aussi des outils permettant de bancariser les données. A noter que chaque domaine a sa classe d'outils, ce qui rend difficile la mise en place d'outils « standardisés » pour l'ensemble de la communauté océanographique/océanologique. Le but est donc de voir si on peut améliorer cette situation très hétérogène en proposant des outils utilisables en mode online et offline permettant de traiter la donnée en amont de la bancarisation.
- Les outils en aval qui sont utilisés pour intercomparer les données, pour traiter la donnée, pour rapprocher les données satellitaires et in-situ et aussi pour faire des analyses plus poussées permettant d'utiliser l'IA ou autres algorithmes innovants (éventuellement CES couleur de l'eau).

Doit-on se diriger plutôt vers la mise à disposition d'outils ou plutôt vers des formats pivot avec l'utilisation d'outils bureautiques mais avec des recommandations et des gabarits prédéterminés ?

La communauté autour de la base de données Pelagos, a choisi d'utiliser les outils bureautiques en créant des gabarits Excel de téléversement, permettant l'ingestion de ces fichiers dans la base. Il est donc possible de consigner dans un fichier l'ensemble des observations d'un site. Quant à la base PhytOBS, elle est nourrie via des exports des bases Pelagos et Quadrige au moyen de fichiers au format pivot BioODV.

Pour BenthOBS, la nature des données à ingérer dans la base est plus complexe car les utilisateurs veulent intégrer à la fois des données de macrofaune mais aussi des informations sur la granulométrie du sédiment, ou autres. Il a donc été choisi d'avoir un gabarit Excel contenant plusieurs onglets où chaque onglet est dédié à une catégorie de donnée ou de métadonnées. La réflexion pour importer les données de la base Quadrige vers la base BenthOBS sont toujours en cours car le format Excel n'est pas une solution envisageable, ce format étant trop complexe. Par contre, une extraction via un CSV avec l'ensemble des champs nécessaires est fortement envisagée. Il reste donc à identifier l'ensemble des paramètres à extraire.

Pour la base Quadrige, les outils développés pour l'ingestion des données (Quadrige2) deviennent obsolescents et le développement de nouveaux outils (Quadrige3) est en cours d'initialisation. Ces outils beaucoup plus thématiques permettront une meilleure supervision. A terme, il pourrait être utile au moment de la saisie et presque en temps réel, de faire un retour d'informations sur le positionnement de la mesure au regard des séries temporelles passées pour capter d'éventuelles informations de contexte très utiles à la qualification par la suite et/ou de limiter les erreurs de saisies.

L'utilisation d'outils tel que « LIMS » (http://www.lims.fr/) ou « LABCOLLECTOR » (https://www.labcollector.com/), qui sont des cahiers de paillasse automatisés, n'est pas forcément très courante et l'interfaçage de ces outils avec les instruments de mesures est encore très peu développé (il existe quelques instruments qui peuvent s'interfacer mais ce n'est pas la majorité). A noter, tout de même, que des développements dans LIMS sont en cours pour intégrer des référentiels sous forme de dictionnaire afin qu'ensuite ils puissent être moissonnés d'une instance de LIMS à une autre. Ces développements permettront d'avoir un langage commun entre le producteur de la donnée et l'institution qui va bancariser cette donnée, mais aussi d'avoir plus d'interaction au moment de la saisie de l'information et de permettre d'ingérer plus d'information.

Dans la proposition PIA3 GAIA DATA, la mise en place de ce type d'outil, afin d'harmoniser les entrées, est prévue. L'objectif est donc d'identifier ces outils car il est primordial d'arriver à résoudre le problème de la grande hétérogénéité en amont de la bancarisation de la donnée. Il est donc important de converger vers des méthodes/outils en nombre limité et que ces outils soient « user-friendly ». Il aussi prévu, dans GAIA DATA, de déployer sur les 8 centres structurants des VRE (Virtual Research Environment) et des VAP (Virtual Analysing Plateform) pour l'ensemble de la communauté scientifique. Celles-ci permettront d'avoir accès, de façon ergonomique, à des capacités de calculs importantes déportées sur les centres HPC et HPDA. Ce type de plateforme est facilement déployable sur des HPC, mais aussi en local sur PC personnel ou dans les laboratoires. Ce qui va permettre de développer les codes en local et de les déployer ensuite sur des HPC et ce, sans modification de code et en utilisant la containerisation, la parallélisation massive permettant de travailler sur de grands ensembles de données.

Au moment de la collecte de la donnée, il est nécessaire d'acquérir le plus de métadonnées possibles, car celles-ci pourront être nécessaires a posteriori (ex. 25 plus tard) lors d'avancées majeures dans la thématique considérée (éventuelle rupture liée à l'évolution du milieu).

L'un des objectifs d'Odatis est de mettre en place ce type de pipelines « standardisé » pour faciliter la saisie, l'analyse et le traitement de données.

4.3. Imagerie

Actuellement, en océanographie/océanologie (mais aussi dans d'autres domaines), la thématique qui est en train de véritablement d'exploser en termes de volume de données acquises est l'imagerie (photos et vidéos). Il existe de plus en plus de systèmes d'acquisition peu couteux et avec des résolutions d'image de plus en plus importantes. Cette multitude d'instruments apporte aussi une multitude de formats (propriétaires ou pas) et une multitude de logiciels permettant le traitement de la donnée brute acquise par le capteur. De plus, l'imagerie dans son ensemble ne correspond pas à une seule communauté mais à un ensemble de communautés, car il existe de nombreuses sous-catégories d'images allant du microscope aux satellites (voir section 3.3 et présentation de CB : 202009_ODATIS_Atelier_CBorremans_Imagerie.pdf[16]). Il y a donc une certaine

[16] https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202009/202009_ODATIS_Atelier_CBorremans_Imagerie.pdf

urgence à prendre en compte ce nouveau défi avant que la quantité de données ne devienne trop volumineuse pour être ingérée.

En termes d'outils, l'outil BIIGLE ressort au niveau de la communauté marine (benthos et plancton), et aussi le besoin d'avoir une suite d'outils interopérables. Pour ce qui est des formats, il en existe plusieurs (propriétaires ou pas) provenant de divers capteurs. On peut noter qu'il existe tout de même quelques formats pivots pour la vidéo (H264mp4) et l'image (RAW, jpeg2000), mais qu'il manque un consensus à ce niveau. Pour l'imagerie du plancton, il n'y a aucune cohérence car chaque constructeur d'instrument sort son propre format.

Si on se place dans la position d'un centre de données qui va gérer des images ou des vidéos**, alors** il est nécessaire de répondre à **plusieurs questions** :

- Quels types d'image va-t-on recevoir (Image brute, Imagette traitée, …) ? Que faut-il archiver et dans quels formats ?
- Comment traiter ces images (ou vidéo) ? En amont d'ODATIS ou est-ce que les CDS vont y contribuer ? A quels formats ? Avec quel compromis avec des outils de compression (algorithme sans pertes d'informations dans l'image) si le volume est trop important ?
- Que va-t-on devoir conserver après le traitement ? comptage ? espèce ? mosaïque d'image ?
- Est-ce qu'en plus des images ou vidéos, il y a d'autres choses à conserver ((par ex couverte géographique de type SIG pour des mosaïques d'image)?
- Où ODATIS et ses CDS doivent intervenir ?

Il y a différentes étapes où les centres de données sont importants pour les utilisateurs des données d'imagerie, avec un « pipeline » à mettre en place en particulier pour :

- Les données brutes : **sauvegarde centralisée** nécessaire. Mais il reste à définir à quel format ? Compressées ou pas ? etc. ;
- Les pré-traitements qui **demandent des capacités de calcul** importants (transcodage, mosaïque et 3D) ;
- L'analyse des images qui dépend fortement des thématiques scientifiques. L'analyse doit être automatisée au maximum, et **profiter de capacité de calcul à distance**, avec une phase d'analyse par les experts en labo ou en collaboratif sur des outils de type web ;
- Les bases de référence pour l'algorithmie.

Concernant la volumétrie à gérer, l'imagerie quantitative est assez différente de l'imagerie microscopique (du médical mais aussi appliqué à la cytométrie par exemple). L'imagerie quantitative produit beaucoup d'images peu volumineuses, à la différence de l'imagerie microscopique qui elle génère moins d'images, mais très volumineuses. L'imagerie microscopique a donc une problématique de volumétrie de stockage.

Cependant, l'imagerie quantitative qui s'applique au plancton, benthos, etc. extrait à partir d'image ou de vidéos des imagettes puis des données dérivées. Si on considère uniquement les données dérivées (espèce, concentration, abondance, …) cela représente quelques dizaines voire centaines de Go. Pour les données brutes avec les métadonnées associées c'est plutôt de l'ordre de 200 Go à

l'échelle de la communauté française. Si on considère les images extraites individuelles, le volume représente quelques To. Mais pour les images brutes avant le traitement d'identification des organismes, il y a un facteur 100 par rapport aux images individuelles (ex. uniquement pour Villefranche environ 100To, environ 10Po pour la communauté française).

Cette problématique de la volumétrie en augmentation constante est un point d'attention sur lequel il faut déjà se focaliser actuellement. En effet, plus nous attendons avant de traiter ce problème de volumétrie, de son stockage et d'archivage, plus le passif deviendra grand et plus le travail a posteriori, une fois qu'un consensus aurait été acté, sera conséquent.

Concernant les traitements de l'imagerie quantitative, le passage de l'image brute à l'image individuelle d'organisme n'a actuellement absolument aucune standardisation, et, il semble utopique, pour le moment, de standardiser cela au niveau national et de sortir cette phase des laboratoires. Dans les laboratoires, il existe différents protocoles et formats. De l'image individuelle à la donnée individuelle (extraction des informations de l'image), cela se standardise un peu plus, mais ce n'est pas encore tout à fait le cas. Enfin pour le passage de la donnée individuelle aux données écologiques, le **logiciel EcoTaxa** centralisé permet de fédérer la communauté, mais d'autres outils existent et sont utilisés d'autres pour d'autres communautés, cependant ces outils devraient aussi être centralisés.

Pour la partie pipeline/workflow/chaine de traitement des vidéos sous-marines, Biigle fédère une certaine communauté autour du benthos. Pour le moment, le traitement tourne plutôt en local, il faudrait envisager de monter en charge en passant par une plateforme nationale.

A noter que, pour la cytométrie, les premiers cytomètres permettant de faire de l'imagerie commencent à apparaitre et vont avoir tendance à se développer dans les années à venir. Ils produisent des imagettes assez similaires à ce qui est ingéré dans EcoTaxa, donc EcoTaxa pourrait aussi être utilisable par cette communauté. Dans Jerico-S3, il y a le projet de mettre les imagettes du cytomètre dans EcoTaxa. Ces imagettes, générées par les cytomètres, ont 500 champs de libre qui peuvent accueillir des données annexes (comme les profils optiques par exemple).

Concernant l'archivage, est-ce nécessaire aujourd'hui de le faire et rapidement ? Oui, et plus on attend plus cela va être compliqué ! Cependant pour que l'archive soit pérenne et qu'elle soit réutilisable, il est nécessaire de connaitre les formats adéquats pour cet archivage. En termes de moyens, via les grandes infrastructures informatiques (CINES, IN2P3, etc.) ou les infrastructures d'organismes (CNES, IFREMER, …), il y a la capacité d'archiver quelques Po voire un peu plus pour de l'archive statique (environ 100Po sur robot avec bande).

ODATIS peut être un facilitateur pour les laboratoires à la mise en place, au niveau national (en coopération avec IFB) et avec les grandes infrastructures, d'une procédure de backup des données brutes puis, dans un deuxième temps, une archive de la chaîne complète. (ex : Le CINES conserve l'herbier historique du MNHN). Il est aussi possible d'étudier la possibilité d'utiliser au niveau national DATARMOR avec une instance EcoTaxa (besoin déjà identifié pour les scientifiques de l'IFREMER). **Une action est donc à engager à ce niveau entre ODATIS et les différents scientifiques**

impliqués dans les laboratoires (action assez urgente à faire avant la fin de l'année avec l'IR Data Terra).

Dans ce cadre, GM propose donc **qu'un groupe constitué de FA, GG, JOI, (et FC, Jean-Baptiste Romagnan pour l'instance d'EcoTaxa à l'IFREMER) recense, rapidement, les besoins actuels (stockage, archive, capacité de calcul), ainsi que la dynamique de son évolution et de le transmettre à GM et JS pour interagir ensemble avec les infrastructures (IFB). L'inventaire devra également être effectué pour l'imagerie sous-marine par CB.**

Pour l'imagerie in-situ benthique, l'utilisation de BIIGLE semble s'imposer mais ce n'est pas le seul outil et **les différents workflows seront donc à étudier dans le cadre du projet AMII**.

JOI précise que les **données finales** sont envoyées uniquement à EMODnet Biology. GM précise qu'il faudrait les avoir aussi au niveau national sans compliquer la procédure en faisant deux flux. Donc la **solution à valider/discuter avec MH** : PELAGOS serait le gestionnaire français des données finales et pourrait les transmettre à EMODnet Biology. JOI serait d'avis de garder une copie locale des données brutes dans chaque laboratoire ou SNO ou CDS (échelle correcte à définir), car les méthodes évoluent et un retraitement de la donnée peut être nécessaire. De plus, les données brutes peuvent être utilisées par d'autres communautés avec un champ d'étude différent.

Il faut également avoir une sauvegarde (backup) de cette donnée (brute) dans un centre différent et/ou centre national (de type CINES ou un CDS), ceci afin d'éviter la perte de ces données en cas d'incident grave sur le site d'origine. Il est nécessaire de se poser la question de quelle granularité de la donnée brute il faut conserver. En effet, cela peut aller d'un facteur 1 (image brute non traitée) à un facteur 100 (vignettes issue de l'image brute d'origine).

4.4. Format de données

Pour ODATIS, l'enjeu est de proposer et de fournir aux utilisateurs des formats décrits sur le site web et, si possible, qu'ils soient compatibles avec ce qui est fait au niveau international pour faciliter le travail de nos utilisateurs. Encore une fois, il ne s'agit pas de réinventer, mais de donner une bonne vision de ce qui se fait et de fournir des outils ou des routines afin de mettre les données dans les formats préconisés.

Pour la physique et la biogéochimie, les formats recommandés sont le NetCDF V4 et/ou le format CSV (ODV avec entête SDN). Qu'en est-il pour les données de biologie marine ?

Pour les données de biodiversité, il existe deux aspects sur les formats de données : format de gestion et format d'échange.

L'aspect format de gestion de donnée est assez classique, avec un courant aussi pour le NetCDF. Cependant les outils et librairies pour créer les fichiers sont très orientés algorithmes python et les biologistes ne sont pas du tout familiers avec leurs utilisations. Dans la communauté biodiversité, les utilisateurs sont plus familiers du langage R avec des données formatées en fichier CVS tabulés et/ou fichiers de type Excel.

En format d'échange,

- un standard semble tout de même s'imposer : le DarwinCore Archive. Celui-ci est vraiment un standard d'échange et il est vivement recommandé de conserver le format de gestion des données utilisées en amont de la phase d'échange. Le DarwinCore Archive est un fichier zippé contenant à la fois des données au format texte et des métadonnées associées au format XML décrit en EML. Son intérêt est de partager plusieurs types d'information : des données d'occurrence (comptage, abondance), des données taxonomiques et des données d'échantillonnage. En plus de ces 3 cœurs de standard, il est aussi possible d'avoir des extensions du standard (comme Audubon Core pour les images). Le DarwinCore Archive se rapproche du concept de data package un peu comme le NetCDF associe les données et les métadonnées.
- Il existe aussi le BioODV (utilisé pour alimenter les bases PhytOBS, BenthOBS, …) créé dans le projet SDN par le VLIZ. A partir du format BioODV il est possible d'exporter en format DarwinCore Archive. Par contre, la passerelle inverse n'est pas encore totalement opérationnelle avec des outils mis à disposition. Est-ce que ce profil DarwinCore Archive généré à partir du fichier BioODV est suffisant pour échanger avec la communauté de biodiversité (PNDB, GBIF, …) ? La réponse semble être oui, mais une étude plus approfondie semble nécessaire pour vraiment être certain de cette compatibilité. A noter que le PNDB, s'intéresse aussi au format BioODV et qu'il pourrait intégrer dans METASHARK une possibilité d'import/export dans ce format et ainsi fournir une passerelle à double sens entre ces deux formats. AR fait remarquer que le BioODV a tout de même des limites rencontrées lors de l'export des données de Quadrige vers EMODnet. En effet, ce format ne permet pas d'exporter plusieurs profondeurs dans un même fichier car l'information de profondeur n'est pas indiquée (obligation de scinder le jeu de données en plusieurs jeux à profondeur constante). Ceci est un point de vigilance à garder en mémoire.

Il faut aussi avoir conscience que chaque format possède des particularités en termes de structure de fichier et qu'il est nécessaire d'adapter le jeu de données à ce format, ce qui peut s'avérer délicat car certaines informations du jeu de données peuvent être altérées lors de l'export afin de rentrer dans le format... (ex : méthode et protocole non standard, arrondi des données, etc.). Il se peut aussi que la granularité du format entraîne l'intégration de biais, qui ne sont pas explicitement notifiés, et qui, à la longue, risquent d'être perdus par manque de traçabilité. Il y a donc une extrême vigilance à avoir lorsque des jeux de données transitent d'un entrepôt à un autre, car leurs contenus peuvent être altérés. (NDLR : intégration dans les entêtes de fichier d'un paramètre de traçabilité, de conformité avec le jeu original, de la base de données source, modification de la donnée à posteriori ou autre… nécessité pour le « end-user » d'avoir connaissance qu'une altération du jeu a été effectué et qu'il est nécessaire d'aller vers la base de données source si nécessaire). (voir IPT du GBIF : https://www.gbif.org/ipt – gestion de version).

Du coté ODATIS, au vu des volumes (assez faibles) de ces bases de données, il est tout à fait envisageable de faire des exports périodiques (snapshot) des bases et de leur attribuer un DOI, ce qui permettrait une traçabilité et une gestion des versions (comme cela est fait pour les données de physique : Argo).

Pour les producteurs de données, il est nécessaire que les CDS proposent deux mécanismes pour rendre disponible la donnée :

- Un mécanisme par téléchargement de fichiers ;
- Un mécanisme dynamique par service de requêtage de la base de données.

Dans l'idéal ces deux services doivent être possibles.

Pour ODATIS, les deux formats, BioODV et DarwinCore Archive, devraient être disponibles. En effet, le BioODV est un format plus facile à utiliser par les utilisateurs finaux (fichier à plat), et le DarwinCore Archive semble plus pertinent pour les échanges entre les différents systèmes. A noter que le NetCDF (conteneur trop complexe) semble ne pas être pertinent pour ces communautés et qu'il n'est pas un choix retenu pour les données de biologie marine à ce jour. Cependant, une partie de la communauté océanographique aimerait avoir ces données en NetCDF (ex : colocalisation avec d'autres données, etc.).

Enfin, il existe des outils :

- de SeaDataNet comme NEMO qui permet de convertir les données en BioODV, OCTOPUS qui permet de contrôler le fichier BioODV , MiKado qui permet de générer les metadonnées,
- du PNDB comme METASHARK qui permet de faire la saisie des métadonnées pour générer un fichier XML en EML afin de faire un data package type DarwinCore Archive,
- du GBIF avec IPT qui permet de créer du DarwinCore Archive.

5. Autres points & préparation du prochain Atelier Technique

Les sujets qui semblent important à traiter dans les prochains ateliers sont les suivants :

- Un atelier sur les référentiels taxonomiques pour les données bio en lien avec le GT langage commun du SAR/SIMM. Les procaryotes (bactérie, etc.) sont sous-représentés dans les référentiels taxonomiques. (Inviter des personnes associées au base et référentiels génomiques).
- Prise en compte des données générées par les laboratoires de génomiques (identification de groupes fonctionnels ou d'espèce via des données de séquences). Quels sont les standards permettant de consigner ces données ? Quel est le référentiel taxonomique à utiliser (barcoding) ? Pour le meta-barcoding, comment prendre en compte l'amplification de l'ADN par la PCR qui fait entrer un biais par rapport aux espèces qui ont plus d'ADN que d'autres ?
- Avec la meta-génomique il n'y a plus d'identification d'espèce ou de groupe fonctionnel, mais des occurrences de séquence ou de gène. Comment intégrer ce nouveau type de donnée ? Quels sont les entrepôts de référence ? Est-ce que le vocabulaire standardisé, qui est utilisé pour la méta-génomique, est « transposable » et/ou adaptable (protocoles, méthodes, etc.) dans nos domaines ? Peut-il être vraiment intégré dans nos référentiels et nos vocabulaires ?

Les besoins en formation à destination de la communauté bio :

- Formation EMODnet bio en français (formation aux outils IPT, DarwinCore) -> (le GBIF peut faire ce type de formation),
- Formation au format BioODV et au vocabulaire BODC (SDN) -> MF peut faire cette formation

(Voir dans les autres pôles (THEIA, AERIS) s'il y a un intérêt à ces formations et s'il ne vaut pas mieux les faire au niveau Data Terra - faire un sondage).

Un CES imagerie (divisé en sous catégories d'imagerie) est à prévoir pour répondre à la question : Où déposer ce type de données ?

6. Relevé synthétique des conclusions et actions

La gestion des données de biologie marine nécessite une liste des référentiels suivants :

- Référentiel taxonomique de type WoRMS ;
- Référentiel taxonomique ad hoc, avec des groupes morphologiques de type SANDRE ;
- Référentiel sur les instruments ;
- Référentiel sur le niveau d'expertise ;
- Référentiel sur les méthodes d'observations, d'échantillonnages et procédures ;
- Référentiel sur les limites de détection, de saturation de l'appareil, des limites de détection ;
- Référentiel de calibration des instruments (n° de série, …) ;
- Référentiel sur les habitats ;
- Référentiels sur les traits de vie ;
- Référentiels des paramètres (taille, poids, …).

Recommandations :

- Garder la traçabilité et la gestion de version dans les bases de données.
- Deux formats de mise à disposition des données : BioODV et DarwinCore Archive
- Questionnement préalable à la « gestion des images ou des vidéos » :
    - Quels types d'image va-t-on recevoir (Image brute, Imagette traitée, …) ? Que faut-il archiver et dans quels formats ?
    - Comment traiter ces images (ou vidéo) ? En amont d'ODATIS ou est-ce que les CDS vont y contribuer ? A quels formats ? Avec quel compromis avec des outils de compression (algorithme sans pertes d'informations dans l'image) si le volume est trop important ?
    - Que va-t-on devoir conserver après le traitement ? comptage ? espèce ? mosaïque d'image ?
    - Est-ce qu'en plus des images ou vidéos, il y a d'autres choses à conserver ((par ex couverte géographique de type SIG pour des mosaïques d'image)?
- Producteurs de données, il est nécessaire que les CDS proposent (idéalement) deux mécanismes pour rendre disponible la donnée :
    - Un mécanisme par téléchargement de fichiers ;
    - Un mécanisme dynamique par service de requêtage de la base de données.

Action :

- Prévoir une réunion avec l'ensemble des scientifiques impliqués dans les laboratoires et Data Terra, pour l'archivage.
- Ecotaxa :
  - Constitution d'un groupe **de FA, GG, JOI, (et FC, Jean-Baptiste Romagnan pour l'instance d'EcoTaxa à l'IFREMER)**
  - **Recensement les besoins actuels (stockage, archive, capacité de calcul)**, ainsi que la dynamique de son évolution et de le transmettre à GM et JS pour interagir ensemble avec les infrastructures (IFB)
  - Solution à valider/discuter avec MH : mise en place d'un 2eme flux de données (Flux « national ») pour les données Ecotaxa. (Flux distinct du flux actuel Ecotaxa > Emodnet Bio). Ce flux national, gestion par PELAGOS.
- Imagerie sous-marine :
  - **i**nventaire (stockage, archive, capacité de calcul) devra également être effectué pour l'imagerie sous-marine **par CB.**
  - Retour sur le projet AMII (action CB): Etude des différents workflows (non BIIGLE)
  - Mise en place d'un **CES imagerie (divisé en sous catégories d'imagerie)** pour répondre à la question : Où déposer ce type de données ?
  -
- Formation :
  - Voir dans les autres pôles (THEIA, AERIS) s'il y a un intérêt à ces formations (cfr ci-dessous) et s'il ne vaut pas mieux les faire au niveau Data Terra.
  - Faire un sondage Data Terra ?

Besoins en formation à destination de la communauté bio :

- Formation EMODnet bio en français (formation aux outils IPT, DarwinCore) -> (le GBIF peut faire ce type de formation),
- Formation au format BioODV et au vocabulaire BODC (SDN) -> MF peut faire cette formation