Compte rendu de l'atelier technique des 14 et 15 juin 2021 - Biodiversité : métadonnées FAIR & TP VRE (Galaxy-E & MetaShARK)

CR Atelier Technique juin 2021

Titre court
CR Atelier Technique juin 2021 – Biodiversité : métadonnées FAIR & TP VRE
Titre long
Compte rendu de l'atelier technique des 14 et 15 juin 2021 – Biodiversité : métadonnées FAIR & TP VRE (Galaxy-E & MetaShARK)
Auteur
Joël Sudre, CNRS – UMS CPST ; Cécile Nys, OceanScope/Ifremer ; Elie Arnaud, MNHN – PNDB.

Dissémination	Copyright
Publique	© Pôle ODATIS

Historique

Version	Auteurs	Date	Commentaires
0.1	Joël Sudre	19 juillet 2021	Version initiale
0.2	Cécile Nys	27 juillet 2021	Relecture et corrections
0.3	Elie Arnaud	2 août 2021	Relecture et corrections
0.4	Cécile Nys	18 août 2021	Révision des corrections
1.0	Cécile Nys & Joël Sudre	23 août 2021	Version pour diffusion en ligne

Table des matières

Table des illustrations

1. Accueil et tour de table des participants

Tableau 1. Liste des participants à l'atelier ODATIS #11 des 14 et 15 juin 2021

Liste des participants à l'Atelier Technique #9	
ARNAUD Elie (MNHM, PNDB) – EA	JOSSE Marie (MNHM – PNDB) – MJ
BRESSAN Hélène (BRGM) – HB	MAUDIRE Gilbert (IFREMER) – GM
BRIAND Dominique (IFREMER) – DB	MENDES Fabrice (OASU) – FM
CHAMPENNOIS Victor (CNRS - LOCEAN) – VC	MERCIER Caroline (ODATIS) – CM
DRU Philippe (IMEV) – PD	NYS Cécile (ODATIS) – CN
HARSCOAT Valérie (IFREMER) – VH	PIERKOT Christelle (CNRS – IRDT/CPST) – CP
HENNEBAUT Brendan (IFREMER) – BH	ROYAUX Coline (MNHM / PNDB) – CR
HOEBEKE Mark (CNRS - SRB) – MH	SANANIKONE Julien (MNHM / PNDB) – JSK
KHVOROSTYANOV Dimitry – DK	SUDRE Joël (CNRS - IRDT/CPST) – JS
LE BRAS Yvan (MNHM / PNDB) – YLB	VERNET Marine (IRDT/CPST) – MV
LIBES Maurice (CNRS – OSU PYTHEAS) – ML	

2. Introduction – (Joël Sudre)

JS présente l'ordre du jour (*voir : Agenda et accès aux présentations*[1]). Cet atelier est issu de la collaboration du Pôle ODATIS avec le Pôle National de Données de Biodiversité[2] (PNDB). Suite à l'atelier de mars 2021, sur l'environnement virtuel de recherche (VRE) pour la physique et les métadonnées FAIR, il a été décidé, avec la collaboration du PNDB, et en la personne d'Yvan LE BRAS d'organiser un atelier similaire, mais axé sur les outils et services utilisés dans la communauté de la biodiversité. Lors de cet AT il est prévu également d'échanger sur leurs usages en termes de principe FAIR afin de découvrir et de créer des passerelles techniques entre ces deux communautés et leurs bonnes pratiques. L'organisation des deux demi-journées en visioconférence est donc une collaboration avec Yvan LE BRAS et son équipe de l'Unité Mixte de Service Patrimoine naturel (UMS PatriNat – en particulier avec Elie ARNAUD et Coline ROYAUX). La première demi-journée est dédiée aux outils et services du PNDB, d'une introduction à Galaxy-E et à MetaShARK. La seconde à une prise en main sous forme d'atelier pratique de l'environnement virtuel de développement Galaxy-E et des outils et « workflows » dédiés aux principes FAIR au PNDB.

[1] https://www.odatis-ocean.fr/activites/ateliers-techniques/atelier-technique-juin-2021
[2] https://www.patrinat.fr/fr

3. Outils et services du PNDB,

3.1. Introduction à Galaxy-E (Yvan LE BRAS)

YLB présente l'Infrastructure de Recherche PNDB (voir 202106_ODATISAtelier_YLeBras_PNDB_VRE_MTD_Galaxy-E.pdf[3]). Ce pôle est dédié aux données et aux connaissances relatives à la biodiversité (état, dynamiques, interactions, rétroactions, etc.). Cette IR est constituée de 9 organismes, 10 universités, la Fondation pour la Recherche en Biodiversité (FRB) et son portage est effectué par le Muséum National d'Histoire Naturel (MNHM). Ses objectifs sont de développer, offrir et mettre en synergie :

- Outils et services visant la description, la mise à disposition, la validation, l'analyse et la réutilisation des données de biodiversité ;
- Fonder un cadre scientifique intégratif : temps long, tous milieux, de la molécule aux anthroposystèmes, pressions anthropiques, etc.) ;
- Compléter le dispositif Système Terre – Environnement en lien avec le Système d'Information sur la Biodiversité (SIB) national porté par l'Office Français de la Biodiversité (OFB)/Ministère de la Transition Ecologique (MTE).

Sa stratégie est de s'appuyer sur l'existant et de favoriser :

- Pérennité, interopérabilité, ouverture, réutilisation ;
- Croisement de données biodiv / autres (climat, env,…), sans se substituer aux porteurs de bases de données ;
- Développer la description des jeux de données (métadonnées, qualité des données) ;
- Faire monter les communautés en compétences ;
- S'inscrire dans le paysage complexe des infrastructures au niveau national, européen et international.

Le PNDB à une offre de services qui comprend :

- L'accès aux jeux de (méta)données avec des services et des outils associés ;
- Une aide à la bancarisation ;
- De l'animation des communautés autour de la science ouverte ;
- Une stimulation des interactions producteurs/utilisateurs.

Le positionnement à l'échelle nationale et international du PNDB ayant déjà fait l'objet de présentations, le lecteur est invité à relire le CR de l'atelier de Mai 2021[4].

[3] https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202106/202106_ODATISAtelier_YLeBras_PNDB_VRE_MTD_Galaxy-E.pdf
[4] https://www.odatis-ocean.fr/activites/ateliers-techniques/atelier-technique-mars-2021

Ses services et outils sont du type :

- Structuration/standardisation de la donnée, :
  - Avec inférence de métadonnées,
  - Métadonnées très fines/détaillées ;
- Accès direct à des données ouvertes brutes ;
- Analyse/couplage de données hétérogènes sur infrastructure cloud / HPC / distribuée via plateforme ET outils open sources et à fort degré « FAIR ».

La gestion des données repose sur :

- Des conseils sur les entrepôts de données les plus pertinents ;
- Une approche « tout est métadonnée » pour :
  - le maDMP (projet Equipex+),
  - les data papers (projet FNSO 2019).

Pour le PNDB, toutes les métadonnées et données sont ouvertes (uniquement licence ouverte Etalab v2 compatible CC-BY 4.0), tous les scripts/outils/plateformes sont sous licence ouverte (GNU GPL/MIT/CECILL, …). Ceci représente 160 codes ouverts pour les outils « Galaxy Ecology » (PNDB) et plus de 8000 accessibles via l'« app store » Galaxy.

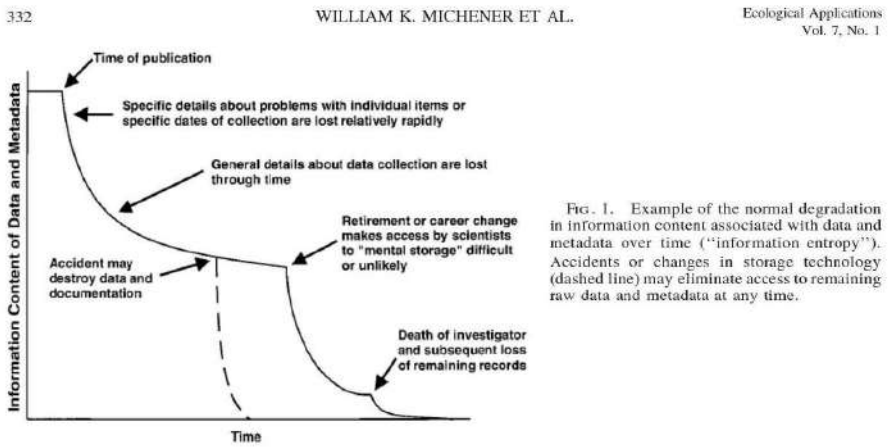

Figure 1. Dégradation de l'information sur les données au fil du temps (Michener et al., 2006, Doi : 10.1016/j.ecoinf.2005.08.004)

Afin d'éviter la dégradation des informations sur les (méta)données associées à leurs dégradations dans le temps (voir fig. 1), le PNDB a mis en place un outil : MetaShARK permettant de récupérer au sein du pôle ses (méta)données.

Le format de mise à disposition est l'Ecological Metadata Language (EML) qui permet d'ingérer des (méta)données soit en mode single, soit en mode batch. Cet outil permet aussi d'enrichir sémantiquement les (méta)données et de les unifier (voir la présentation : planches 8 et 9). Il est à noter que cet outil permet la standardisation par la métadonnée et non par la donnée ce qui est fait plus usuellement à ODATIS. Pour les (méta)données de biodiversité cet outil permet de :

- Faciliter l'accès aux données brutes par tous types d'acteurs : Les participants aux atlas de la biodiversité communale n'ont pas accès à la donnée, ou qu'une partie filtrée après passage par SI d'association puis SI politique publique.
- Faciliter le calcul d'indicateurs de Biodiversité : On ne peut pas se contenter de données d'occurrence ou de présence seule pour les calculs d'indicateurs à l'échelle des communautés ou écosystèmes.
- D'éviter le problème du fait qu'un standard de données peut obliger à contraindre / modifier le sens d'une variable primaire : Pour certaines données de recherche notamment, il faut rentrer dans le cadre même si la correspondance n'est pas parfaite.

En ce qui concerne le volet analyse, le PNDB prône l'utilisation de Github, Conda, la conteneurisation, la mise en cloud via l'utilisation de l'environnement de recherche virtuel : Galaxy-E (aux moyens de machines virtuelles locales ou sur le cloud). Cet environnement permet d'utiliser un environnement Jupyter/Pangeo Python (comme pour ODATIS), mais aussi des applications R shiny, Rstudio, etc. Son offre, plus large que l'environnement pour la physique est adapté à sa communauté mais aussi à la communauté plus physicienne voulant interagir avec la communauté de biodiversité (biodiversité marine en particulier). Elle permet aussi le passage de codes Python dans l'environnement Galaxy en créant des applications intégrées à ce système (sans manipulation de code de la part de l'utilisateur, une fois la transformation de Python en module Galaxy effectuée).

Les planches de 15 à la fin, permettent d'avoir la visio intégrée de cette VRE ainsi que les « workflows » d'ingestion des (méta)données dans l'univers Galaxy-E.

3.2. Introduction à MetaShARK : métadonnées FAIR (Elie ARNAUD & Yvan LE BRAS)

YLB et EA présentent plus en détail l'un des outils centraux du PNDB : l'outil MetaShARK (voir 202106_ODATISAtelier_EArnaud_FAIR_metadata.pdf[5]). L'outil MetaSHARK a un plan de développement sur deux ans qui débute actuellement afin de prendre en compte les remarques et les ajouts, simplifications, etc. qui lui permettront d'être plus utile et fonctionnel pour les utilisateurs des différentes communautés. Cette présentation est axée sur les aspects techniques de MetaShARK qui est construit comme un écosystème comprenant les briques de base de la figure 2 :

- Au premier niveau : EML maintenu par le NCEAS, Langage R et un outil en langage web (html, javascript) ;
- Au second niveau : EML Assembly Line maintenu par EDI (Environmental Data Initiative) qui est en étroite collaboration avec le PNDB) et Shiny (package R qui permet à partir du langage R de programmer des interfaces graphiques et des outils) ;
- Au dernier niveau, il y a l'outil MetaShARK construit sur l'ensemble de ces briques de bases.

[5] https://www.odatis-ocean.fr/fileadmin/documents/activites/ateliers/atelier_202106/202106_ODATISAtelier_EArnaud_FAIR_metadata.pdf

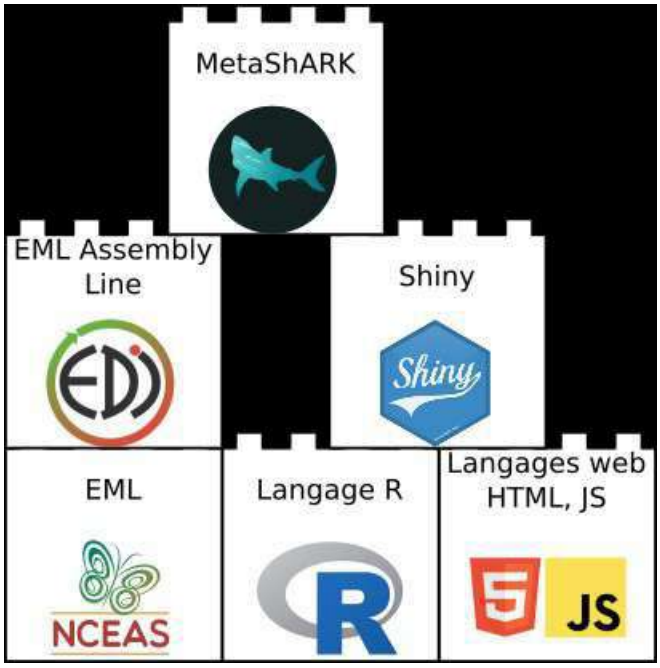

Figure 2. Empilement des briques de bases pour la création de l'outil MetaShARK

L'EML Assembly Line permet de générer des métadonnées par script à partir de données, de faire du stockage intermédiaire en tableaux (avec « templates » associés) et au final de produire un assemblage conforme aux spécifications de l'EML (en XML). L'édition des « templates » est effectuée manuellement et permet de lier à ces fichiers, les fichiers de données. L'objectif de MetaShARK est d'assister l'utilisateur pour aller chercher les fichiers de donnée avec un affichage « tout en un » de son data package (data package comprenant données, métadonnées et tous les fichiers intermédiaires).

Le fichier EML qui a été créé, validé et préparé par l'utilisation de MetaShARK pour son échange est ensuite envoyé au PNDB via son portail web. Cette dernière étape de mise en ligne n'est pas encore mise dans MetaShARK mais cela est prévu dans son plan de développement.

A la suite de cette présentation technique de MetaShARK, EA nous invite à faire une démonstration en directe de l'outil MetaShARK. L'outil MetaShARK est accessible sur le lien suivant : https://metashark.test.pndb.fr/[6].

[6] https://metashark.test.pndb.fr/

4. Atelier Pratique pour la prise en main de la VRE GALAXY-E (Coline ROYAUX & Yvan LE BRAS)

YLB présente le projet GALAXY Europe qui est une plateforme web pour le partage et le traitement des données de recherche. Elle permet un accès facilité à l'analyse Cloud et au calcul haute performance (HPC) par l'interfaçage de divers langages informatiques. Elle est basée sur les principes fondamentaux d'accessibilité, de reproductivité de transparence et de peer review. Cet environnement est né de la nécessité pour la biodiversité génomique d'avoir un moyen pour gérer les fichiers extrêmement volumineux ([1 – 100] Go voire 1 To) produits par le séquençage génomique d'échantillon. Il existe 3 instances internationales de Galaxy : Une américaine, une australienne et une européenne.

Un article vraiment fondateur pour cet environnement qui repose sur CONDA (BIOCONDA et biocontainers) est le suivant :

> Practical computational reproductibility in the life sciences (B Grünning et al., Doi : https://doi.org/10.1016/j.cels.2018.03.014[8]).

Un aperçu du "backend" de Galaxy est présenté à la figure 3.

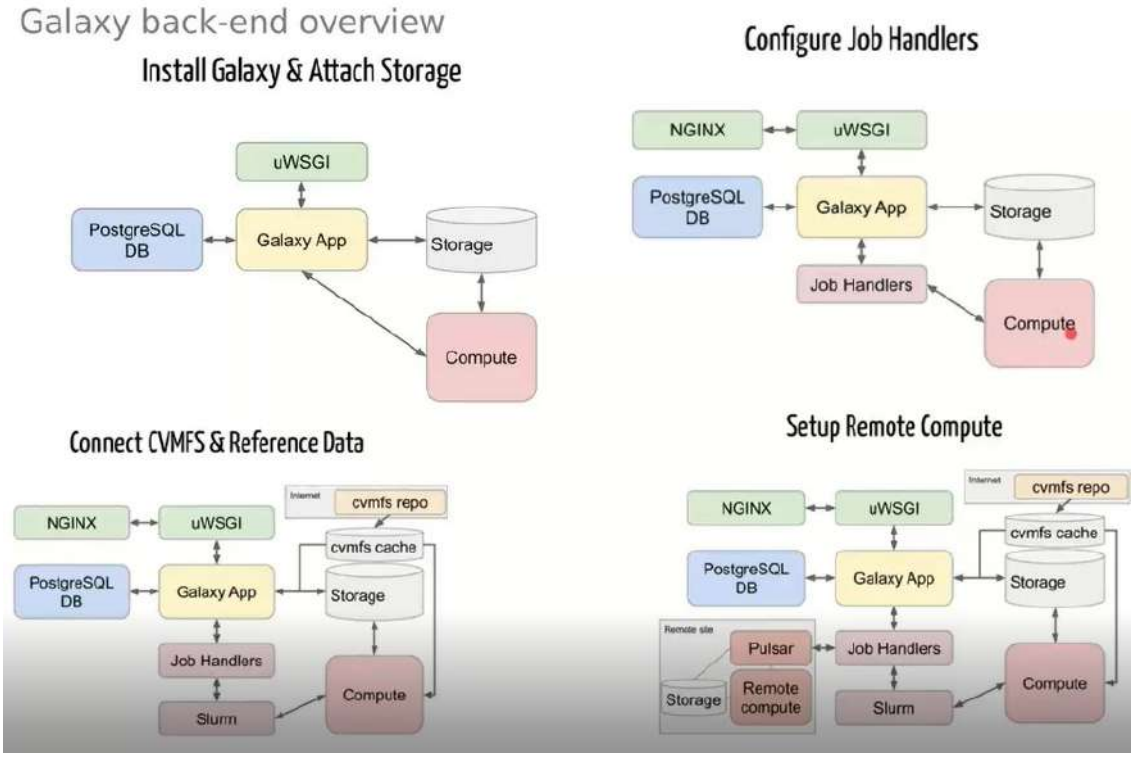

Figure 3. Aperçu du back-end de Galaxy-E

[7] https://doi.org/10.1016/j.cels.2018.03.014
[8] https://doi.org/10.1016/j.cels.2018.03.014

L'environnement GalaXY-E est aussi un lanceur d'outils en application interactive comme (JupyterLab avec un Jupyter Notebook, l'environnement de développement en R, R Shiny, l'outil de visualisation des NetCDF Panoply, etc.). Galaxy permet aussi la mise en place et l'utilisation d'un moteur de « workflows », avec la possibilité de lancer soit en asynchrone ou en synchrone les « workflows » et ceux de façon très simple pour l'utilisateur. Il est aussi possible de transformer un script synchrone en scripts asynchrone.

Suite à cette présentation très succincte de l'environnement GALAXY-E, CR présente un « worrflow » en démonstration au moyen d'un tutoriel utilisant Galaxy-E afin de calculer des métriques de biodiversité et en utilisant un enchaînement d'outils permettant de faire ces analyses. Cette démonstration permet de prendre en main divers outils disponibles dans Galaxy-E et d'apprendre à les utiliser.

5. Conclusion

En guise de conclusion, ce premier atelier technique ODATIS avec une forte collaboration avec le PNDB a permis de prendre en main les outils MetaShARK et Galaxy-E qui sont des outils collaboratifs très utilisés dans la communauté de la biodiversité et qui pourraient être utilisés aussi par la communauté de la biodiversité marine. Les échanges fructueux de ces 2 demi-journées, on encore permis de révéler qu'un rapprochement entre l'IR DATA TERRA et ses pôles (ODATIS dans le cas présent) et l'IR PNDB serait extrêmement profitable pour les différentes communautés scientifiques qu'ils représentent et permettrait de décloisonner les données environnementales et les données de biodiversité.