De la conception d'un système d'observation à large échelle au déploiement et à l'exploitation de son système d'information : application à l'observation des habitats coralligènes et à la colonisation de récifs artificiels (ARMS)

Romain David

Thèse de Doctorat d'Aix-Marseille Université
Spécialité Ecologie Marine
Soutenue par Romain DAVID le Vendredi 6 Juillet 2018

De la conception d'un système d'observation à large échelle au déploiement et à l'exploitation de son système d'information
Application à l'observation des habitats coralligènes et à la colonisation de récifs artificiels (ARMS)

D. Ienco, IRSTEA, Montpellier (rapporteur)
T. Saucède, Biogéosciences, univ. Bourgogne, Dijon (rapporteur)
L. Berti-Equille, AMU, Marseille (examinateur)
T. Tatoni, AMU, I.M.B.E., Marseille (examinateur),
R. Vigne-Lebbe, M.N.H.N.-U.P.M.C., Paris (examinateur)
J.-P. Féral, C.N.R.S., I.M.B.E., Marseille (directeur)

Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse Jean-Pierre Féral, de m'avoir proposé ce sujet et m'avoir épaulé, y compris financièrement (sur ses contrats ou en obtenant des crédits dédiés de la part de l'INEE), pendant ces années riches en expériences et en formations. Il m'a aussi fait confiance en me proposant la direction d'un "work package" dans le cadre du programme européen CIGESMED dont il était responsable.
J'ai pu, grâce à ce soutien et cet engagement, faire mes premiers pas dans le monde de la recherche, et petit à petit comprendre l'organisation, les méthodes de travail, les tenants et les aboutissants de la carrière d'un chercheur en écologie marine.
Je voudrais aussi remercier tous les membres de mon comité de thèse qui m'ont conseillé et proposé de nouvelles orientations afin d'enrichir mon travail au cours de ces dernières années, ces conseils et idées m'ont été précieux. Merci également à tous les membres du jury d'avoir accepté de donner de leur temps et de leur énergie pour évaluer ce travail.

Ma gratitude se porte aussi tout particulièrement vers mes collègues et/ou étudiants de la Station (anciens et actuels) avec lesquels j'ai eu le plus d'interactions, en particulier Anne Chenuil, responsable du programme DEVOTES, avec qui j'ai eu le plaisir de mettre en place l'expérience reposant sur les ARMS, Pascal Mirleau chercheur et plongeur, Abigail Cahill (post-doctorante), Emilie Egea (Ingénieur de recherche), Giulia Gatti (post-doctorante et plongeuse), Aurélien De Jode (Doctorant), Alexandra Weber (doctorante), Marjorie Selva (technicienne) et Sandrine Chenesseau (Technicienne et plongeuse).
Un grand merci à mes collègues de l'axe thématique "gestion de la biodiversité et des habitats naturels" avec lesquels j'ai collaboré sur de nombreux aspects de la DCSMM ou de la gestion d'espaces naturels à savoir Xavier Fizzala, Florent Renaud, Sandrine Serre, Pauline Vouriot chargés de mission DCSMM pour le descripteur 4 et Laure Thierry de Ville d'Avray (doctorante). Un coup de chapeau spécial aux étudiants qui ont réalisé leurs stages dans le cadre des programmes CIGESMED et DEVOTES et qui se sont investis sans compter pour assurer la récolte et l'analyse de données et d'échantillons de terrain (analyse photo notamment) et/ou les différents traitements de matériel prélevé. Parmi eux, j'ai eu le plaisir d'encadrer dans le cadre de leur stage Dorian Guillemain, Laure Thierry De Ville D'Avray, Walid El Guerrabi, Jacky Dubar, Ohane Legendarme. J'ai aussi eu le plaisir de travailler avec d'autres étudiants comme Zinovia Erga, Sophie Dubois (et pardon pour ceux que j'oublie). Nous avons tous ensemble formé une équipe de choc : sans vous, ce travail ne serait pas ce qu'il est.

J’ai rencontré des personnes remarquables, professionnelles et humaines à la fois à l’OSU Pythéas, à la station marine d’Endoume et à l’IMBE. Ceux-ci, dans des contextes bien différents, ont été à mon égard d’une bienveillance souvent rassurante. Pour cette attention, je remercie tout particulièrement Cyrille Blanpain, Carole Borchellini, Alrick Dias, Dominique Estival, Giulia Gatti, Christian Marschal, Emmanuelle Renard, Léïta Tschanz, Marc Verlaque, Lilita Vong. Un remerciement tout particulier à Joëlle Massei, notre gestionnaire pour m’avoir aidé avec bienveillance et patience dans la gestion administrative des programmes et actions que j’ai pu mettre en œuvre ces dernières années.
J’ai aussi des remerciements spéciaux à faire, notamment à Bernard De Ligonnès, pilote du pointu *l’Armandia* et d’une aide et d’une disponibilité à toute épreuve. Bonne retraite à toi, je garderai un souvenir impérissable de tous ces moments passés sur l’eau ou ailleurs avec toi : sans aucun doute parmi mes meilleurs souvenirs de ces années.

Cette thèse a aussi pu avancer grâce à des financements institutionnels. La construction du premier prototype pour le consortium IndexMed a été financée par l’INEE qui a soutenu le projet "CHARLIEE" (CHAnger de Regard En Liant dans IndexMed l’Environnement et les Étoiles - PEPS Blanc) en 2014, puis par le CNRS *via* le défi "VIGI-GEEK (VIsualisation of Graph In transdisciplinary Global. Ecology, Economy and Sociology data-Kernel)" en 2015, enfin par un financement alloué directement par la direction pour l’interdisciplinarité du C.N.R.S. en 2016. En 2017 et 2018, c’est l’action GRAMINÉES qui a été labellisée et soutenue par le GDR MaDICS. Ce soutien a permis d’obtenir un cofinancement de la F.R.B. Les données utilisées dans le cadre de cette thèse ont été obtenues dans le cadre du projet européen CIGESMED <www.cigesmed.eu> (conventions ANR n° 12-SEAS-0001-01, 02 et 03 pour la France, GSRT-12SEAS-12-C2 pour la Grèce, projet TUBITAK No : 112Y393 pour la Turquie). En complément, en 2016, 2017 et 2018, le Labex DRIIHM a soutenu un programme propre aux OHM, et cofinancé les séminaires et ateliers IndexMEED. Merci aussi pour leur soutien aux séminaires au labex OT-Med, à l’OHM Littoral Méditerranéen, à l’OSU Pytheas, à la Fédération de Recherche Eccorev et à la Fondation pour la Recherche sur la Biodiversité.
Je souhaite adresser mes plus vifs remerciements aux partenaires des différents programmes de recherche CIGESMED* et DEVOTES**, avec lesquels j’ai pu interagir. Ces programmes ont été un cadre riche et motivant pour effectuer ce travail.
Mes remerciements aux contributeurs les plus actifs lors des séminaires organisés dans le cadre de CIGESMED, français, grecs et turcs qui, lors de nos débats et lors de séminaires intensifs ont contribué à améliorer le protocole et faire avancer les travaux de ce programme.

Le soutien de France Grilles et notamment de Vincent Breton, Geneviève Romier et Yannick Legré a permis d'expérimenter et d'utiliser les ressources informatiques sur l'Infrastructure Grid Nationale et ont financé une partie des formations et séminaires sur les outils de calculs parallélisés auxquels j'ai pu assister. Dans ce cadre, je souhaite dire ma reconnaissance à Gergely Sipos, Jan Bot et Roberta Piscitelli pour le soutien utile fourni lors de l'atelier "design your e-infrastructure" EGI <www.egi.eu>.

Je souhaite par ailleurs remercier l'équipe de la F.R.B. et particulièrement Aurélie Delavaud, Anna Cohen Nabeiro, Robin Goffaux et leur direction pour avoir permis par leurs appuis financier, technique, administratif et scientifique la réalisation des ateliers et séminaires de IndexMEED ces deux dernières années et ainsi avoir tant favorisé son développement. Merci aussi à tous les organisateurs, membres du conseil scientifique d'IndexMEED, intervenants et participants aux différents séminaires, et aux ateliers cartographie de compétences, curation puis visualisation de données (IndexMed puis IndexMEED) : sans votre investissement et votre confiance, nous n'aurions pas réussi à atteindre les ambitieux objectifs que nous nous étions fixés (notamment Anne-Sophie Archambeau, Fanny Arnaud, David Auber, Nicolas Bailly, Loup Bernard, Cyrille Blanpain, Vincent Breton, Denis Couvet, Alrick Dias, Sophie Gachet, Robin Goffaux, Karina Gibert, Manuel Herrera, Dino Ienco, Romain Julliard, Julien Lecubin, Yannick Legre, Michelle Leydet, Grégoire Loïs, Victor Méndez Muñoz, Jean-Charles Meunier, Isabelle Mougenot, Sophie Pamerlon, Jean-Claude Raynal, Geneviève Romier, Dad Roux-Michollet, Alison Specht, Christian Surace, Thierry Tatoni, et tant d'autres).

Merci aussi à l'équipe technique du S.I.P. (Service Informatique de l'O.S.U. Pytheas) notamment Cyrille Blanpain, Julien Lecubin et Alrick Dias sans qui le développement du prototype n'aurait pas pu aboutir et sur qui j'ai aussi pu compter pour m'aider à organiser puis animer les différents ateliers interdisciplinaires.

Et enfin, ce qui compte le plus pour moi, ma famille : merci à mes grands-parents, cousins et cousines d'avoir constitué l'univers dans lequel j'ai pu me construire et puiser les exemples dont j'ai eu besoin lors de ce travail. Merci à ma famille, particulièrement mes parents, et ma fratrie (Sophie, Emilie, Julie, Benjamin, Marie et Quentin) ; J'ai bien conscience pendant ces six années de vous avoir délaissés ; vous êtes mon véritable chez moi, et j'en avais déjà largement conscience, s'il ne devait me rester plus qu'un seul dernier soutien, ce serait bien de vous qu'il proviendrait. Plus particulièrement, Merci à David, mon beau-frère, Marie et Emilie, mes soeur, Sophie ma cousine et Dr Sophie Gachet pour leurs relectures attentives et leurs conseils avisés. Merci à mon frère Benjamin pour son soutien moral et financier, et pour la droiture, le courage et la volonté que tes aventures t'ont demandés, ils ont été pour moi un exemple. Pour conclure, merci à Doriane, compagne de vie de ces 12 dernières

années, pour tes encouragements, ta détermination et ta volonté dans ta nouvelle profession qui ont été pour moi un moteur pour terminer ce travail.

Sommaire

Résumé

Dans le domaine de l'environnement marin, la fréquence des campagnes de collecte de données recueillies lors de programmes de recherche, de suivis environnementaux [obligations européennes entre autres], d'études d'impact, (missions de terrain, capteurs optiques ou radar, suivi de la qualité des eaux, recensement automatique ou semi-automatique des taxons, etc.) conduisent à l'accumulation d'un volume considérable de données. Des protocoles d'observation sont constamment développés dans de nombreux cadres, et produisent des données très hétérogènes centrées sur l'utilisation souvent spécifique à un métier qu'envisage leur producteur. Du fait de leur hétérogénéité, celles-ci sont difficiles à agréger pour avoir une vue d'ensemble (on parle parfois "d'empilement de bases de données"). De plus, l'accès aux données n'est pas organisé et se révèle souvent difficile, voire même impossible. Cet accès pour de multiples utilisateurs et l'agrégation de données à large échelle[1] sont pourtant incontournables pour mieux cerner les enjeux de protection de la biodiversité et des ressources marines, et anticiper leur détérioration irréversible.

Afin de mieux protéger la biodiversité marine et surtout afin que les enjeux de conservation et de préservation des ressources soient mieux pris en compte par les politiques publiques, il est nécessaire de renforcer la cohérence des systèmes d'observation et d'acquisition de nouvelles connaissances, et d'organiser l'accès aux données et résultats de recherche pour tous les utilisateurs potentiels. Faire des propositions pour améliorer la cohérence entre systèmes d'observation et systèmes d'information est l'objectif cadre de ce travail.

La rationalisation des moyens à investir pour préserver le bon état environnemental n'est possible qu'en réalisant un état des lieux des connaissances produites, des compétences disponibles et des verrous à lever pour la mise en place efficace de systèmes d'observation à large échelle (les blocages en terme de socle de connaissance autant que les blocages méthodologiques, sociologiques, scientifiques et fonctionnels décrits dans les systèmes aujourd'hui mis en œuvre). Nécessairement, les systèmes d'observation doivent être couplés à une architecture de systèmes d'information. Cette architecture doit permettre d'organiser la création et l'accès à la connaissance, de faciliter sa conservation, d'harmoniser les méthodes et systèmes de gestion et d'analyse de données.

[1] Par systèmes d'observation à large échelle, on entend un système d'observation pertinent pour effectuer des suivis à une échelle pan régionale, pan méditerranéenne, précurseurs de systèmes à plus large échelle.

La construction d'une démarche et d'un projet en réseau financé sur le long terme constitue un préalable à la pérennisation de systèmes de suivi à grande échelle. Les réseaux de suivi ensuite générés deviennent alors multi-usagers et devront permettre cette rationalisation des moyens investis dans la production de nouvelles connaissances. Ils doivent produire suffisamment de descripteurs fiables pour élaborer une indication performante, qui est-elle même nécessaire aux démarches de reportage (D.C.S.M.M., D.H.F.F., D.C.E., Tableau de bord des mers françaises, O.N.B., etc.).

Cette thèse a pour objectif de i) proposer des méthodes, protocoles et recommandations pour construire et/ou soutenir la mise en place de réseaux de suivis utiles et pérennes de la biodiversité, à l'échelle d'une zone biogéographique ou sur le plan international, s'appuyant sur les acteurs locaux (dispositifs allant des suivis de gènes aux suivis d'espèces et d'habitats), ii) favoriser les utilisations multiples et novatrices des données tout en préservant les droits de l'auteur/inventeur des dispositifs, augmenter et améliorer les différents accès aux données (brutes, traitées et de synthèse).

Deux dispositifs « cas d'étude » ont été choisis pour ce travail : les habitats coralligènes à l'échelle de la Méditerranée et la colonisation de récifs artificiels (ARMS) dans différentes mers régionales en focalisant :

- sur la construction de réseaux de suivi et d'observation pérennes et utiles pour différents types d'usages,
- sur le partage efficace des connaissances à long terme avec ses différents utilisateurs potentiels (scientifiques, gestionnaires, élus, amateurs, grand public…) et sur l'interopération des systèmes d'informations,
- sur les méthodes, outils et interfaces d'analyse de données exploitant les nouvelles avancées dans le domaine du *Big Data*, de la gestion des données hétérogènes et de leur analyse sous forme de graphes.

Les habitats côtiers étant la cible principale de ce travail, le test des différents protocoles montre qu'une expérimentation à large échelle doit absolument être décrite très explicitement dans des termes standardisés au-delà même du champ disciplinaire de l'écologie marine (si possible en les organisant en micro thésaurus à "aligner" avec les standards en cours de développement). Cette expérimentation doit se baser sur des méthodes de mesure les plus simples possibles à mettre en œuvre. Les tests effectués par différents opérateurs ont montré l'importance d'une formation, puis d'une mise à l'épreuve itérative sous la forme de confrontation des résultats sur une même observation, que ce soit sur le terrain ou pendant les analyses *ex situ*. Les temps d'apprentissages sont d'ailleurs à adapter aux types d'objets et/ou aux méthodes concernés (taxon, habitat, détermination, comptage, mesure...).

Le travail sur l'architecture des systèmes d'information et les débats concernant les cycles de vie de la donnée ont permis de mettre en évidence l'inefficacité d'un système centralisé,

et l'inévitabilité d'un système de gestion modulaire, orienté "métier[2]" et décentralisé. Il en découle que la non gestion actuelle des autorités est un verrou pour la traçabilité de la donnée et la reconnaissance des producteurs.

Grâce à l'organisation de l'accès aux données sous forme de flux paramétrables et ouverts, il a été proposé dans le cadre de cette thèse un mécanisme de couplage de données de différentes origines (des observations de terrain et des données décrivant les contextes) reposant sur la requalification des facteurs descriptifs hétérogènes en facteurs équivalents et simplifiés dont le choix repose sur un arbitrage collaboratif entre spécialistes. En se basant sur un prototype, une nouvelle méthode d'analyse de données environnementales et l'utilisation de méthodes de fouille de données basé sur les graphes ont été mises en démonstration et développées pour devenir générique. Des exemples de visualisation des données et différents types de démonstrations possibles partant des données ont été construits grâce à l'organisation d'ateliers de curation et de visualisation de données sous forme de graphes.

En conclusion, les premiers tests fonctionnels ont montré que l'information produite doit pouvoir être contrôlée en temps réel et de manière itérative, et que les processus de curation de la donnée doivent nécessairement être mis en place en même temps que la conception des procédés d'observation. En complément, les définitions de standards et l'accessibilité des données de contexte[3] nécessitent un travail collaboratif plus poussé, produit sur le long terme, à une fréquence soutenue, et demande d'être considérés pour toutes leurs utilisations possibles. Enfin, pour passer du prototype proposé à une infrastructure de recherche capable d'alimenter des systèmes d'aide à la décision dans le domaine environnemental, l'animation de groupes de travail interdisciplinaires (recherche thématique et recherche informatique) opérationnelle sur le plan international est indispensable ; celle-ci doit s'appuyer sur un personnel qualifié et dédié, et avoir pour objectif le décloisonnement des recherches. Elle doit favoriser l'augmentation du temps de travail pluri-/interdisciplinaire en commun et des moyens dédiés à long terme aux processus de fouille et de curation des données pour l'aide à la décision dans le domaine environnemental.

En perspective, l'utilisation de la grille de calcul défini lors des ateliers pour faire de la fouille de graphes de manière parallélisé est proposé, avec le challenge de passage à l'échelle avec des données distribuées et très hétérogènes formant des graphes de plus d'un milliard de noeuds et plusieurs centaines de milliards de liens. Il sera possible de i) développer ces

[2] Se dit d'une base de données, d'une application ou d'un système d'information qui ont des caractéristiques spécifiques à un métier, et évoluant avec le métier pour lequel ils sont conçus.

[3] Donner un contexte à une donnée correspond à l'attribution d'un qualificatif correspondant aux conditions d'enregistrements ou aux conditions particulières, biotiques ou abiotiques, dans lesquelles la donnée a été mesurée (p.e. température, profondeur, pente, orientation).

travaux en s'intéressant aux différents algorithmes de fouille de ces graphes, ii) s'intéresser aux verrous bloquant le passage à l'échelle sur de très grands jeux de données et iii) tester leur mise en œuvre pour l'aide à la décision dans le domaine de la gestion de la biodiversité. Pour atteindre ces objectifs, les organismes s'occupant de recherche en écologie et ayant financé ces avancées doivent continuer à soutenir les activités du consortium IndexMEED créé dans le cadre de ce travail, ou développer des groupes interdisciplinaires semblables et les financer sur le long terme.

Summary

From designing a large-scale observation system to deploying and operating its information system: application to the observation of coralligenous habitats and the colonization of artificial reefs (ARMS)

In the field of the marine environment, the frequency of data collection campaigns collected during research programs, environmental monitoring [European bonds among others], impact studies, (field missions, optical or radar sensors monitoring of water quality, automatic or semi-automatic census of taxa, etc.) lead to the accumulation of a considerable amount of data. Observation protocols are constantly developed in many settings, and produce very heterogeneous data centered on the use often specific to a profession that their producer is considering. Because of their heterogeneity, these are difficult to aggregate to get an overview (sometimes called “stacking databases”). In addition, access to data is unorganized and often proves difficult, if not impossible. This access for multiple users and the aggregation of data on a large scale are nevertheless essential to better understand the issues of protection of biodiversity and marine resources, and to anticipate their irreversible deterioration.
In order to better protect marine biodiversity and especially so that the issues of conservation and preservation of resources are better taken into account by public policies, it is necessary to strengthen the coherence of observation systems and acquisition of new knowledge, and organize access to data and search results for all potential users. Making proposals to improve the coherence between observation systems and information systems is the main objective of this work.

The rationalization of the means to invest for preserving the good environmental state is only possible by realizing an inventory of the knowledge produced, the available competences and the locks to be lifted (the blockages in term of base of knowledge as well as the methodological blockages , sociological, scientific and functional described in the systems currently implemented) for the efficient implementation of large-scale observation systems. Necessarily, observation systems must be coupled to an information system architecture. This architecture must make it possible to organize the creation and access to knowledge, to facilitate its conservation, to harmonize methods and systems of management and data analysis.
The construction of a long-term network approach and project is a prerequisite for the sustainability of large-scale monitoring systems. The monitoring networks then generated

then become multi-users and should allow this rationalization of the means invested in the production of new knowledge. They must produce sufficient reliable descriptors to develop a powerful indication, which is itself necessary for reporting (M.S.F.D.,H.F.F.D., W.F.D., French Seaboard, N.O.B., etc.).

This thesis aims to i) propose methods, protocols and recommendations to build and / or support the establishment of networks of useful and sustainable monitoring of biodiversity at the scale of a biogeographic zone or at the international level, relying on local actors (devices ranging from gene tracking to species and habitat monitoring); ii) fostering multiple and innovative uses of data while preserving the rights of the author / inventor of the devices; improve different access to data (raw, processed and summary)

Two "case study" were chosen for this work: coralligenous habitats at the Mediterranean scale and the colonization of artificial reefs (ARMS) in different regional seas by focusing

- the construction of sustainable and useful monitoring and observation networks for different types of use
- on the effective sharing of long-term knowledge with its various potential users (scientists, managers, elected officials, amateurs, general public ...) and on the inter-operation of information systems
- methods, tools and data analysis interfaces exploiting new advances in the field of Big Data, the management of heterogeneous data and their analysis in the form of graphs.

As coastal habitats are the main target of this work, the testing of the different protocols shows that a large-scale experiment must absolutely be described very explicitly in standardized terms even beyond the disciplinary field of marine ecology (if possible in organizing in micro thesauri to "align" with standards being developed) and rely on the simplest possible methods of measurement to implement. The tests carried out by different operators have shown the importance of training and then iterative testing in the form of comparing the results on the same observation, whether in the field or during the ex situ analyzes. The learning times are also adapted to the types of objects and / or methods concerned (taxon, habitat, determination, counting, measurement ...).

The work on the architecture of information systems and the debates concerning the life cycles of the data made it possible to highlight the inefficiency of a centralized system, and the inevitability of a modular management system, “job oriented” and decentralized. It follows that the current non-management of the authorities is a lock for the traceability of the data and the recognition of producers.

Thanks to the organization of access to data in the form of configurable and open flows, it has been proposed as part of this thesis a mechanism for coupling data of different origins (field observations and data describing the contexts) based on the requalification of

heterogeneous descriptive factors into equivalent and simplified factors whose choice is based on a collaborative arbitration between specialists. Based on a prototype, a new method of environmental data analysis and graph-based data mining concept organization was put into demonstration and developed to become generic. Examples of data visualizations and different types of possible data-based demonstrations were constructed through the organization of curation and data visualization workshops in the form of graphs.

In conclusion, the first functional tests have shown that the information produced must be able to be controlled in real time and in an iterative manner, and that the curation processes of the data must necessarily be put in place at the same time as the design of observation processes. In addition, the definition of standards and the accessibility of context data require more collaborative work, produced over the long term and sustained frequency, and be considered for all their possible uses. Finally, to move from the proposed prototype to a research infrastructure capable of feeding decision-making systems in the environmental field, the animation of interdisciplinary working groups (thematic research and computer research) operational on the international level is essential ; it must rely on a qualified and dedicated staff, and aim at de-compartmentalizing research and fostering the increase of multi- / interdisciplinary working time in common and dedicated means for long-term data mining processes and curation of data for decision support in the environmental field. In perspective, the use of the calculation grid defined during the workshops to carry out graph mining in a parallel way is proposed, with the challenge of scaling up with distributed and very heterogeneous data forming graphs of more than one billion nodes and several hundred billion links. It will be possible to i) develop this work by looking at the different search algorithms of these graphs, ii) look at the locks blocking scalability on very large data sets and iii) test their implementation. for decision support in the field of biodiversity management. To achieve these objectives, organizations engaged in ecology research and having funded such advances must continue to support the activities of the IndexMEED consortium created as part of this work, or develop similar interdisciplinary groups and fund them in the long term.

Avant-propos

Dans cet avant-propos, j'ai trouvé essentiel de présenter le contexte dans lequel s'inscrit ce travail (Chapitre 1), pour guider le lecteur dans la structure pluridisciplinaire de ce manuscrit. Ce travail a été développé en partie dans le cadre de programmes de recherche européens (**CIGESMED** pour Coralligenous based Indicators to evaluate and monitor the "Good Environmental Status" of the MEDiterranean coastal waters et **DEVOTES project** pour DEVelopment Of innovative Tools for understanding marine biodiversity and assessing good Environmental Status) ayant entre autre pour objectif de mettre au point et de tester des protocoles permettant de produire, d'analyser, de gérer, et de diffuser des données écologiques et environnementales accessibles et réutilisables (Les aspects de ces programmes connectés avec cette thèse sont respectivement présentés en annexe 3 et 4).
La construction, le développement et la mise en œuvre des protocoles s'appuient sur des observateurs et des opérateurs de différents pays (Turquie, Grèce, Italie et France pour les données de CIGESMED, et Arabie Saoudite, Italie, France et Espagne concernant les données provenant du programme DEVOTES). En cela, ce travail est un prototypage de ce qui est faisable à large échelle dans le domaine de l'observation des substrats durs en milieu côtier.
Je me suis intéressé non seulement à la mise en œuvre de l'observation à large échelle dans le domaine côtier sur substrat dur (chapitre 2) mais aussi à la gestion puis aux utilisations des données qui sont produites (voir chapitre 3). En cela, il s'appuie sur différentes compétences encore peu développées dans le domaine de l'écologie marine, notamment en décrivant les processus de construction des accès à l'information puis d'utilisation des données (bases de données réparties, qualité de données hétérogènes et multi-sources, constructions sémantiques, construction des accès et des services associés à la donnée, systèmes d'analyse, de visualisation puis de fouille de données). Tous ces aspects sont autant de problématiques scientifiques à développer pour répondre aux objectifs de cette thèse. Ceci a nécessité de faire appel à la participation des spécialistes de certaines disciplines des Sciences et Technologies de l'Information et de la Communication (S.T.I.C.). Les S.T.I.C. sont souvent réduites par les "thématiciens"[4] à un aspect technique regroupé sous le terme "Informatique". Pour accéder à ce niveau d'interdisciplinarité, il a fallu construire un consortium interdisciplinaire et s'appuyer sur des personnes ressources expérimentées provenant du monde des S.T.I.C., mais aussi travaillant dans des domaines ou le

[4] Ce terme, utilisé par les communautés S.T.I.C. regroupe tous les scientifiques dont les disciplines, comme l'écologie marine, l'économie, la sociologie, l'archéologie, l'anthropologie, qui ne sont pas directement dans le domaine des S.T.I.C.

développement et l'usage des grands volumes de données (le Big Data) est un état de fait et non un balbutiement comme dans le domaine de l'écologie. Ce consortium, dont la vocation est de développer et tester des concepts génériques dans le domaine de l'écologie au sens large appelé IndexMed (annexe 5) puis IndexMEED (annexe 6) rassemble donc des écologues, des chercheurs dans différents domaines précités des S.T.I.C., mais aussi des archéologues, anthropologues, sociologues (travaillant souvent aussi avec des données très hétérogènes), et des astronomes et physiciens des particules (habitués aux grands volumes de données, à l'utilisation de données distribuées et aux systèmes permettant de les exploiter). Ce consortium qui s'est auto-développé autour des différents verrous opérationnels mis en exergue lors de mon travail de recherche a trouvé aujourd'hui une certaine notoriété. Son premier développement a été soutenu par la mission pour l'interdisciplinarité du CNRS et l'obtention du financement de projets "défi" et PEPS (Projet Exploratoire Premier Soutien). En continuité de ce travail, IndexMEED (Indexing for Mining Ecological and Environmental Data) développe son action au sein de plusieurs Groupements De Recherches (G.D.R.) comme MaDICS (Masses de Données, Informations et Connaissances en Sciences), Sémandiv (SÉMANtique de la bioDIVersité et dans une moindre mesure EcoStat (ECOlogie STATistique) sur le plan national, ou comme R.D.A. (Research Data Alliance) ou le T.D.W.G. (Taxonomic Databases Working Group) sur le plan international.

La collaboration avec ces différents experts a été l'occasion d'organiser des ateliers interdisciplinaires dont une partie des résultats a été présentée lors de conférences internationales (annexe 7) et développée dans des publications présentés dans ce manuscrit (annexe 1 pour les publications les plus importantes dans leur intégralité et annexe 7 pour les autres). Afin d'éliminer autant que possible les jargons propres à chaque discipline, il a fallu proposer des définitions accessibles à un non spécialiste et les moins ambivalentes possibles. Celles-ci sont regroupées dans le glossaire proposé à la fin de ce mémoire.

L'objectif de ce manuscrit n'est pas de faire une présentation exhaustive des systèmes d'observation en milieu côtier d'une part, ou des systèmes d'information sur l'environnement d'autre part. Mais du fait de l'interdisciplinarité du cadre et afin que chaque communauté comprenne, sans vision réductrice, les difficultés et les enjeux liés aux problématiques scientifiques des autres disciplines, il a fallu présenter aux non spécialistes des S.T.I.C, sans simplification excessive, la complexité des systèmes d'observation et leur enjeux dans le domaine du milieu marin côtier. Inversement, il a fallu présenter aux "thématiciens" les aspect S.T.I.C., les concepts et développements réalisés avec un vocabulaire le moins technique et le plus explicite possible. De manière globale, je voulais que ce travail soit compréhensible par n'importe quelle personne n'ayant pas de compétence particulière dans l'un des

domaines précités, et pour cela, qu'il soit parfaitement appréhendé en tenant toujours compte du cadre actuel.

J'ai donc jugé nécessaire autant que possible de brosser un tableau général concernant les méthodes d'observation usitées ou en développement (chapitre introductif avec des encadrés explicatifs pour les non spécialistes) et de faire un exposé sur les méthodes d'observations au début du chapitre 1). J'ai ensuite présenté la mise en oeuvre des protocoles et de leurs résultats (chapitre 2) ; les systèmes d'information multi-utilisateurs basés sur des données hétérogènes dans le domaine de l'environnement sont abordés à partir du chapitre 3 (avec des développements sur les défis de l'interopérabilité et de la mise en place de plans de gestion de données, et un exposé de grands principes concernant l'analyse de données en début de chapitre 4). Pour renforcer ce tableau général, chaque aspect conceptuel utile à la compréhension globale du travail et qui peut être "nouveau" pour un non spécialiste, est défini et replacé dans son contexte en début de chaque chapitre, et des encadrés font des focus sur des éléments de connaissance à même d'éclairer le lecteur sur des sujets particuliers.

Deux questions sont au coeur de la thèse : comment construire un système d'observation et ses protocoles (chapitre 2) couplé à un système d'information réparti et permettant l'analyse de données hétérogènes (chapitre 3)? En quoi ces méthodes peuvent être plus pertinentes en s'appuyant sur une description de l'architecture conçue dans le chapitre 4 et sur des méthodes d'exploration de données basées sur les graphes (décrites au début du chapitre 5). La réflexion sur la pertinence des services basés sur la donnée a aussi été menée dans le cadre du Work Package 6 du programme CIGESMED ("Data management, mapping and assimilation tools") dont j'ai été responsable.

Pour construire les réponses à ce questionnement, ce travail repose sur l'exploitation de deux cas d'études (des protocoles de suivi définis puis mis en œuvre dans le cadre de CIGESMED et de DEVOTES[5]) et s'appuie sur le développement et le test d'un prototype d'exploitation de données ainsi qu'une organisation des compétences et une articulation des disciplines lors de différents ateliers pluridisciplinaires (chapitre 5).

Certains résultats sont donc des propositions de méthodes d'exploitation des données, dont les choix dépendent non seulement d'une argumentation scientifique, mais aussi de raisons pratiques (financement, formation, efficacité de mise en œuvre, organisation humaine, etc.), et ce parce que l'observation à large échelle a clairement des objectifs opérationnels. La conception des systèmes d'observation puis d'analyse de données sur de larges échelles doit faire consensus entre les observateurs et es utilisateurs. Elle repose sur un travail

[5] Ce manuscrit fait référence de nombreuses fois aux différents programmes de CIGESMED, DEVOTES, IndexMed et IndexMEED. Ceci est nécessaire pour que le lecteur associe sans confusion les méthodes, résultats et discussions à chacun de ces programmes,.

nécessairement collaboratif. Afin de faciliter la poursuite de ces travaux, les procédés de mise en œuvre de l'analyse de données hétérogènes basée sur les graphes, et les outils conçus pour animer ce travail collaboratif, sont présentés comme un des résultats de cette thèse et ont été construits de manière à être transposables dans d'autres disciplines environnementales (Annexes 8 et 9). Les premières conclusions ont ensuite été replacées dans les questionnements plus généraux qui sont actuellement l'objet de travaux collaboratifs scientifiques dans des groupements internationaux (raison pour laquelle il était impératif de s'investir dans certains groupes de travail de ces assemblées internationales, que ce soit dans le cadre de l'écologie marine ou dans celui des S.T.I.C.). Ce questionnement et les résultats qui en découlent sont réutilisés pour étayer une série de premières recommandations pour de prochaines mises en œuvres de systèmes d'observation et d'information à large échelle dans le domaine de l'écologie au sens large (fin du chapitre 5). Tous ces résultats et les conclusions qui en découlent ont été publiés dans des journaux et des proceedings à comité de lecture (en annexe 1), c'est-à-dire en anglais. Cependant, j'ai préféré ne pas faire une thèse sur articles et rédiger mon manuscrit en français pour plusieurs raisons. Tout d'abord, je souhaite que mon travail puisse être repris et développé, ou au moins aide à la prise de conscience en France de l'importance des travaux sur la donnée sur le plan international. Mon travail doit aussi être accessible à mes compatriotes non anglophones car il s'adresse particulièrement à eux (Aix Marseille Université est financé par leurs impôts !). Finalement, je pense que ce manuscrit complète utilement les publications sur lesquelles il est basé, son format lui permet d'être plus explicite et précis, et améliore la compréhension de l'articulation entre tous les aspects de mon travail.

Mots clefs : coralligène, observation, base de données, méditerranée, graphes, systèmes d'information

Key words: coralligenous, observation, database, mediterranean, graphs, information systems

Chapitre 1 : Introduction : éléments de contexte et enjeux concernant l'observation en milieu marin côtier et les systèmes d'information à large échelle

1. Concepts généraux et enjeux concernant l'observation en milieu marin côtier

1.1 L'utilisation des navires, bouées appareillées et satellites pour l'observation

Les impacts de nouveaux usages et leur intensification en zone côtière (exploitation de granulats, fixations d'éoliennes en mer, développement d'infrastructures de forage, etc.) s'ajoutent aux impacts déjà récurrents des techniques de pêche, des effluents, polluants et macrodéchets d'origine terrestre. L'amélioration des connaissances du milieu marin est essentielle pour mettre en place une « économie bleue » (volet maritime de la stratégie Europe 2020), associant la recherche et l'innovation technologique, l'utilisation durable des ressources, la compétitivité et la création d'emplois en faveur d'une croissance intelligente, durable et partagée.

Les engins autonomes à large rayon d'action ont permis des progrès significatifs pour cartographier, mesurer et comprendre les environnements marins. Les suivis de biodiversité sont automatisés à leur tour depuis plus récemment avec l'émergence des outils de screening moléculaire (protéomique, métabolomique, génomique, métagénomique, etc.) et le développement de méthodes de reconnaissance automatique [e.g., les suivis génomiques du plancton lors des expéditions de Tara entre 2009 et 2013 (par exemple, Pesant S. *et al.*, 2015), ou les recherches sur la reconnaissance automatique du zooplancton (par exemple, Gasparini *et al.*, 2007 ou Gorsky *et al.*, 2010)]. Leur utilisation permet aujourd'hui d'engranger un grand volume de données. Par ailleurs, les méthodes de suivi par télédétection ou réalisées à la surface représentent aussi une part significative de notre connaissance du milieu marin, surtout sur les 15 à 20 premiers mètres de profondeur. Les données produites par ces engins et satellites sont de plus en plus diverses et volumineuses, et ont fait rentrer les disciplines de l'environnement marin dans l'ère du big data. Cette utilisation est plus limitée sur le littoral et dans les eaux côtières, zones les plus sensibles aux pressions et aux

effets des changements climatiques, et où vit une partie considérable de la biodiversité marine aujourd'hui menacée (Coll *et al.*, 2010). Pour autant, en Mer Méditerranée, même si elles ne sont pas aussi volumineuses, une large partie des données historiques concernant l'écologie marine se concentre dans la zone côtière (Fredj *et al.*, 1992).

1.2 L'observation in situ des fonds sous-marins

Les variables concernant le benthos

Les suivis benthiques sont généralement plus délicats à effectuer avec des engins autonomes, même si l'utilisation de ROV (Remotely Operated Vehicles) (Pelletier *et al.*, 2012) se développe. Ceux-ci demandent soit un pilotage de la surface soit une installation / désinstallation nécessitant des plongeurs. Ces outils, parfois difficiles à stabiliser car pilotés depuis la surface, sont plutôt exploratoires et restent peu utilisés pour des suivis scientifiques du benthos, car les variables mesurables par un ROV n'ont pas la même précision qu'avec un opérateur en plongée. Pour les données correspondant à l'étude des substrats durs, les données le plus souvent récoltées correspondent à l'abondance des colonies et leur recouvrement. Pour chacun des jeux de données, la précision des inventaires va de la catégorie benthique à l'espèce selon les compétences de l'observateur, sachant que certaines déterminations de familles et de genres sont déjà une affaire de spécialiste (Tableau 1).

Tableau 1 : Variables selon le type d'inventaire : les variables calculables à partir des inventaires de macrofaune ne sont pas les mêmes que celles liées aux inventaires sur substrat dur. (Source : WP2 du projet "PAMPA Indicateurs de la Performance des AMP pour la gestion des écosystèmes côtiers, des ressources et de leurs usAges[6])

Variables	**Macrofaune**	**Benthos/Substrat**
Présence/absence	X	X
Abondance ou autre statistique liée à l'abondance	X	X
Biomasse ou autre statistique liée à la biomasse	X	X
Fréquence d'occurrence, % couverture	X	X
Indices de diversité taxonomique (richesse spécifique totale ou moyenne, Shannon, Clarke et Warwick)	X	X
Indices de diversité fonctionnelle (richesse, régularité, divergence)	X	X
Indices de diversité trophique	X	
Niveau trophique moyen	X	
Indices de rugosité, complexité de l'habitat		X

A ces variables viennent s'ajouter les caractéristiques des espèces recensées dans les « référentiels espèces » et les conditions de collecte (propres à l'unité d'observation), que l'on nomme le plus souvent données de contexte.

Les méthodes d'étude destructrices

Concernant le benthos, les sédiments meubles sont mieux connus que les substrats durs, car leur étude de la surface a pu être systématisée grâce à différentes sortes de dragues et de bennes (voir quelques exemples figure 1_1 et figure 1_2) de prélèvements utilisables d'un navire, et permettant des études et des suivis avec un grand nombre de réplicas.

[6] https://wwz.ifremer.fr/pampa/

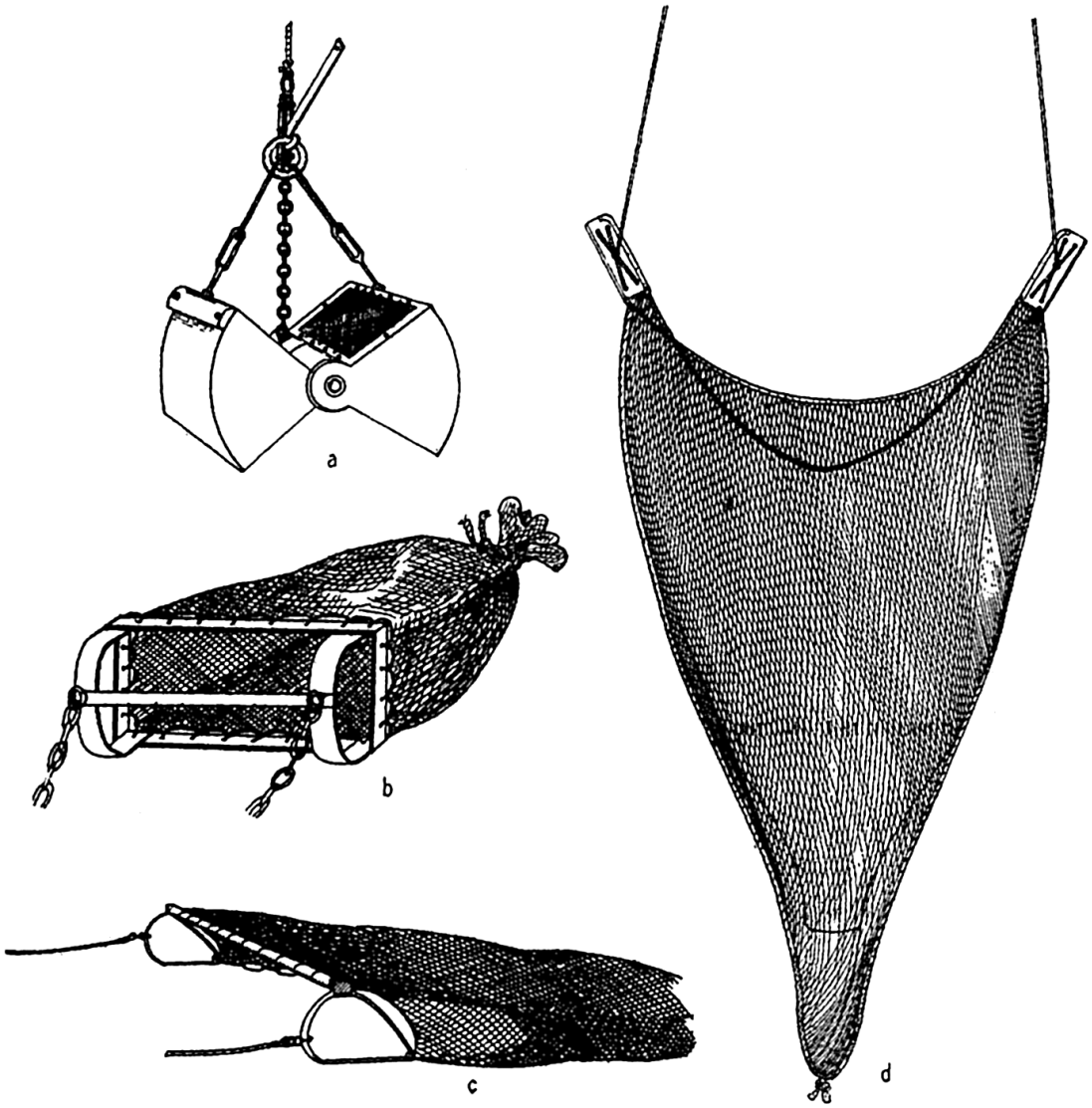


Figure 1_1. Principaux types d'engins de prélèvement benthique :

(a) Benne, (b) drague, (c) chalut à perche, (d) chalut à panneaux. Seules les bennes, sous certaines conditions, permettent des évaluations quantitatives. Les autres engins, traînants, sont plus ou moins sélectifs et tous sont destructeurs.

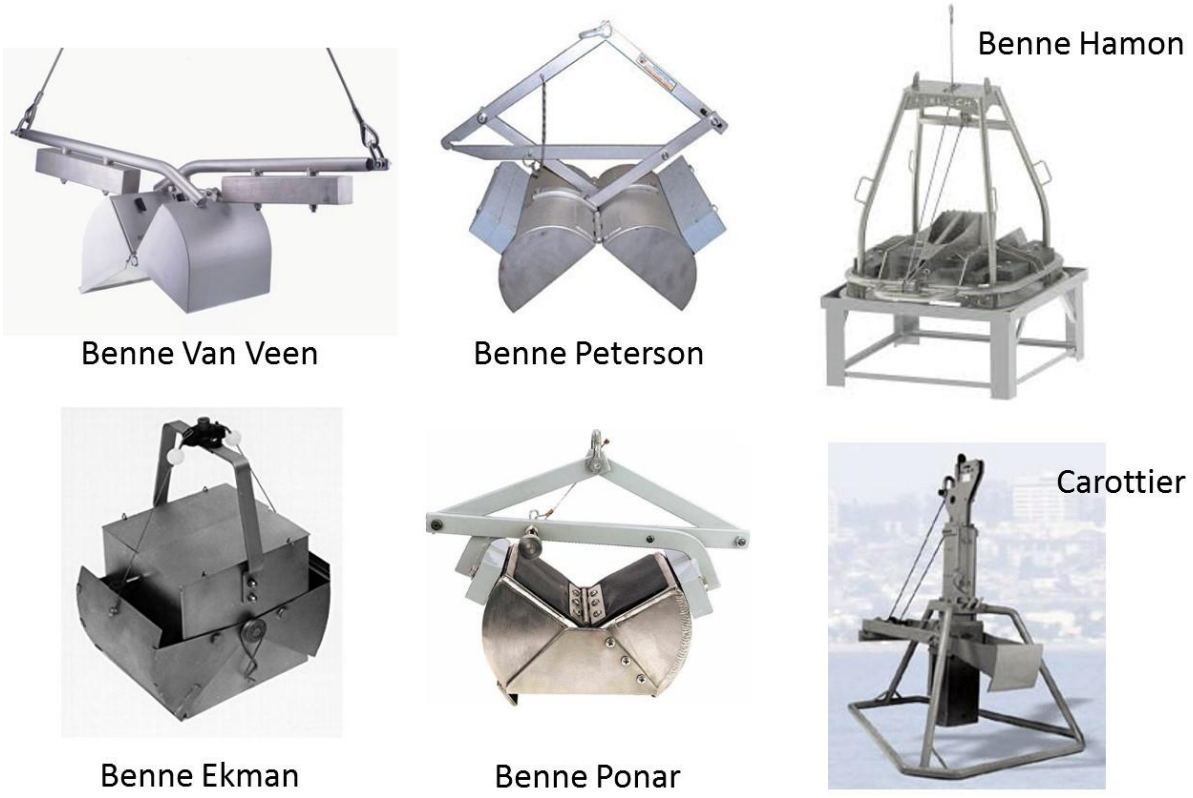


Figure 1_2. Exemples d'engins utilisés pour effectuer des prélèvements quantitatifs.

Les méthodes basées sur des prélèvements de substrat ont l'avantage de pouvoir travailler plus longtemps sur les échantillons, et de faire des déterminations précises (sous réserve de l'existence et de la disponibilité des spécialistes des groupes) et/ou des échantillonnages importants. Les échantillons peuvent être conditionnés pour être présentés à des spécialistes et étudiés de différentes manières, combinées ou non (microscopie classique ou électronique, approches moléculaires telles que la génomique, la protéomique, la métabolomique...)

Elles sont surtout développées pour les substrats meubles, pour lesquels des engins de prélèvement comme des bennes peuvent être mis en place depuis la surface. Ces engins permettent des approches quantitatives, basées sur un grand nombre de réplicas. D'autres méthodes comme les dragues et les chaluts permettent de prospecter des surfaces plus grandes et d'étudier le substrat, mais de manière qualitative.

L'inconvénient de toutes ces méthodes réalisées depuis la surface est de cibler plus difficilement un substrat précis, surtout lorsque la mosaïque d'habitats présents est particulièrement complexe. Ces méthodes sont difficiles à mettre en oeuvre sur des substrats durs, et encore plus sur des substrats biogènes, plus complexes et plus fragiles. Les prélèvements les plus ciblés y sont réalisés par des plongeurs. L'aspect destructeur de ces

prélèvements peut paradoxalement paraître plus évident aux gestionnaires que l'impact d'un engin traînant sur une beaucoup plus grande surface de substrat meuble. Les prélèvements sont donc censés être effectués avec parcimonie, d'autant plus que les milieux et/ou les espèces concernés revêtent une importance patrimoniale acceptée à l'heure actuelle comme plus étant forte que pour les substrats meubles. Certaines de ces techniques ont été utilisées dans le cadre des programmes scientifiques dont fait partie cette thèse (CIGESMED et DEVOTES), tels les grattages en plongée de petite surface, ou les échantillonnages d'espèces (réputées fréquentes) pour des études de génétique des populations en relevant des paramètres de contexte. La conception, le test et la mise en œuvre de protocoles de prélèvement et d'observation des assemblages d'espèces *in situ* vont permettre, grâce à une description commune de variables de contexte, de comparer des habitats moins différents, et ont pour objectif d'améliorer la détection des changements, qu'ils soient d'origine humaine ou naturelle. Cette comparaison devrait enfin permettre lorsque c'est possible de remplacer des méthodes destructrices par un équivalent plus acceptable y compris par des gestionnaires du milieu.

Les méthodes non destructrices reposant sur la plongée

Le recensement visuel sous-marin ou "Underwater Visual Census" (U.V.C.) :

L'U.V.C. constitue une technique non destructrice d'évaluation de la biodiversité et de son abondance de plus en plus utilisée et dont les limites ont été maintes fois décrites (Harmelin-Vivien *et al.*, 1985 ; Kulbicki *et al.*, 2010, MacNeil MA *et al.*, 2008). L'U.V.C. est utilisée par la communauté scientifique depuis les années 70 dans les différentes mers du globe et adaptée aux A.M.P. pour les suivis des peuplements de poissons (Claudet et Pelletier, 2004 ; Claudet *et al.*, 2006).

Elle est souvent mise en oeuvre pour suivre les évolutions de populations de poissons notamment en Méditerranée et a été comparée avec les récents systèmes d'observation video rotatifs (STAVIRO[7]) (Prato *et al.*, 2017) : avec l'apparition de matériels moins onéreux et plus performants, la vidéo sous-marine s'est développée et présente des avantages pour le suivi et la gestion des écosystèmes marins. Ces techniques non destructrices permettent une bonne observation des peuplements ichtyologiques (Pelletier *et al.*, 20012). Par exemple, le STAVIRO, à l'exception de son installation, ne requiert pas de plongeurs professionnels, ce qui permet de s'affranchir de l'effet plongeur (Pelletier et Leleu, 2008) et limite les coûts d'observation. Un autre modèle de station d'observation installé sur le substrat, appelé MICADO[8] enregistre des images selon des intervalles programmés, de

[7] https://wwz.ifremer.fr/webtv/Thema/Ressources-halieutiques/STAVIRO-et-MICADO
[8] https://wwz.ifremer.fr/webtv/Thema/Ressources-halieutiques/STAVIRO-et-MICADO

l'aube au crépuscule. Ces techniques nécessitent encore de dépouiller l'échantillon recueilli et de s'appuyer sur des techniciens pour analyser les images, ce qui est relativement coûteux en temps. D'autre part, la prise d'images sous-marines et les réglages du matériel ne sont pas des exercices triviaux et doivent être faits par un personnel qualifié et expérimenté.

Les stations vidéo ont l'avantage de pouvoir être déployées de jour comme de nuit, sur une large gamme de profondeurs (Cappo *et al.*, 2006), permettant l'observation des espèces qui ont tendance à fuir les plongeurs. Les inconvénients majeurs des techniques actuelles de vidéo en station peuvent être identifiés comme suit (source : programme PAMPA) :

- Il est très difficile de calculer une densité absolue du fait de la difficulté à évaluer le volume observé ;
- Il y a toujours un risque de surestimation de l'abondance lié aux possibles doubles comptages ;
- Le temps d'analyse *a posteriori* peut s'avérer conséquent au vu du nombre d'observations réalisables ;
- Il existe un effet (mineur) lié au passage du bateau et à la pose de la caméra ;
- Une station rotative vidéo (R.V.) par exemple comporte en général quelques rotations de l'ordre de la minute, soit un temps d'observation total d'au maximum quelques dizaines de minutes concentrées à un moment de la journée, à moins de réitérer l'installation à différents moments de la journée et/ou de la nuit (ce qui ne permet d'observer qu'une partie de la faune, celle qui sort de ses abris à ce moment-là).

L'objectif des suivis est souvent de faire des comparaisons avec des études et des situations antérieures. Pour y parvenir, utiliser un matériel et une méthodologie comparables aux suivis précédents est impératif (les capteurs, appareils photos, vidéos évoluent ainsi que les formats de données qu'ils produisent).

Concernant les suivis du benthos en plongée, l'observateur peut noter un certain nombre de paramètres relatifs aux peuplements : la diversité, l'abondance, la taille individuelle. L'observation en plongée présente le gros avantage de permettre à l'observateur de compléter ces informations grâce aux données de l'environnement (habitat, courant, visibilité, etc.) recueillies simultanément, et se fait le plus souvent en utilisant une surface de référence normalisée (quadrats ou transects).

Il existe de plus un biais reconnu dû à l'observateur, sa compétence et son entraînement. Ce dernier biais peut être réduit si un apprentissage et une re-validation itérative des compétences sont réalisés au préalable. Cette variabilité « observateur » peut être mieux

appréhendée lorsque deux plongeurs font leurs observations sur le même transect. A noter que l'observation se limite à la surface externe et que toutes les espèces peu visibles ou endogènes ne sont pas prises en compte. A titre d'exemple, Hong (1980) dénombra plus de 1600 taxons sur des prélèvements de coralligène, alors qu'un plongeur expérimenté n'en recensera environ qu'une centaine lors d'une plongée. La plus grande partie des espèces n'est jamais observée en plongée malgré une présence probable.

Enfin, il existe aussi des stations d'expérimentation, posées puis récupérées par des plongeurs (exemple d'une cloche benthique, figure 2) qui peuvent selon le site rester plus ou moins longtemps en place (quelques heures à quelques jours pour les appareillages fragiles, à quelques mois sans intervention en plongée pour les balises de mesure telles que les capteurs de température ou les réseaux comme le réseau national SOMLIT -Service d'Observation en Milieu Littoral[9]). Ces systèmes un peu plus normés nécessitent *à minima* deux plongées (installation puis retrait) et souvent des plongées d'entretien au moins mensuelles (nettoyage des capteurs sur les bouées pour éviter le *fouling* par exemple). Dans le cadre de SOMLIT, des capteurs sur divers supports permettent des mesures automatiques haute fréquence (température, salinité, oxygène, turbidité, chlorophylle du phytoplancton) jusqu'à soixante mètres avec transmission journalière. Ces données sont essentiellement des données physiques et n'apportent pour le moment que peu d'éléments concernant la biodiversité elle-même.

[9] http://www.SOMLIT.INSU.fr

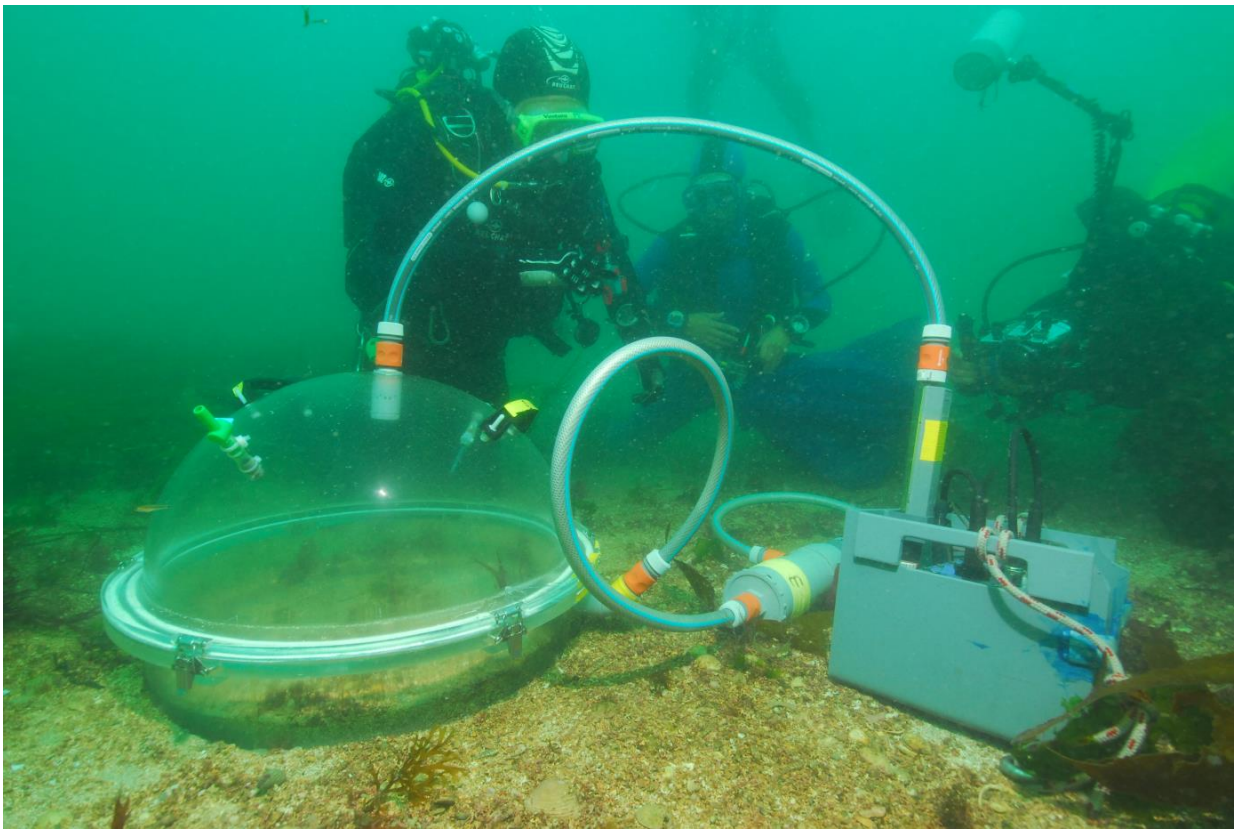


Figure 2 : exemple d'une expérimentation en plongée avec installation puis retrait d'une cloche benthique pour mesurer le métabolisme du sédiment. A noter que les prises vidéo et photo peuvent permettre de détecter* a posteriori *une erreur dans la mise en place de l'expérimentation (par exemple, dans ce cas, un tuyau qui pourrait être débranché) © F. Zuberer.

Zoom sur la plongée scientifique : un développement nécessaire pour le littoral et la zone côtière

L'utilisation de capteurs immergés apporte aussi un complément de connaissances indispensable à une description plus réaliste de ce milieu en 3D. L'observation, l'expérimentation, la maintenance, la récupération ou le remplacement des capteurs immergés rendent incontournable l'intervention des plongeurs.

Figure 3 : Prélèvement de pontes de gorgones rouges (Paramuricea clavata) – © R. David.

En effet, ces capteurs ne peuvent pas toujours se substituer à l'homme et une partie complémentaire de l'expertise, principalement dans le domaine des sciences de la vie, s'effectue obligatoirement en plongée : recherche de nouvelles ressources, recensements d'espèces, cartographie d'habitats, expertise de l'état écologique d'un milieu ou de l'effet de mesures compensatoires, ou d'évitement des impacts. Par ailleurs, la taille des objets pertinents en biologie et en écologie est souvent inférieure à la résolution spatiale des images acquises par télédétection.

Pour la plongée à caractère scientifique en France (Figure 3), la gamme de profondeur 0 – 50m est plus souvent utilisée, car le nombre de personnes formées à l'espace lointain est faible et le matériel utilisé plus difficile à maîtriser (mélanges gazeux à trois gaz). Le développement des systèmes de surveillance intelligents s'appuie sur les compétences des chercheurs en matière de plongée et la démocratisation de nouvelles technologies moins onéreuses. En France, le cadre juridique de la plongée scientifique est fixé par la loi (décret n°2011-45 du 11 janvier 2011). Il impose le Certificat d'Aptitude à l'Hyperbarie (C.A.H.). Il définit les matériels, types de plongée, risques, normes de sécurité, rôles, responsabilités, aptitudes et formations pour la pratique de toute plongée réalisée dans le cadre des institutions de recherche. Le décret instaure quatre classes de plongeurs limitées par une profondeur maximale (12, 30, 50 et > 50 m). Tout plongeur scientifique est astreint aux dispositions qui concernent la formation, l'encadrement, les équipements, comme la pratique des opérations de plongée ou le suivi médical. Les standards de formation français ont servi de base à l'établissement des standards européens. La plongée scientifique demande des infrastructures et la mise en place de nouveaux moyens d'investigation en plongée dans des zones plus profondes. Le développement des recycleurs et des mélanges gazeux, tout en augmentant la sécurité, permettra des interventions dans la zone des 100 m. Il est intéressant de noter que nous en savons moins sur les fonds marins que sur le sol de la lune : au-delà de 200 m de profondeur, moins de 10 % du relief des fonds marins est connu, selon l'Organisation hydrographique internationale.

1.3 La gestion de la donnée scientifique d'observation sous marine : état des lieux d'un domaine peu développé

Concernant la gestion de la donnée, qu'elle ait été prise par le plongeur ou effectuée par un automate, la photographie et *a fortiori* la vidéo sous-marine ont l'avantage de permettre un archivage des fichiers source qui peuvent ensuite être réétudiés ou redéchiffrés en fonction des avancées des méthodes de reconnaissance, qu'elles soient manuelles ou automatiques. Ces images représentent des volumes considérables qu'il est difficile de conserver utilement, c'est à dire en les documentant suffisamment pour comprendre ce qu'elles représentent (quoi, où, quand et comment) et quels sont les facteurs de contexte qui pourraient les enrichir. Compte-tenu de la relativement faible qualité de certains enregistrements, un pré-tri doit être effectué lors de cette démarche d'archivage. Parfois, les enregistrements constituent également des archives qui peuvent être analysées de nouveau *a posteriori*, non seulement pour limiter l'effet observateur mais aussi pour faire de la formation d'opérateurs. Lors d'une

plongée d'observation, les enregistrements peuvent enfin confirmer une observation du plongeur ou fournir des informations complémentaires sur l'habitat et le comportement de la faune. De manière générale, ils peuvent aussi être utilisés comme un outil pour améliorer la qualité des protocoles en plongée (comportement, placement des observateurs, équipement…)

La récolte de données basée partiellement ou totalement sur la plongée est rendue plus difficile que celle d'autres données sur la biodiversité étant donnés les coûts de déplacement (une sortie bateau monopolise en général un à plusieurs marins en surface et au moins deux plongeurs) et le maintien d'un matériel onéreux, fragile et à durée de vie limitée (caissons photos, détendeurs…) Ces investissements sont parfois pris en charge *via* des réponses à appels à projet ponctuels, sélectionnés pour leur originalité et donc proposant un nouveau protocole dont l'équivalence avec les précédents est difficile à jauger.
Pourtant, les suivis à long terme, condition *sine qua non* d'une meilleure appréciation des variations naturelles des variables surveillées d'une part et d'autre part de l'influence de facteurs anthropiques sur ces variables, devraient être répétés et donc reproductibles par d'autres, grâce à une documentation testée et des outils de conservation et d'analyse de données accessibles à toute personne souhaitant compléter les dispositifs existants avec une donnée comparable. L'approche scientifique *stricto sensu* crée des protocoles nécessairement différents des précédents, à des fins de publication originale, ce qui fait que les types, formats et contenus des données évoluent fortement. Les évolutions techniques de plus en plus rapides augmentent encore cette variabilité dans les types, formats et qualités des données. Cette variabilité très importante nuit à la comparabilité dans le temps et dans l'espace et empêche le véritable développement de systèmes d'information alimentés par des réseaux d'observation à large échelle (Hajj-Hassan, 2016). En conséquence, aujourd'hui, la conception et la mise à disposition de la documentation de la donnée est souvent négligée et parcellaire, et la donnée est moins structurée dans le domaine scientifique que dans celui de la gestion d'espaces naturels (qui doit pourtant reproduire des protocoles au moins approuvés par des scientifiques). Cette lacune vient probablement du fait que le travail de documentation de la donnée qui doit être fait par le scientifique n'est aujourd'hui pas un élément important comptant dans l'évaluation de son travail et favorable à sa carrière scientifique.

On constate aujourd'hui la complexité grandissante de la mise en œuvre de protocoles peu standardisés, avec des coûts de production grandissants, et dont les données ne sont pas utilisées *a posteriori*. De plus, étant donné la prédominance de la peur de se faire déposséder du bénéfice de la publication à partir de jeux de données (produits par un chercheur dans le

cadre d'un projet de recherche), les notions de "plan de gestion de données" sont encore considérées comme secondaires dans le domaine de la recherche en écologie. A cela, il faut ajouter que les investissements matériels informatiques et la disponibilité des compétences pour gérer la donnée sur le long terme sont souvent insuffisants et précaires, ce qui fait que de nombreuses données historiques n'existent plus qu'au format papier dans des thèses et sur des disques durs dont l'obsolescence est atteinte en moins de dix ans, ou parfois même disparaissent purement et simplement.

Ces dernières années toutefois, grâce à l'impulsion de l'Etat, de l'Europe et de la modification des conditions d'éligibilité des appels à projet de recherche (comme l'existence d'un plan de gestion de données de qualité, de critères de reproductibilité des données et des expériences), des essais de mise en réseau et d'utilisation des infrastructures déployées depuis plus de 20 ans dans d'autres domaines tels que l'astronomie, l'astrophysique, la météorologie ou la physique des particules se développent. Cela laisse espérer le développement de protocoles communs entre gestionnaires et scientifiques, la mise en place de réseaux de surveillances pérennes (voir le focus sur les "réseaux de surveillance" page 37) et l'utilisation directe de résultats scientifiques prendront en compte les difficultés de terrain dans le domaine de la gestion.

Néanmoins, cette mutation nécessaire est difficile : il n'y a encore que peu de cohérence de ces réseaux de surveillances en plongée au niveau européen, sans doute parce que les financements étant sur projets, le suivi de ces systèmes s'appuie sur un personnel en situation précaire, dont l'embauche au-delà de quelques années est rendue impossible. De plus, sans doute du fait de la multitude des acteurs et instituts différents s'intéressant à ces données et à la concurrence que cela engendre, aucun pilotage clair n'est mis en œuvre de manière consensuelle pour étayer compétences et bonnes pratiques sur le long terme.

Pour autant, les enjeux de la conservation et de l'interopération des données sur le long terme font consensus et sont renforcés en France par l'application de nouvelles directives comme la DCSMM (Directive Cadre Stratégie pour le Milieu Marin) et une nouvelle organisation comme l'Agence Française de la Biodiversité.

Focus sur les "réseaux de surveillance" et laboratoires scientifiques ayant une activité de production et de gestion de données concernant le benthos en Méditerranée française.

Les premiers réseaux de suivi sur le long terme du benthos en Méditerranée concernent les substrats meubles et le phytobenthos (par exemple Boudouresque, 1971 ou Bourcier,

1988). La mise en place de réseaux de surveillance concernant le benthos de substrat dur est somme toute assez récente, et liée à la prise de conscience provoquée par les marées noires de la fin du siècle dernier (dont l'Erika en 1999).

Deux réseaux benthiques à l'échelle de la façade méditerranéenne française existaient avant la création du RÉférentiel BEnthique Méditerranéen (REBENT) : le R.S.P. (Réseau de Suivi des Posidonies) créé en 2001, et le R.S.G. (Réseau Survie des Gorgones) créé en 1984. Lors de sa mise en place, et en plus des deux réseaux existants, le REBENT a intégré des réseaux plus locaux comme le Réseau Littoral Méditerranéen (R.L.M.), le Réseau de Surveillance Lagunaire (R.S.L.), mais aussi le réseau de surveillance Caulerpe (Suivi de l'expansion de *Caulerpa taxifolia* et de *Caulerpa cylindracea*), et Seagrass-Net, le réseau de surveillance des magnoliophytes marines.

Un certain nombre de données provenait aussi de suivis du Centre d'Océanologie de Marseille (COM) à Cortiou (suivi depuis 1965) et de suivis scientifiques dans les étangs de Berre et de Vaine organisés par le GIPREB (Groupement d'Intérêt Public pour la Réhabilitation de l'Etang de Berre) depuis juin 2000. Le reste des données provenait de suivis dans les A.M.P. (Aires Marines Protégée) lorsqu'elles avaient pu mettre en place des protocoles en s'appuyant sur leur réseau de laboratoires scientifiques [Parc National de Port-Cros (P.N.P.C.), Réserve de Scandola, Réserve de Cerbère-Banyuls s'appuyant notamment sur le laboratoire Ecosystèmes Aquatiques Tropicaux et Méditerranéens FRE E.P.H.E.-C.N.R.S. 2935 (Université de Perpignan), Réserve Naturelle des Bouches de Bonifacio (R.N.B.B.)] et si elles bénéficiaient d'un employé de réserve qualifié pour les réaliser. La plupart des espèces patrimoniales y sont suivies (*Pinna nobilis*, *Paramuricea clavata*, *Lithophyllum* sp., *Posidonia oceanica*, Mérous, etc.). Un suivi concerne également l'ichtyofaune en général dans ces A.M.P.. La R.N.B.B. avait mis en place une base de données rassemblant ces données et faisait en plus un suivi particulier de la gorgone rouge (*Paramuricea clavata*). Enfin, sur les sites où ont été implantés des récifs artificiels comme celui de la Côte Bleue, un suivi de l'ichtyofaune était réalisé régulièrement pour essayer de mettre en évidence l'effet réserve. En complément, certaines collectivités avaient mis en place des suivis proches de leurs rejets pour essayer d'en apprécier les impacts (Cassis, Giens, Hyères) ou proches de leurs aménagements (ports de Sausset-les-Pins, Porquerolles, de l'Aygade, de Saint-Tropez, l'anse des Sablettes, la plage de la Capte, etc.) En fin de compte et en toute logique, ce sont les secteurs proches des laboratoires et dans les A.M.P. ayant du personnel que les suivis préexistaient.

Depuis, REBENT a priorisé l'acquisition de données nouvelles dans la zone de balancement des marées et les eaux côtières concernées par la D.C.E. allant jusqu'à 5 miles des côtes (les zones de petit fond ont essentiellement été concernées). Ces suivis

concernent très majoritairement la macro flore benthique (macro algues et phanérogames marines) et les invertébrés benthiques de substrat meuble.

De 2008 à 2011, dans le cadre d'appels à projet appelés LITEAU, le programme PAMPA a eu pour objectif de construire et tester des indicateurs portant sur les écosystèmes, les usages et la gouvernance, pour évaluer la performance de systèmes de gestion des écosystèmes côtiers incluant des Aires Marines Protégées (A.M.P.) Le projet est développé en partenariat avec quatre A.M.P. méditerranéennes : les Réserves Naturelles des Bouches de Bonifacio, de Cerbère-Banyuls, le Parc Marin de la Côte Bleue, et le Cantonnement de Pêche du Cap Roux. Lors du déroulement de ce projet, des formats de données, des référentiels et une base de données multithématique ont été développés (biodiversité, usages, gouvernance) et des outils de calcul et de représentation des résultats, ainsi que des protocoles harmonisés de collecte des données ont été mis en place, pour la première fois, sur de nombreux sites simultanément. Ce projet a développé des indicateurs basés sur le benthos, mais ceux-ci ont essentiellement été testés dans les réserves outre-mer. Les données produites ont été intégrées dans le système de gestion des données de l'IFREMER appelé SISMER (Systèmes d'Informations Scientifiques de la Mer). Ce système de gestion de bases de données contient aujourd'hui en majorité des données issues de programmes de recherche et de surveillance gérés par l'IFREMER.

"Les systèmes d'informations gérés par le SISMER s'étendent du C.A.T.D.S. (données du satellite S.M.O.S.) aux données de géosciences (bathymétrie, sismique, échantillons géologiques) en passant par les données de la colonne d'eau [physique et chimie, données pour l'océanographie opérationnelle -Coriolis- Copernicus C.M.E.M.S. (Copernicus Marine Environment Monitoring Service)], données halieutiques (Harmonie), données d'environnement côtier (Quadrige 2), données d'environnement profond (Archimède)."[10]

Du fait de cette richesse thématique, les données sur la biodiversité du benthos ne représentent qu'une toute petite partie de SISMER et la difficulté de maintenir et faire évoluer ce système en permettant une réutilisation pertinente des données est manifeste et reconnue (communications lors du séminaire de restitution du programme LITEAU 3[11]).

Depuis 2011, la France a initié la mise en œuvre de la directive cadre communautaire "stratégie pour le milieu marin" (D.C.S.M.M.) de 2008. Dès lors, celle-ci s'applique à quatre zones appelées "sous-régions marines" : la Manche-Mer du Nord, les mers celtiques, le golfe de Gascogne et la Méditerranée occidentale. Pour chaque zone, un plan d'action pour le milieu marin (PAMM) a été élaboré et mis en œuvre, en vue de réaliser ou de maintenir un bon état écologique (B.E.E.) du milieu marin au plus tard en 2020. Ce plan

[10] http://data.ifremer.fr/SISMER

[11] http://www1.liteau.net/index.php/projet/liteau-iii

d'action pour le milieu marin (PAMM) est constitué de différentes phases itératives répétées tous les six ans : une évaluation initiale et une caractérisation du bon état écologique. La définition du Bon État Écologique s'appuie sur 11 types de descripteurs (1. Diversité biologique, 2. Espèces non indigènes, 3. Espèces exploitées, 4. Réseaux trophiques marins, 5. Eutrophisation, 6. Intégrité des fonds marins, 7. Conditions hydrographiques, 8. Contaminants, 9. Questions sanitaires, 10. Déchets marins, 11. Énergie marine) chacun géré par différents acteurs du milieu marin (ANSES, B.R.GM, CNRS, Ifremer, MNHN, SHOM en tant que chefs de files). Là encore, on peut remarquer que les "descripteurs écologiques" ne constituent pas la majorité des travaux de suivi et de mesure[12]. Seuls les descripteurs 1, 4 et dans une moindre mesure 2 sont basés sur des composantes liées à la biodiversité. Parmi ceux-ci, seul le D1 comporte, pour partie seulement, des programmes de suivi et des programmes de mesure concernant le benthos de substrat dur. Etant donné que les moyens concernant la mise en œuvre de ces programmes sont très fluctuants, et vu le manque de recul qu'ont les spécialistes sur l'efficacité des protocoles mis en œuvre sur une large échelle, il est peu probable que les objectifs de B.E.E. soient réellement atteints en 2020.

Sur le plan européen, malgré l'harmonisation des contenus des descripteurs (tous constitués d'un ensemble de sous-descripteurs choisis pour leur pertinence et pour la capacité à pouvoir les produire), les programmes de suivi sont développés de manière autonome par chaque pays mettant en œuvre son PAMM. Les données produites, même si elles respectent de mieux en mieux les standards et les référentiels mis à disposition et reconnus par les instances européennes, sont encore très hétérogènes. Les dispositifs de suivi sont basés sur des protocoles développés localement et choisis dans chaque pays en fonction des moyens et des compétences disponibles. Concernant les suivis du benthos de substrat dur, les spécialistes compétents sur les taxons concernés sont de moins en moins nombreux dans les laboratoires, et de moins en moins disponibles.

Malgré cette hétérogénéité des données, des entrepôts de données se développent à toutes les échelles géographiques, avec pour objectif de créer des accès et des services basés sur l'agrégation de ces données hétérogènes. En plus des systèmes d'information des instituts précités, coexistent des systèmes locaux, départementaux, régionaux[13],

[12] Ici, on entend par mesure une action de gestion ayant pour but d'agir sur l'état du milieu.
[13] http://opendata.regionpaca.fr/

étatiques régionaux thématiques[14] ou généraux[15],[16] nationaux[17],[18], européens[19],[20] et internationaux[21],[22],[23]. Aujourd'hui, chacun de ces systèmes possède un périmètre de thématiques de données différent, et adopte des modèles de structuration des données différents. A cause de cette hétérogénéité et de ce manque d'interopérabilité, les systèmes d'information ne peuvent pas communiquer entre eux. Les producteurs de données les plus volontaires ne peuvent pas alimenter tous ces systèmes d'information. Il en résulte une augmentation continuelle de l'éparpillement, de l'hétérogénéité et du morcellement de la donnée et, force est de le constater, une proportion de plus en plus importante de perte de ces données, de moyens de les lire, voire la disparition des supports de cette donnée. Ces données sont noyées dans une mer de systèmes d'information hétéroclites. Même si elles sont théoriquement accessibles, elles sont souvent perdues car pas ou peu indexées, ce qui fait que personne ne peut les analyser, voire savoir qu'elles existent. Ce constat est particulièrement flagrant pour la donnée concernant le benthos de substrat dur pour laquelle peu de moyens sont disponibles pour permettre une véritable mise en cohérence des systèmes de production et de gestion des données.

2. Enjeux de l'observation du benthos de substrat dur en milieu côtier sur de grandes échelles géographiques

2.1 Généralités sur l'observation à large échelle

La surveillance de l'environnement sous-marin est compliquée à mettre en œuvre sur de larges surfaces. Les dispositifs d'observation se concentrent souvent sur des sites où plusieurs méthodes d'investigation peuvent être mises en oeuvre de manière simultanée ou non. Ces méthodes traditionnelles ont une faible résolution spatiale et temporelle et nécessitent un travail conséquent par unité de surface / temps qu'elles couvrent. Pour mettre en œuvre la Directive-Cadre « Stratégie pour le Milieu Marin» (D.C.S.M.M.), les États membres européens sont tenus d'améliorer les réseaux de surveillance et les méthodes et

[14] http://www.paca.developpement-durable.gouv.fr/silene-le-portail-public-des-donnees-naturalistes-r356.html
[15] http://www.paca.developpement-durable.gouv.fr/cartographie-interactive-a398.html
[16] http://www.occitanie.developpement-durable.gouv.fr/cartes-et-donnees-sig-r6096.html
[17] https://www.data.gouv.fr/fr/
[18] https://inpn.mnhn.fr/accueil/index
[19] https://www.seadatanet.org/Tools/Catalogues-follow-up
[20] http://www.emodnet.eu/
[21] http://www.gbif.fr/
[22] http://www.iobis.org/
[23] http://geobon.org/

techniques d’investigation à partir des navires mais aussi *via* des systèmes d’observation *in situ* (sous la surface). Cela ne peut être réalisé qu’en développant et en testant des systèmes de surveillance innovants et dont le rapport coût / efficacité est largement amélioré, ainsi que des indicateurs de l'état de l'environnement. Des synthèses permettent de faire un état des lieux des connaissances (Coll et al, 2010) et d’utiliser des méthodologies et technologies récemment développées pour améliorer les indicateurs de biodiversité marine et les méthodes de suivi (Danovaro et al, 2017). Les outils innovants y sont discutés concernant les technologies actuellement utilisées ainsi que les avantages et les inconvénients de leur utilisation dans la surveillance de routine. En particulier, de nouvelles informations concernent (i) les approches moléculaires, y compris les microréseaux, la P.C.R. quantitative en temps réel (q.P.C.R.) et les outils métagénomiques (*metabarcoding*) ; (ii) des méthodes optiques (à distance) et acoustiques ; et (iii) les instruments de suivi *in situ*. Leurs applications dans la surveillance du milieu sous marin dans le cadre de la DCSMM y sont de mieux en mieux documentées et s’appuient sur des études de cas afin d'évaluer leurs utilisations potentielles dans la surveillance marine de routine future (Danovaro et al, 2017). Ces technologies et protocoles récemment développés et plus accessibles financièrement peuvent présenter des avantages évidents en termes de précision, d'efficacité et de coût rendant réalisables une observation puis une évaluation réaliste.

2.2 Les limites actuelles de l'observation automatisée des habitats benthiques durs en milieu côtier

Les observations à large échelle de l'espace océanique côtier sont basées sur l'orthophotographie, mais le traitement des images (transformation de la photo en polygones dans un SIG) reste manuel jusqu'en 2000. En France, l'IGN a procédé à la numérisation de toutes les prises de vue aériennes réalisées depuis 1945 et les rend accessibles dans la BD ORTHO® Historique. L'IGN a initié la production d'une édition couvrant la France entière des années 50, avant les grands aménagements des années 60. Ces bases de données ont notamment servi à estimer la surface des herbiers de posidonie (*Posidonia oceanica*) à partir des années 80. Néanmoins, la numérisation de ces habitats est toujours manuelle, et les erreurs d'interprétation des photos amenant à sous-estimer ou surestimer les tailles des mattes de posidonies peuvent être nombreuses (amas de feuilles de posidonies mortes, herbiers clairsemés). Avec l'avènement de l'observation *via* les satellites, l'augmentation des capacités de stockage et l'amélioration continue des capteurs photo et vidéo, de nouvelles bases de données ouvertes accessibles aux chercheurs en écologie marine ont vu le jour. Celles ci permettent par exemple aujourd'hui de cartographier et de suivre les efflorescences algales. Cette grande masse de données fournie par l'observation de l'espace (*remote sensing* – télédétection) a nécessité le développement d'une méthode d'analyse automatisée. Cette méthode est basée sur la segmentation de l'image en zones en fonction de leurs caractéristiques (en général un ou plusieurs spectres), puis à l'attribution de chaque zone à une classe correspondant à un type d'objet. Ces techniques très utilisées aujourd'hui en milieu terrestre ne sont utilisables que dans la zone peu profonde (maximum 40 mètres) et dépendent de la turbidité de l'eau. Elles sont d'autant moins adaptées à des habitats qui peuvent être verticaux (et profonds) comme les habitats coralligènes. D'autres techniques de classification automatisée ont été développées notamment pour le plancton (basées sur la cytométrie en flux), ou sur l'extrapolation d'images sonar qui permettent aujourd'hui de cartographier les profondeurs avec une précision de l'ordre du mètre. Elles restent inappropriées pour les peuplements et habitats benthiques.

L'observation automatisée peut aussi être faite *in situ*. Elle est alors basée sur des enregistreurs, souvent spécialisés, d'un facteur abiotique capté par une cellule de mesure dont l'entretien est plus ou moins onéreux, mais jamais facile. Ces capteurs, souvent regroupés sur une station, ne sont pas utilisés à large échelle pour les habitats benthiques. L'apparition de drones permet aujourd'hui de couvrir de plus grandes surfaces, mais ils restent rarement utilisés pour les études du benthos car leur localisation en cours de mesure reste très imprécise et leur pilotage près des substrats, aléatoire. Il s'agit probablement de

technologies prometteuses, à condition de développer ensuite le moyen d'analyser les données brutes produites (photo, vidéo, son, autres capteurs…)

2.3 La nécessaire utilisation de données d'interprétation

Les observations du benthos automatisables sont basées de plus en plus sur la vidéo / la photo. Néanmoins, compte-tenu de la complexité des écosystèmes observés et du nombre de taxons encore à découvrir, de la difficulté de localiser précisément les images enregistrées, et du manque de référentiels et d'expert pertinent pour améliorer la reconnaissance des taxons *via* des photographies du benthos, les systèmes de classification sont aujourd'hui peu opérationnels. Une grande partie de l'acquisition de connaissances se fait donc en plongée, et le socle de connaissances de la vie sous-marine reste encore fortement dépendant de l'enrichissement des collections prélevées *in situ*.

Le développement de la plongée comme moyen d'observation a posé un certain nombre de questionnements aux communautés de systématiciens et d'écologues concernant la précision des relevés basés notamment sur des images vidéo. Certains articles comme celui cité ci-dessous ont été co-signés par un grand nombre d'auteurs (*crowdsourcing*[24]). Ces textes montrent la méfiance des naturalistes spécialistes concernant la description d'espèces sur photos, en se passant de la mise en collection académiquement reconnue :

"The question whether taxonomic descriptions naming new animal species without type specimen(s) deposited in collections should be accepted for publication by scientific journals and allowed by the Code has already been discussed in Zootaxa (Dubois & Nemésio 2007; Donegan 2008, 2009; Nemésio 2009a–b; Dubois 2009; Gentile & Snell 2009; Minelli 2009; Cianferoni & Bartolozzi 2016; Amorim et al. 2016). Photography-based taxonomy is inadequate, unnecessary, and potentially harmful for biological sciences
Article in Zootaxa 4196 (3) : 435-445 · November 2016,
DOI : 10.11646/zootaxa.4196.3.9"

Ce texte en "*crowdsourcing*" appelle donc à la prudence concernant l'utilisation unique ou même prioritaire de la photographie pour des études taxonomiques. Néanmoins, le plus gros des observations, notamment dans le milieu marin, est basé sur des photos (notamment pour les milieux profonds ou peu accessibles comme les habitats coralligènes). Pour exploiter les images de plus en plus nombreuses, il est nécessaire d'ajouter aux descriptions de collections les caractères extérieurs possiblement visibles sur photos. Ces éléments permettent de déterminer un spécimen "photographié" dans les descriptions taxonomiques, en précisant à quel niveau taxonomique cette détermination est faisable, à

[24] on appelle crowdsourcing, le fait de rédiger un document en laissant au plus grand nombre la possibilité de participer. Wikipédia est un exemple très connu de crowdsourcing.

une date donnée (sachant que la découverte d'espèces cryptiques[25] peut révoquer cette détermination). Un référentiel de ce genre de caractères permettrait par exemple de connaître quelles erreurs sont possibles ou probables en fonction de la région (en tenant compte du possible déplacement des taxons d'une région à une autre) et ainsi de valider ou invalider certaines données. La majorité des observations concernant ces milieux étant basée sur des photos, ces données, que l'on peut appeler données issues d'interprétations de photos / vidéos, restent actuellement irremplaçables concernant l'étude des substrats benthiques durs en milieu côtier. L'enjeu est aujourd'hui de pouvoir s'appuyer sur les systématiciens et les taxonomistes dont l'intérêt est de décrire de manière la plus complète possible soit un phylum soit une faune liée à un habitat ou un site, pour obtenir le savoir nécessaire pour constituer des listes de taxons utilisables (c'est à dire reconnaissables) avec le moins d'erreurs possibles sur les supports audio, vidéo ou photo. Les bases de connaissances ainsi constituées seront des matériaux précieux pour développer les nouvelles méthodes de reconnaissance automatique (*machine learning* et *deep learning* sont des exemples utilisés dans de nombreuses disciplines).

2.4 L'échantillonnage pour l'observation basée sur des études « moléculaire », de nouvelles méthodes de suivi ?

En complément de ces approches vidéo et photo, des techniques basées sur des approches moléculaires voient le jour. Celles-ci sont basées sur l'exploitation de prélèvements d'eau, de substrats, de grattages ou de prélèvements de taxons choisis, prélèvements généralement réalisés pour les substrats durs en milieu côtier par des plongeurs (et parfois des ROV, mais leur mise en œuvre est encore plus complexe). Les analyses produites peuvent être génomiques, protéomiques, métabolomiques ou concerner plusieurs de ces aspects à la fois.

Aujourd'hui, la question de l'équivalence, voire même de la complémentarité de ces différents types de suivi doit être posée, et ceci surtout avant toute mise en oeuvre sur le long terme et à large échelle, car lors de leurs déploiements, la prise en compte des coûts de mise en œuvre devient cruciale.

Ces approches sont testées dans le cadre des deux programmes CIGESMED (analyse phylogéographique pour l'algue coralline *Lithophyllum sp*. et le bryozoaire[26] *Myriapora truncata*, analyses métagénomiques pour des prélèvements de grattages de substrats) et

[25] espèces définies comme telle car isolées reproductivement et/ou dont la lignée génétique a une importante différenciation génétique, indiquant une divergence ancienne entre l'une et l'autre, mais qui n'est pas distinguable d'un point de vue morphologique.

[26] étymologiquement "animaux mousses". Les bryozoaires, sont des animaux coloniaux et sessiles. Ces organismes sont en majorité marins.

DEVOTES (analyses métagénomiques de la partie vivante fixée aux ARMS). Utiliser des échantillonnages faits par des plongeurs a pour avantage de pouvoir relever des paramètres environnementaux pour “contextualiser” le prélèvement et éventuellement identifier des espèces cryptiques (notamment *Lithophyllum sp*. dont le statut décrit par Athanasiadis en 1999 pourrait changer) ou des préférendums de certaines communautés pour des conditions particulières.

2.5 Analyses et approches comparatives / intégratives

L’identification de l’influence de ces contextes est un des objectifs de ce travail. Celle-ci est basée sur une proposition de typologie assez simple pour être relevable en plongée en même temps que les échantillonnages. Cette “contextualisation” (c’est à dire une description de tous les facteurs de contexte que l’on pense pouvoir influencer la composition des peuplements) soit au moins par un booléen, soit par un classement dans des catégories) doit ainsi permettre de comparer, dans des conditions les plus proches possibles, la pertinence et la puissance des suivis basés sur des approches moléculaires et celles des suivis présentés dans cette thèse, basés sur des analyses photographiques. Chacune de ces approches aura des avantages et des inconvénients (ou des rapports coûts avantages), dépendants de ces contextes locaux (présentés dans la partie “cartographie”), des environnements naturels et humains (testés notamment avec les récifs artificiels ARMS), des régions où ils sont mis en application, et des moyens d’observation mis en œuvre (matériel, observateurs, opérateurs, formation). L’influence de ces facteurs de contextes est abordée dans le cadre de ce travail en utilisant comme modèle les ARMS d’une part et les habitats coralligènes d’autre part.

2.6 L’unité taxonomique, chaînon nécessaire pour la compréhension systémique du niveau d’intégration du moléculaire au paysage

Une telle approche comparative puis intégrative n’est possible qu’en développant dans chaque système d’analyse (photographique ou moléculaire) un cadre taxonomique qui permet la comparaison des méthodes et ceci à large échelle. Différentes approches doivent être testées, aussi bien d’un point de vue quantitatif (fréquences relatives d’individus entre différentes communautés et selon les contextes) que qualitatif (nombre de taxons distincts détectés par contexte ou par groupe taxonomique) pour connaître le potentiel de chaque système d’observation.

3. Concepts généraux et questionnements concernant les systèmes d'information à large échelle sur la biodiversité

Il existe différents types de données en écologie utilisées et réutilisables *a posteriori* par une entité autre que le producteur de la donnée. On peut les classer en trois catégories.
Les données les plus anciennes sont issues notamment des collections et herbiers réalisés dans le cadre des sociétés naturalistes (et les plus anciennes proviennent de célèbres naturalistes comme Cuvier, Lamarck, etc.) Mais en fin de compte, les données historiques (*i.e.* antérieures aux années 50) ne sont pas très abondantes en zone côtière, et encore moins abondantes concernant le substrat dur.
Entre les années 50 et 90, l'essentiel de la donnée est une donnée d'inventaire essentiellement qualitative (on liste les taxons présents) et parfois quantitative si l'étude a été faite en milieu sédimentaire. Depuis, les réseaux de suivi en milieu côtier méditerranéen se sont essentiellement développés dans le cadre de la D.C.E., et très peu concernent encore le substrat dur (voir focus sur les réseaux de suivi, page 37). Celles-ci sont utilisables depuis 70 ans pour des données d'écologie marine issues de systèmes d'observation, rarement depuis plus longtemps sur les herbiers réalisés en laboratoire avec date, saison et lieu de récolte.
La catégorie de données la plus récente nécessite la récolte de prélèvements (pour des analyses dans des domaines tels que la chimie, l'écologie génétique, la protéomique, la métabolomique…) Ces données sont structurées majoritairement de la même manière dans chaque domaine (la codification d'un gène avec les nucléotides, les atomes des molécules, trois bases pour un acide aminé, etc.) et cette structure est partagée par leurs spécialistes.

3.1 Contexte : des données [de plus en plus] hétérogènes et multi sources

Dans un contexte de production de données en écologie à partir de multiples sources, l'équivalence des systèmes d'observation et la mise en place de systèmes d'inter-calibration deviennent cruciaux. De plus en plus, des approches transdisciplinaires intégratives deviennent nécessaires dans l'étude des systèmes où l'information dans chaque discipline est de qualité inégale et mal distribuée. Cependant, toutes les variables (pressions biotiques, abiotiques, anthropiques et naturelles, services écosystémiques perçus et reconnus, perception de la société, etc.) interagissent dans un large éventail d'échelles spatio-temporelles (Gachet *et al.,* 2005 ; Conruyt *et al.*, 2010 ; Féral *et al.*, 2014). Certaines recherches (Laporte *et al.*, 2014) ont tenté de mettre en évidence des interdépendances logiques dans les systèmes socio-écologiques pour faciliter la compréhension du rôle relatif

des services rendus par la biodiversité et les écosystèmes. Comme le montre la Figure 3, la complexité des prismes d'observation de la biodiversité et l'imbrication des différentes disciplines rend les approches multidisciplinaires très difficiles.

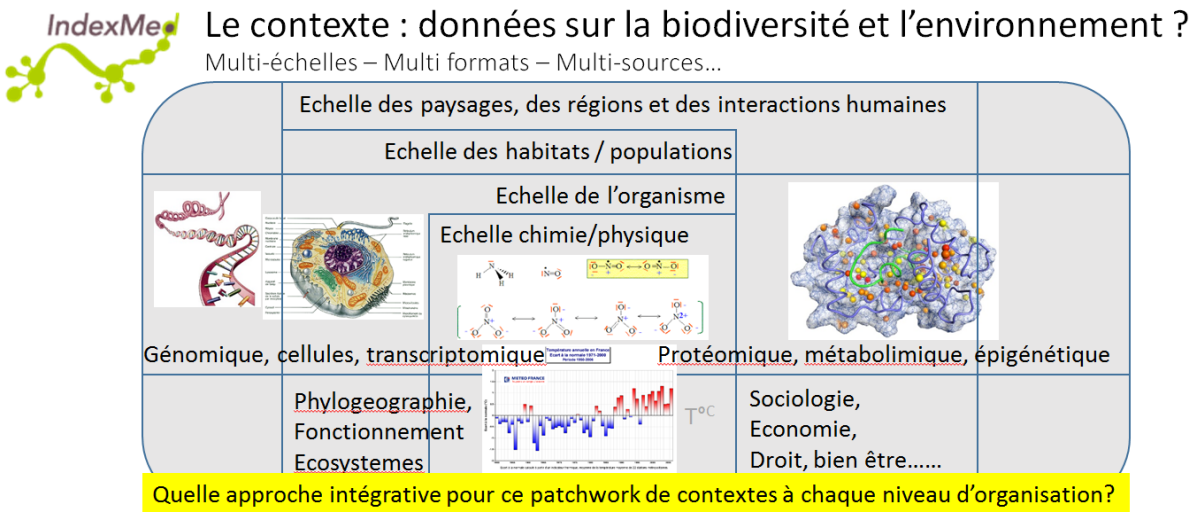


Figure 3 : Les prismes d'observation de la biodiversité sont multiples, du moléculaire à l'échelle des paysages, et aujourd'hui intègrent les relations homme-nature. Chaque facteur est complexe, difficile à mesurer, et interagit avec les autres d'une manière différente en fonction du contexte dans lequel il est mesuré (effets antagonistes, « potentialisateurs », ou de cascades). Les données sur la biodiversité (présence d'espèces, abondances, abondances relatives, biomasses …, code barre, gènes …, valeur économique, patrimoniale …) sont par ailleurs souvent coûteuses à produire, très diversifiées (format, grain géographique et temporel) peu comparables et peu automatisées.

Plusieurs auteurs et initiatives internationales ont également essayé de préciser, à travers une approche hiérarchique de la biodiversité (Noss, 1990), un ensemble minimum commun de variables à mesurer (Essential Biodiversity Variables), complémentaires les unes des autres et couvrant les niveaux interconnectés d'organisation de la biodiversité. Ils devraient permettre de saisir, avec les moyens et outils actuels, l'information maximale possible sur l'état et les tendances de la biodiversité avec le moindre effort (Pereira et al., 2013 ; Kissling et al., 2015). Des initiatives similaires sont en cours pour le climat, les conditions météorologiques et les océans [Connecting GEO[27]] afin de favoriser la découverte et l'analyse de données complémentaires à travers les échelles spatiales et temporelles. Néanmoins, pour chacune de ces variables et à chaque échelle, de nombreux formats et protocoles de mesure coexistent et ceci pour chacun des métiers et spécialités en relation avec le domaine de la biodiversité (Gestion, Chasse, Pêche, Police, Recherche, etc.)

[27] http://www.earthobservations.org/geo_community.php Le GEO (Group of Earth Observation) est un partenariat de plus de 100 gouvernements nationaux et de plus de 100 organisations participantes qui envisagent un avenir où les décisions et les actions au profit de l'humanité sont éclairées par des systèmes d'observation précis de la Terre, coordonnés sur le long terme et à large échelle.

3.2 Le Big Data pour la biodiversité ?

De nombreux acteurs produisent des données sur la biodiversité. Même en considérant uniquement la recherche, les nombreux prismes d'observation se déclinent sur des thématiques toujours plus diversifiées, déclinées par groupes taxonomiques et spécificités régionales (Figure 4).

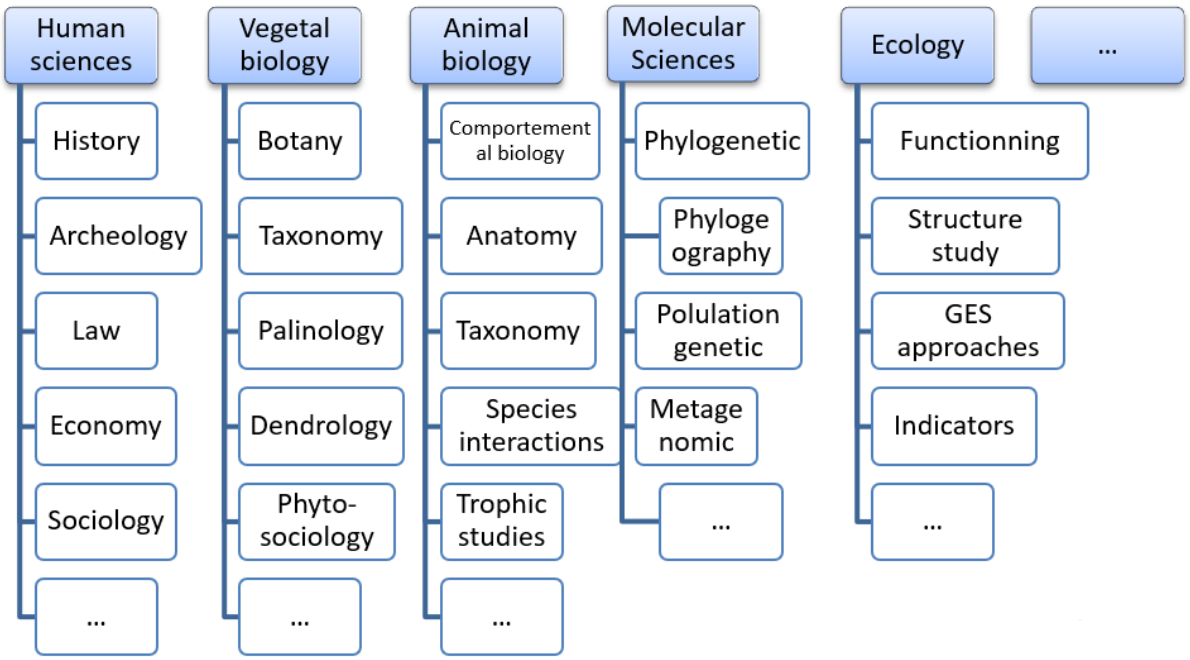


Figure 4 : Organisation en "silo" des systèmes de gestion de données sur la biodiversité : cette spécialisation excessive des recherches dans le domaine de la biodiversité se décline aussi selon les régions et les types d'habitats. Il en résulte des systèmes "mono-échelles, mono-disciplinaires et peu connectés" (Communication de R. David au TDWG, 2016).

Ces données représentent, pour chaque jeu de données, des fichiers de centaines de kilooctets pour des fichiers textes et tableurs, et jusqu'à plusieurs centaines de giga-octets pour des dispositifs de suivi vidéo par exemple. Les bases de données issues de télédétections peuvent aller au-delà. Pour autant, le volume des données généré est bien moins important que dans d'autres domaines comme l'astronomie, ou l'astrophysique (Figure 5), et chaque domaine a sa propre définition du Big Data. Le challenge du Big Data dans le domaine de la biodiversité porterait plus sur la gestion de la diversité de la donnée que son volume, du moins en l'état actuel des choses.

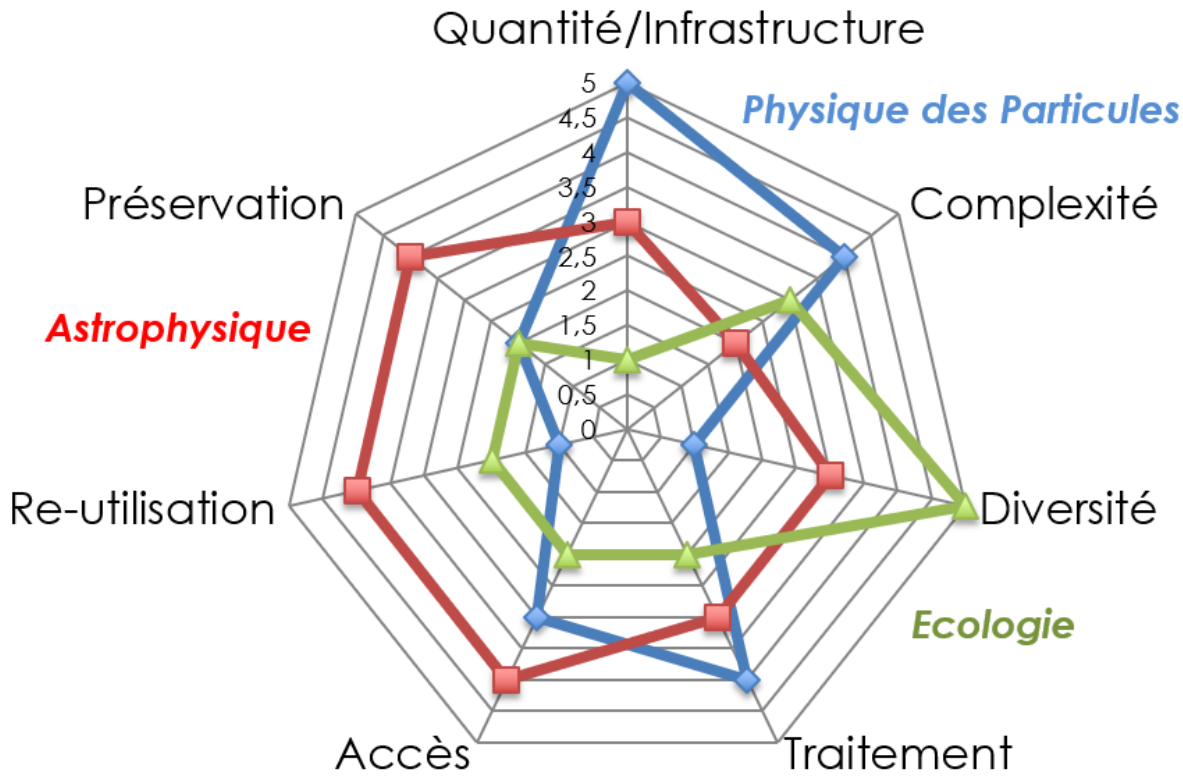


Figure 5 : Comparaison du niveau de difficulté concernant les composantes de la gestion des données dans différents domaines. Dans le domaine de l'écologie particulièrement, c'est dans la gestion de la diversité et de la complexité des données que les premiers défis sont à relever (source Predon[28], moyenne de notes attribuées sur 5, à dire d'expert)

3.3 La quête de l'interopérabilité

Contexte : des besoins d'agrégation dans un système très hétérogène

Avec le développement d'Internet, de systèmes d'information à bas coût et l'avènement de systèmes de publication techniquement plus accessibles (les C.M.S. ou Content Management Systems), la saisie et la diffusion de connaissances naturalistes se développent de plus en plus vite. Néanmoins, on observe qu'entre communautés de recherche et au sein de chacune d'entre elles, les éléments de langage employés diffèrent.

[28] PREDON : groupe s'intéressant à la préservation des données au CNRS animant une action au sein du G.D.R. MaDICS https://www.cppm.in2p3.fr/~diaconu/concrete5.7.5.9/

Depuis 15 ans, pratiquement chaque producteur de données a développé ses solutions de saisies ou d'accès aux données. Les problématiques d'agrégation de ces connaissances deviennent aujourd'hui un enjeu pour avoir une vision globale des liens entre l'activité de l'homme et les variations des différents aspects de la biodiversité, et ceci à toutes les échelles de temps et d'espace.

Le millefeuille des acteurs est aujourd'hui un frein à l'interopération, chaque utilisateur et producteur de données ayant adopté un système qui répond à ses propres besoins, avec ses propres syntaxes, structures et éléments sémantiques, qui relèvent souvent plus d'un "secteur métier" et de son jargon que de véritables concepts établis de manière partagée par toute une communauté. Pourtant, Il est essentiel d'utiliser un vocabulaire contrôlé pour limiter le risque d'erreurs engendrées par la polysémie[29] ou l'équivalence de deux termes dans les données agrégées.

Basées sur ce constat, les premières tentatives d'uniformisation des jargons *via* des interfaces ont vu le jour. Ces interfaces, parmi lesquelles nous pouvons citer le "Thésauform" en France, permettent la co-construction de vocabulaires contrôlés dans le cadre de microthésaurus (Laporte et Garnier, 2012). L'usage de ces outils peine encore à se démocratiser, et les moyens à mettre en œuvre pour les développer font encore cruellement défaut. Sur le plan européen, dans le domaine de l'agriculture, il existe un portail[30] répertoriant les différentes ontologies, et permettant de les parcourir, de les annoter, de faire des comparaisons entre elles (c'est une démarche pilotée par des chercheurs français de l'INRA au travers du programme AnaEE : Analyses et Expérimentations pour les Ecosystèmes). Une démarche allant dans ce sens est en cours pour l'écologie, mais elle n'est suivie que par une très petite part des acteurs œuvrant dans le domaine de la biodiversité. A cela s'ajoute, peut-être notamment à cause de l'utilisation de langues différentes, une évolution en parallèle des différents référentiels, thésaurus et standards sur les plans nationaux, européen ou international. Elle complique leur synchronisation (référentiels d'espèces, définitions de thésaurus avec par exemple les synonymies, les polysémies, ou plusieurs définitions pour un même terme ou un même concept dans les ontologies pour des "secteurs métiers" parfois très voisins).

[29] se dit d'un terme lorsqu'il peut prendre plusieurs sens selon les régions, disciplines ou matières considérées. Il s'agit d'un des freins principaux aux études transdisciplinaires. Plusieurs sens pour un même terme.
[30] http://agroportal.lirmm.fr/

Définitions et concepts autour de l'interopérabilité

Avant de parler d'interopérabilité entre systèmes, on parle souvent de compatibilité, à savoir la capacité d'un système à interagir avec un autre. Cette relation n'est pas forcément bijective, un système « A » étant parfois capable d'accepter une information d'un système « B » sans que cela soit forcément réciproque. Cette relation ne concerne d'ailleurs souvent qu'une partie des informations d'un, voire de chacun des deux systèmes. On parle alors de compatibilité partielle.

Lorsqu'un système devient une référence dans un domaine, il finit par imposer ses formats, processus et fonctionnalités. Les autres systèmes font alors en sorte de pouvoir être compatibles avec lui. L'inconvénient est que l'évolution des systèmes « périphériques » dépendent alors totalement de l'évolution du système « maître ».

L'interopérabilité elle, repose sur la présence d'un standard ouvert, mis en place de manière commune par tous les organismes intéressés, de manière à répondre à un ensemble de besoins communs d'échange. Cette interopération ne dépend alors plus d'un système maître (Figure 6).

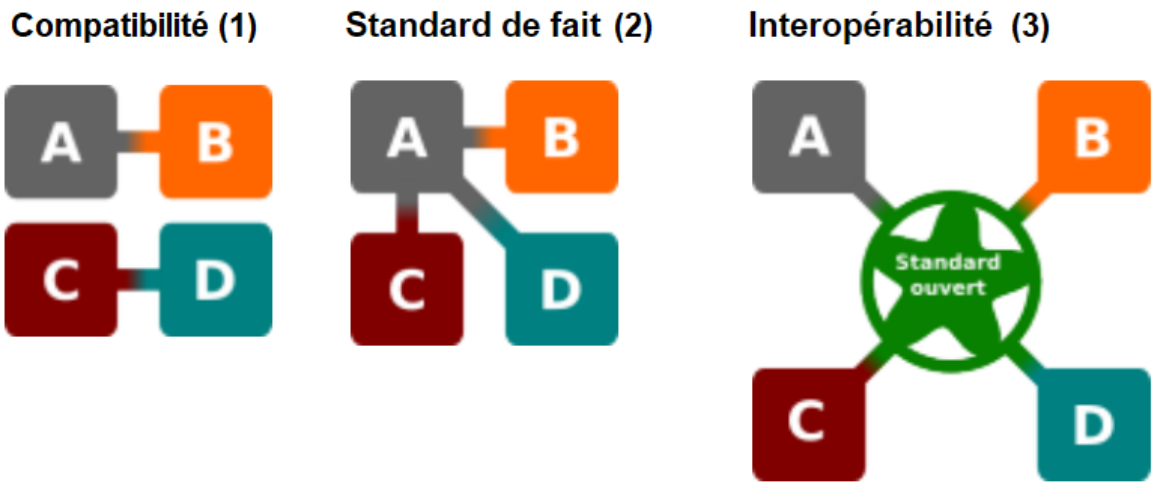


Figure 6 : Degrés d'opérabilité selon l'AFUL (Association Francophone des Utilisateurs de Logiciels Libres) : « La compatibilité (1) est la possibilité pour deux systèmes de type différent de communiquer ensemble (par exemple, le système B dépend du système maître A, le système C dépend du système maître D). Lorsqu'un acteur devient dominant dans un domaine, on parle de standard de fait (2) ; les autres acteurs font en sorte d'être compatibles avec lui (les systèmes B, C et D dépendent du système maître A). Avantage : les systèmes peuvent à peu près communiquer ensemble. Inconvénient : l'acteur dominant contrôle d'une certaine manière cette possibilité. L'interopérabilité (3) est la possibilité pour différents systèmes de communiquer entre eux sans dépendre d'un acteur particulier. Elle repose sur la présence d'un standard ouvert. »

Différentes définitions de l'interopérabilité existent. Par exemple, selon la directive INSPIRE (Infrastructure for Spatial Information in Europe)[31], on entend par «interopérabilité» la possibilité d'une combinaison de séries de données géographiques et d'une interaction des services, sans intervention manuelle répétitive, de telle façon que le résultat soit cohérent et la valeur ajoutée des séries et des services de données renforcée (Article 3, chapitre 1 de la DIRECTIVE 2007/2/CE du 14 mars 2007 établissant une infrastructure d'information géographique dans la Communauté européenne : INSPIRE). En France, la directive est

[31] La directive INSPIRE (article 3) définit une « série de données géographiques » comme « une compilation identifiable de données géographiques », une donnée géographique étant « toute donnée faisant directement ou indirectement référence à un lieu ou une zone géographique spécifique ». le terme « identifiable » signifie que la série doit avoir un sens pour ses utilisateurs potentiels. En particulier ces derniers doivent pouvoir identifier facilement, parmi les thèmes des trois annexes de la directive, celui ou ceux qui sont concernés par la série de données géographiques.

transposée de manière assez récente par l'ordonnance du 21 octobre 2010 et sa prise en compte par les différentes administrations, huit ans après, est encore très incomplète[32].

Une autre définition issue des travaux du groupe de travail « Interop » de l'AFUL (Association Francophone des Utilisateurs de Logiciels Libres) présente l'interopérabilité comme « la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs et ce sans restriction d'accès ou de mise en œuvre ».[33]
Notons que l'interopérabilité ne concerne donc pas uniquement les données, mais aussi par exemple les interfaces, les processus et algorithmes, les protocoles. Elle doit, pour éviter l'obsolescence, être envisagée dans le temps autant qu'avec des échelles géographiques différentes (ce qui équivaut à une rétrocompatibilité).

La mise en œuvre de cette interopérabilité repose donc sur l'ouverture des accès aux systèmes d'information et à leurs contenus et fonctionnalités et sur la mise en œuvre de standards.
Le problème qui émerge aujourd'hui est que des standards concurrents ont été mis en place dans de larges communautés dans des secteurs métiers très voisins. Ceci est particulièrement vrai en France où la structuration des métiers dans le domaine de la biodiversité est extrêmement complexe. La récente mise en place de l'Agence Française de la Biodiversité doit être l'occasion d'améliorer cet état de fait.

La solution pour sortir de ce jeu insoluble de concurrence passe vraisemblablement par une législation qui harmonise les pratiques et les différents standards en tenant compte de tous les besoins « métiers » (recherche, gestion, diffusion, formation…) et des besoins régaliens (police et évaluation des politiques publiques notamment). Avec les directives cadres (D.H.F.F. pour Directive Habitat Faune Flore, D.C.S.M.M. pour la Directive Cadre Stratégie pour le Milieu Marin et D.C.E. pour Directive Cadre sur l'Eau), l'Europe semble avoir pris la mesure des besoins en la matière. Le défi réside maintenant dans la mise en œuvre harmonisée de cette directive par chacun des états membres.

[32] http://eur-lex.europa.eu/legal-content/FR/ALL/?uri=CELEX:32007L0002)
[33] https://aful.org/gdt/interop

Evolution des cadrages et recommandations sur les plans français, européen et international

Après l'appel en 2006 de l'O.C.D.E. à élargir l'accès aux données de la recherche financée sur fonds publics, le C.E.R. (conseil scientifique du Conseil Européen de la Recherche) publie en 2007 des recommandations pour mettre en accès libre les résultats de recherches financées par le C.E.R. En 2010, le rapport « Riding the wave. How Europe can gain from the rising tide of scientific data »[34] reprend et complète ces recommandations puis, en 2012, la Commission Européenne affirme l'importance d'améliorer l'accès aux données de la recherche et demande aux Etats européens de définir des politiques de libre accès aux données scientifiques.

En France, le Référentiel Général d'Interopérabilité (R.G.I.) a été créé par l'article 11 de l'ordonnance n° 2005-1516 du 8 décembre 2005. Son élaboration a été conduite par la Direction Générale de la Modernisation de l'État (D.G.M.E.) à partir de 2006, puis par le Secrétariat Général pour la Modernisation de l'Action Publique (S.G.M.A.P.). Le R.G.I. décrit un ensemble de normes et bonnes pratiques communes aux administrations publiques françaises dans le domaine informatique. Sa diffusion et *a fortiori* son adoption restent néanmoins encore « confidentielles » dans les secteurs métiers en relation avec l'écologie.

Sur le plan international, de nouvelles opportunités sont créées par les formats de données ouverts (*open data*) en écologie (Reichman *et al.*, 2011) et les normes de qualification utilisables dans la gestion des données sont développées avec un consortium tel que Biodiversity Informatics Standards (anciennement Taxonomic Database Working Group)[35] (Darwin Core Task Groupe) (Wieczorek *et al.*, 2012).

Etat des lieux de l'interopération dans le domaine de la biodiversité

L'*open access*, ou accès ouvert, est une clef importante pour l'interopération de systèmes d'information. Son développement souffre pourtant de la présence simultanée de barrières financière, légale, et technique. « Les informations diffusées doivent être structurées et documentées selon les normes et les standards des disciplines scientifiques pour pouvoir être découvertes, identifiées, citées et réutilisées » (*cf.* Charte de l'INRA, Institut National de Recherche Agronomique[36], pour le libre accès aux publications et aux données)).Dans les faits, l'investissement nécessaire est reporté et beaucoup de producteurs de données

[34] High-Level Group on Scientific Data, Riding the wave How Europe can gain from the rising tide of scientific data, Final report of the High level Expert Group on Scientific Data, October 2010, url : http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204

[35] www.tdwg.org

[36] http://www.inra.fr/

croyant avoir mis en place l'*open data* ne permettent pas cette interopérabilité et donc n'appartiennent pas, de fait, au domaine de l'accès ouvert.

Il existe des systèmes d'information et des institutions organisant l'accès aux données sur la biodiversité et l'environnement. Ils travaillent sur le maintien et l'évolution des référentiels et des standards. Ceux-ci se limitent souvent à un aspect « inventaire » de la biodiversité (collecte, observations, référentiels et distribution) et négligent les aspects et liens fonctionnels entre les différents prismes d'observation. Des initiatives telles que CoL [Catalogue of Life], Data-ONE [Data Observation Network for Earth], EMODnet [European Marine Observation and Data Network], GEO-BON [Group On Earth Observations Biodiversity Observation Network], EU-BON [European Biodiversity Observation Network], GBIF [Global Biodiversity Information Facilities], LifeWatch, OBIS [Ocean Biogeographic Information System], ou le T.D.W.G. [Biodiversity Information Standards] ainsi que Darwin Core ou A.B.C.D. [Access to Biological Collections Data] sont des exemples d'outils et de réseaux bien connus et nécessaires pour organiser l'accès à la connaissance, sa normalisation et l'amélioration de l'interopérabilité des systèmes d'observation à large échelle. Cependant, les approches intégratives, particulièrement dans la zone de gestion côtière, nécessitent une meilleure interopérabilité à chaque échelle (Féral et David, 2013). La participation des différents Etats à ces organisations est toutefois très hétérogène : par exemple, l'adhésion à LifeWatch étant proportionnelle au PIB, la France, l'Allemagne et plusieurs autres pays n'en sont pas adhérents, et les réflexions sur l'exploitation de la donnée, l'organisation des systèmes d'information et le développement des standards nécessaires avancent en ordre dispersé. Par exemple, en France, depuis 2014, un format de données appelé Données Élémentaires d'Échanges (D.E.E.) a été mis en œuvre dans le cadre du S.I.N.P. (Système d'Information sur la Nature et les Paysages). Il peut être déploré que ce standard, qui ne concerne que les données d'occurrence, se soit développé en marge des démarches internationales précitées (le GBIF France a tout de même participé à son élaboration).

3.4 Entrepôts de données et accès aux données

Dans le domaine de la biodiversité, c'est le GBIF qui est le plus avancé concernant la structuration d'accès à la donnée agrégée à large échelle. A l'heure actuelle, ce système d'information regroupe plus de 850 millions d'occurrences d'espèces directement accessibles[37], et ces données sont de plus en plus réutilisées à des fins de recherche, même si le nombre d'organismes publiant" (1110) semble bien inférieur au nombre probable de producteurs de ce type de données. Le GBIF améliore le développement de cette base

[37] https://www.gbif.org

ouverte grâce notamment à des outils d'intégration de données (I. P. T. pour Integrated Publishing Toolkit) régulièrement mis à jour, et dont l'ergonomie et la simplicité ont permis le succès. Ils encouragent la complétion d'un maximum de métadonnées dans les formats et standards internationaux (E.M.L.[38] pour Ecological Metadata Language notamment) grâce à la génération de *Data Papers* depuis ces métadonnées, lorsque celles-ci sont suffisamment complètes. Ces *Data Papers* permettent de citer un jeu de données au même titre qu'une autre publication, valorisant ainsi le producteur, et pourraient devenir la norme. Jusqu'alors, ce système agrégatif ne permettait pas d'associer d'autres données que celles correspondant strictement à des observations de terrain (un taxon, un lieu, une date, un observateur) et une grande partie des informations issues de protocoles complexes ne pouvaient pas y être « entreposées ».

Récemment, le GBIF a proposé une première version d'un nouveau format d'échange appelé « sampling-event data » (Littéralement "données d'échantillonnage - événement") : https://www.gbif.org/sampling-event-data. Ces « sampling-event data » décrivent les occurrences d'espèces dans le temps et dans l'espace en les associant aux descriptions de l'effort d'échantillonnage. De telles données sont aujourd'hui produites dans le cadre des milliers d'enquêtes environnementales, écologiques et sur les ressources naturelles, mais peu ou pas accessibles (il reste parfois uniquement un rapport sous format PDF au bout de quelques années). Les « sampling-event data » peuvent être des études ponctuelles ou des programmes de surveillance sur plus long terme. Ces données sont généralement quantitatives, calibrées et respectent certains protocoles (qui sont donc documentés et rattachés à ce nouveau format) afin que les changements et les tendances des populations puissent être détectés en reproduisant le protocole. Cette innovation obtenue par consensus est une avancée majeure par rapport aux données d'observation et de collecte opportunistes, qui constituent aujourd'hui la part la plus importante des données sur la biodiversité accessibles à tous. Enfin, ce format permet aussi, pour l'instant en format libre, d'y associer toute information de contexte ou de type moléculaire – génomique / métagénomique / protéomique / métabolomique. Les modalités de standardisation de ces nouveaux types d'information sont en cours de discussion et font l'objet de nombreux groupes de travail dans le cadre d'assemblées comme le T.D.W.G. Afin d'assurer une rétrocompatibilité du nouveau système, ce format intègre les prérequis du Darwin Core comme présenté par le GBIF dans la figure 7.

[38] La norme la plus populaire pour la spécification des métadonnées est le langage EML (Ecological Metadata Language), spécialement développé pour le domaine de l'écologie. Il est basé sur des travaux réalisés par la Société écologique d'Amérique (Michener *et al.*, 1997, Applications écologiques)

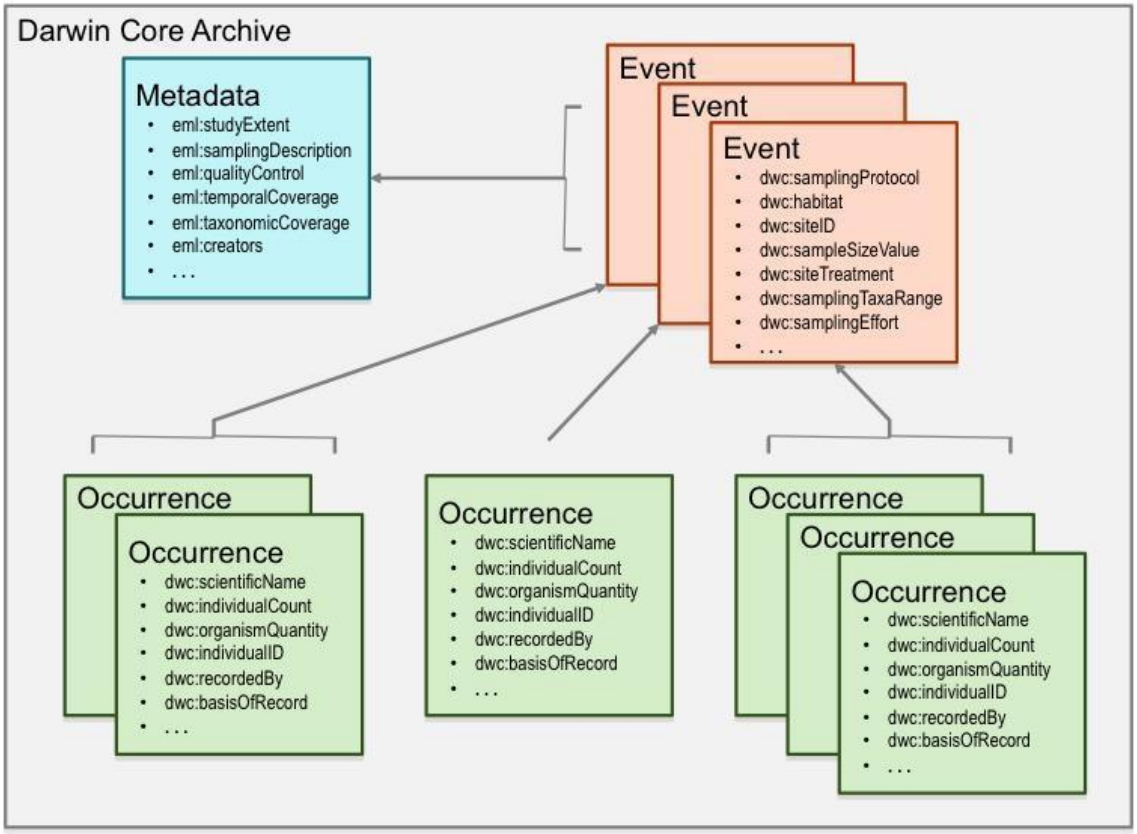


Figure 7 : présentation sur le site du GBIF d'un schéma simplifié de la structure de « sampling-event data » Le format est détaillé dans le guide des meilleures pratiques pour les « sampling-event data »[39]. Le guide de publication d'IPT répertorie un certain nombre de jeux de données d'exemples d'événements d'échantillonnage[40]

Il manque aujourd'hui le recul nécessaire pour juger de l'efficacité et de l'usage réel qui sera fait de ces nouveaux formats ; néanmoins, qu'il s'agisse d'études ponctuelles ou de programmes de surveillance, ils sont plébiscités et encouragés car les données provenant d'échantillonnages sont utilisables pour la recherche à large échelle et la gouvernance naissante au niveau mondial (I.P.B.E.S.[41]). Les « sampling-event data » permettent de rendre comparables des analyses écologiques clés : écologie des populations et métapopulations, études phénologiques, écologie communautaire et tous autres

[39] https://github.com/gbif/ipt/wiki/BestPracticesSamplingEventData#sampling-event-data.
[40] https://github.com/gbif/ipt/wiki/samplingEventData#exemplar-datasets.
[41] I.P.B.E.S. : Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (Plateforme Intergouvernementale sur la Biodiversité et les Services Écosystémiques) https://www.ipbes.net/

changements dans la structure de la communauté définis par les nouvelles gammes de “variables essentielles de la biodiversité” (E.B.V.) (Pereira et al, 2013).

La mise au point de ces outils à l’international s’appuie au sein du GBIF sur des équipes de développement et de gestion de projets structurées, suffisamment dimensionnées et expérimentées, dont les actions et les postes sont pérennisés, ce qui est un fait rare dans de nombreuses infrastructures traitant de données sur la biodiversité. Récemment en France, les effectifs de l’équipe du GBIF France ont intégré l’U.M.S. “Patrinat”[42] au sein de la nouvelle Agence de la Biodiversité. Cette avancée organisationnelle devrait aider à démontrer le besoin de moyens pérennes en terme de développement et de déploiement de méthodes et outils mis en communs à large échelle, c’est-à-dire au niveau international.

D’autres infrastructures plus spécialisées apportent aussi leur contribution à ces efforts de standardisation, mais plus le schéma structurant et typant les données est complexes, plus il est difficile d’obtenir un consensus, ce qui fait qu’une majeure partie des données biologiques est aujourd’hui conservée avec une typologie peu organisée et standardisée, ou parfois tournée vers un objectif métier unique rendant moins aisée sa réutilisation (absence ou profusion de documentation, logiciels propriétaires, algorithmes de calculs opaques …)

Parmi eux, OBIS (Ocean Biogeographic Information System) se présente comme un centre mondial d'échange de données et d'informations en libre accès sur la biodiversité marine pour la science, la conservation et le développement durable. Son système de qualification de la qualité des données et de leurs métadonnées est plus complexe que celui du GBIF, même si un corpus commun de standards de métadonnées et de données est utilisé de part et d’autre. Néanmoins, ces systèmes communiquent entre eux de manière très partielle et chaque infrastructure possède des lacunes différentes. Un autre réseau appelé EMODNET (European Marine Observation and Data Network), plus centré sur des thématiques de type océanographie physique, propose un accès notamment à des cartographies d’habitats à l’échelle européenne. Il est à noter que la précision des données proposées est variable car elle est issue pour partie de données calculées et de modèles prévisionnels. Un des problèmes majeurs de l’échange de données « agrégées » est que la qualité et donc “l’utilisabilité” de ces données restent encore peu documentées, alors que les utilisations peu rigoureuses et les agrégations douteuses se multiplient.

Tous ces réseaux échangent de manière parcellaire une partie de leurs observations, les freins principaux étant d’une part la normalisation des formats d’échange entre les parties prenantes (sans aller jusqu’à la standardisation) et le coût humain du traitement et de l’administration de ces données.

[42] http://patrinat.mnhn.fr/

Concernant l'entreposage à long terme, et pour répondre aux nouvelles conditions d'éligibilité des appels à projets nationaux et européens, des archives ouvertes comme "HAL" (Hyper Articles en Ligne[43]) permettent aux chercheurs de présenter leur *curriculum vitae* et d'archiver leur production scientifique. Elles ont été mises en place par les universités en s'appuyant sur les infrastructures proposées en France par le C.I.N.E.S. (Centre Informatique National de l'Enseignement Supérieur) ; celles-ci permettent (description officielle) :

- « D'assurer une large diffusion des résultats de la recherche.
- D'accroître la visibilité de la production scientifique des chercheurs, accessible librement.
- D'indexer cette production par la plupart des moteurs de recherche.
- D'offrir des services tels que la constitution de listes de publications.
- De garantir la pérennité des données stockées dans l'archive.
- De répondre aux exigences de la Commission Européenne dans le cadre du programme H2020 ».

L'autre fonctionnalité intéressante de HAL est de permettre l'archivage de versions de travail et donc, dans le cadre de la diffusion d'une prépublication qui pourrait être copiée, de valider l'antériorité d'un travail et de permettre un recours par le producteur spolié.

Enfin, certaines entreprises développent des « entrepôts de données » sous forme de réseaux sociaux (Academia, ResearchGate, RIO…) qui permettent de partager tout type de données et même des résultats « négatifs ». Les modèles économiques de ces structures sont souvent décriés et certains chercheurs appellent leurs collègues à les boycotter. Faut-il vraiment ignorer ces outils ? Voici la réponse laissée par un chercheur sur un forum :

« Je ne le ferai pas pour deux raisons : d'abord, nombre de chercheurs, notamment au sud de la Méditerranée où il n'existe pas d'archives ouvertes publiques (ou bien où les chercheurs ne connaissent pas ces dernières), utilisent ces services privés. C'est souvent via ces sites qu'il est possible de prendre connaissance de l'existence de ces travaux. Ensuite, même si manifestement les consultations via HAL sont plus nombreuses, ces interfaces permettent de recevoir à intervalles réguliers des demandes d'articles (non déposés, seulement signalés) via ResearchGate et, un peu moins, par Academia. HAL aurait tout intérêt à se doter d'un bouton de demande d'article, même si certains doutent de l'efficacité d'un tel dispositif qui pourrait au contraire retenir les chercheurs de déposer leurs textes en libre accès, en se limitant au dépôt de leurs notices (éventuellement le texte étant déposé mais sous embargo). Les recommandations émergentes sont d'utiliser exclusivement et prioritairement HAL pour les dépôts, et d'utiliser les réseaux sociaux scientifiques (et les autres, notamment Twitter) pour leur fonction d'aide à la diffusion, en particulier dans les

[43] https://cv.archives-ouvertes.fr/

sous-communautés disciplinaires et thématiques. Ces sites récupèrent automatiquement (notamment ResearchGate qui fait cela très bien) les nouveaux textes déposés sur HAL : il suffit de valider et d'activer ainsi le relais. »

A noter aussi que les dépôts via HAL, même s'ils demandent certaines informations obligatoires, n'imposent en fin de compte que très peu de formalisme, ce qui ne garantit en aucun cas la documentation suffisante de données pour une réutilisation potentielle. Cet état de fait est encore plus criant pour les réseaux sociaux pour lesquels les informations demandées sont minimalistes.

4. CIGESMED, premier programme cadre de cette étude

4.1 Objectif de la thèse dans le cadre de CIGESMED

Il existe actuellement différents indices aidant au suivi des habitats marins, utilisés par les réseaux de surveillance au niveau national. Pourtant, les indices, leur utilisation et la signification de leurs valeurs varient souvent d'un endroit à l'autre (Borja *et al.*, 2009). Une synthèse des impacts sur les habitats marins méditerranéens, réalisée en 2010, a conclu notamment qu'il y avait un manque de connaissance sur la distribution spatiale des habitats, et particulièrement un manque de données dans l'est de la Méditerranée. Une production de données standardisées est la condition *sine qua non* d'une évaluation correcte de l'impact anthropique sur les habitats marins (Claudet *et al.*, 2010). Le manque de connaissance ainsi pointé est d'autant plus marqué au niveau des habitats coralligènes, encore mal connu.

Dans ce contexte, le projet CIGESMED (programme franco-gréco-turc décrit dans l'annexe 3) a associé gestionnaires et scientifiques travaillant sur les habitats coralligènes, afin de concevoir et préparer un réseau d'observateurs utilisant des méthodes standardisées et inter-calibrées. Le work package 2 qui consistait à élaborer des protocoles et le work package 6 consistant à utiliser de nouvelles approches pour trier, organiser, fouiller les grands ensembles de données hétérogènes produites et développer un système d'information utilisable à différents niveaux par des scientifiques, des décideurs, des gestionnaires de l'environnement et par le grand public, ont été un des cadres de cette thèse.

Définition des habitats coralligènes

Le terme « coralligène », étymologiquement “producteur de corail”, a été utilisé pour la première fois par Marion en 1883 pour décrire les fonds durs appelés *broundo* par les pêcheurs de Marseille (Marion, 1883). Marion pensait que le corail rouge (*Corallium rubrum*) était typique de ces fonds biogènes durs. Actuellement, le terme « coralligène » fait débat, car la présence de corail rouge n'est ni obligatoire ni exclusive dans ce type d’habitat. Les habitats coralligènes à forte densité de *Corallium rubrum* ne sont que l'un des types possibles de ces habitats [RAC/SPA UNEP – MAP (2006)]. En 2006, Ballesteros recommande d'utiliser les termes "habitats coralligènes" car il en existe de nombreux types.

Dans le contexte européen actuel, les habitats coralligènes sont considérés comme des habitats “d'intérêt communautaire” (Directive Habitats 92/43 / C.E.E., code de l'habitat : 1170-14) et devraient être promus comme habitats « prioritaires ». Ces milieux sont actuellement considérés comme le deuxième « hotspot » de biodiversité marine en Méditerranée (la prairie de Posidonie serait la première selon Boudouresque, 2004), car plus de 1700 espèces utilisent ou vivent dans ces habitats (Ballesteros, 2006).

Ils sont aussi considérés comme des zones écologiques de grande valeur par la Convention de Barcelone : en 2008, cette convention a proposé un plan de gestion des habitats coralligènes. Pourtant, en France, en Europe comme dans les autres pays limitrophes de la Méditerranée, il n'existe actuellement aucun instrument réglementaire pour leur protection. De manière générale, la Directive-cadre sur la stratégie pour le milieu marin de l'Union européenne (D.C.S.M.M., en anglais Marine Strategy Framework Directive, M.S.F.D.), impose que chaque Etat élabore une stratégie et un plan d'action (P.M.A., Plan d’Action pour la Méditerranée, dans notre cas) pour atteindre et maintenir le « bon état écologique » de ses habitats marins. Ce « bon état écologique », tel que défini par la D.C.S.M.M., est évalué par onze descripteurs sur l'état et les pressions mesurées dans chaque milieu. Le premier colloque sur les habitats coralligènes a été lancé en 2009 (P.N.U.E.-MAP-CAR / SPA, 2009) pour donner suite au plan d'action pour la conservation de la végétation marine (adopté en 1999 par les parties prenantes de la Convention de Barcelone).

Depuis la publication de Marion en 1883, de nombreuses études faisant référence aux habitats coralligènes ont été publiées : Laborel (1961), Laubier (1966), Hong (1980), Sartoretto (1996) ou Ballesteros (2006). Néanmoins, les habitats coralligènes, dont on commence à percevoir la diversité des fonctions écologiques, ont un potentiel d’interaction fort avec les autres habitats p.e. les prairies de posidonies) quand ils leur

sont contigus. Sont-ils un catalyseur de la richesse en biodiversité des habitats environnants ? Pour l'instant, ces assemblages d'habitats complexes à très faible dynamique de construction sont encore trop peu documentés pour le savoir.

Dans le cadre des ateliers organisés par les participants au programme CIGESMED, une nouvelle proposition de définition tenant compte des disparités des assemblages d'espèces entre l'est et l'ouest de la Mer Méditerranée a été faite :

"Reefs in dim-light conditions mainly bio-constructed on hard substratum by calcifying coralline algae widespread throughout the Mediterranean Sea, including patchwork of habitats complicated by the action of bioeroders. These complex biogenic formations provide a number of different conditions of light, food and shelter. They are often considered as biodiversity hotspots gathering numerous sessile and sedentary species such as sponges, bryozoans, corals and gorgonians depending on the region and on the depth, to which hundreds of sciaphilic species are associated. These complex environments are a reservoir of natural resources (fisheries, red coral) and form highly valued landscapes sought by divers".[44]

"Le coralligène", un patchwork d'habitats riches et variés

Que ce soit à l'échelle méditerranéenne (Fredj et al., 1992 ; Coll et al., 2010) ou au niveau d'une mer régionale voire locale et parfois en ne considérant qu'un seul faciès (Casas-Güell E. et al., 2016), cet habitat montre une grande diversité aussi bien en terme de composition spécifique que de structure.

Les habitats coralligènes sont considérés comme des paysages sous-marins complexes, une sorte de puzzle écologique (UNEP – MAP – RAC/SPA, 2009) ayant une structure très complexe (Figure 8) et permettant le développement de plusieurs types de communautés (Laborel, 1961 ; Laubier, 1966 ; Laborel, 1987 ; Ballesteros, 2006). La grande complexité structurelle et écologique du coralligène engendre une multiplication d'unités biocénotiques (Hong, 1980 et Hong, 1982) et donc de nombreux faciès.

[44] C'est cette définition du coralligène qui est prise en compte dans ce travail. Elle n'inclut donc pas certains faciès du « pré-coralligène » décrits par plusieurs auteurs (par exemple, Pérès et Picard, 1964 ; Gili et Ros, 1985 ; Ross *et al.*, 1984 ; Gori *et al*., 2011), si ceux-ci ne comportent pas de concrétion basale d'algues corallines. Dans les faits, même dans ces faciès marqués généralement par des algues photophiles telles que *Udotea petiolata* ((Turra) Borgesen, 1926) et *Halimeda tuna* ((J. Ellis & Solander) J.V. Lamouroux, 1816), ces concrétions basales sont fréquentes.

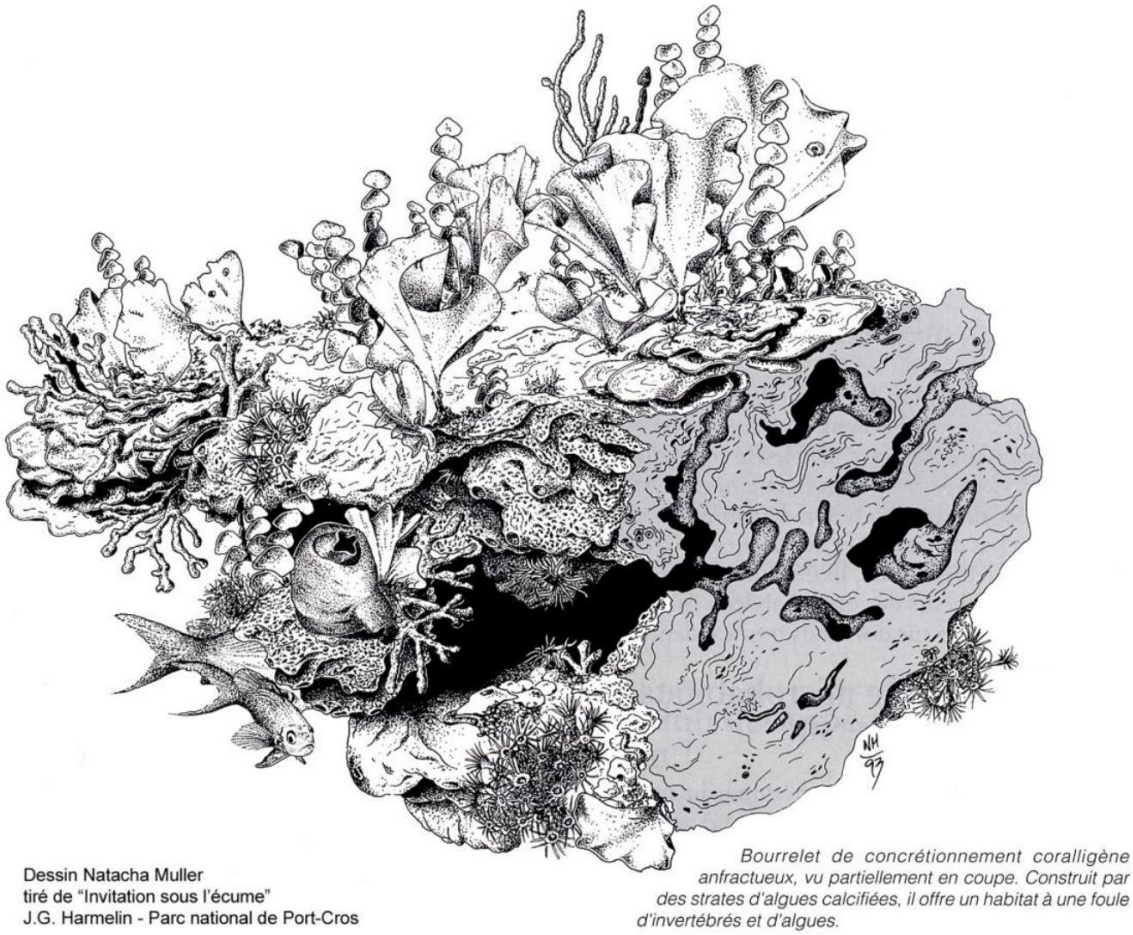


Figure 8 : Dessin d'un bloc de coralligène (d'après Laubier, 1966) représentant un bourrelet de concrétionnement coralligène anfractueux, vu partiellement en coupe. Construit par des strates d'algues calcifiées, il offre un habitat à une foule d'invertébrés et d'algues.

La description de ces faciès en considérant les différentes biocénoses[45] est conditionnée par la prédominance d'une ou de plusieurs espèces remarquables (gorgones, éponges dressées, bryozoaires). Cette prédominance - souvent sous forme d'association[46] - est favorisée par la valeur particulière et locale de certains facteurs physico-chimiques et géomorphologiques (courant, sédimentation, lumière et profondeur, rugosité, pente et orientation) dont la caractérisation est encore peu aisée, et qui peuvent changer à une échelle

[45] Groupement d'organismes vivants, liés par des relations d'interdépendance dans un biotope dont les caractéristiques dominantes sont relativement homogènes ; chaque biocénose comprend notamment la phytocénose, limitée aux végétaux, et la zoocénose, limitée aux animaux (d'après UNEP, PAM, CAR/ASP, 2006).

[46] Aspect permanent d'une biocénose avec une dominance végétale dans laquelle les espèces sont liées par une compatibilité écologique et une affinité chorologique (d'après UNEP, PAM, CAR/ASP, 2006).

métrique. La description de ces faciès selon leurs aspects bionomiques seulement reste difficile et leur distribution est très relative aux micro-conditions environnantes (Virgilio et al., 2006), essentiellement la luminosité naturelle et la courantologie, qui influencent et conditionnent fortement la création et l'emplacement de différents écotones[47] et enclaves.
La complexité structurale des habitats coralligènes favorise l'alimentation, le frai et la protection de certaines espèces de poissons (Claudet et al., 2006). La présence de support solide (algues corallines, bryozoaires…) permet le développement d'une grande biodiversité d'espèces épi-benthiques, des ressources alimentaires potentielles pour les poissons vivant sur des fonds différents (Martins et al., 2013). Les nombreux trous, anfractuosités ou fissures fournissent des abris pour les espèces benthiques, qui peuvent s'installer ou se cacher des prédateurs (e.g. crustacés : langoustes (*Palinurus elephas*), cigales (*Scyllarus arctus* et *Scyllarides latus*), homards (*Homarus gammarus*), scorpénidés (*Scorpaena* spp.), serrans (*Serranus cabrilla, Serranus scriba*) ou congres (*Conger conger*) (Humphries et al., 2011 et *cf.* convention de Barcelone, IV.3.1. - Biocénose coralligène).

Interactions biotiques au sein des habitats coralligènes

La forte complexité structurale des différents types d'habitats coralligènes et le fait que ces habitats soient souvent sous forme de patchs insérés dans d'autres milieux, multiplient les espaces de transition. Laubier en 1966, considérant cette complexité, a défini ces concrétions coralligènes comme un "carrefour écologique". Cette diversité et complexité de l'habitat induit aussi une plus grande diversité morphologique des espèces y trouvant refuge (Farré et al., 2015).
Les organismes constructeurs du coralligène et les bioérodeurs endolithes[48] sont en perpétuelle compétition. De ce fait, la vitesse de croissance de ces bio-concrétions est très lente (moins de 1 mm par an) et certaines concrétions ont mis plusieurs milliers d'années pour atteindre leur taille actuelle (Sartoretto et al, 1996).

4.2 Pourquoi les habitats coralligènes comme cas d'étude ?

L'habitat coralligène est un bioherme endémique de la Mer Méditerranée qui constitue un patrimoine biologique très fragile. La prise en compte de données et facteurs d'origine humaine est de plus en plus importante pour organiser sa gestion et permettre le maintien de son bon état écologique. Déjà, en 1983, Hong mettait en évidence un déclin du nombre d'espèces bio-constructrices en faveur de bioérodeurs (notamment l'augmentation des

[47] Transition entre deux écosystèmes ou habitats ou zones aux conditions biotiques et/ou abiotiques différentes.
[48] Espèces qui vivent enfouies dans la roche ou les concrétions coralligènes en y creusant des cavités ou en utilisant celles d'autres espèces.

éponges perforantes comme les cliones) sur des concrétions coralligènes soumises à un gradient de pollutions domestiques et industrielles.

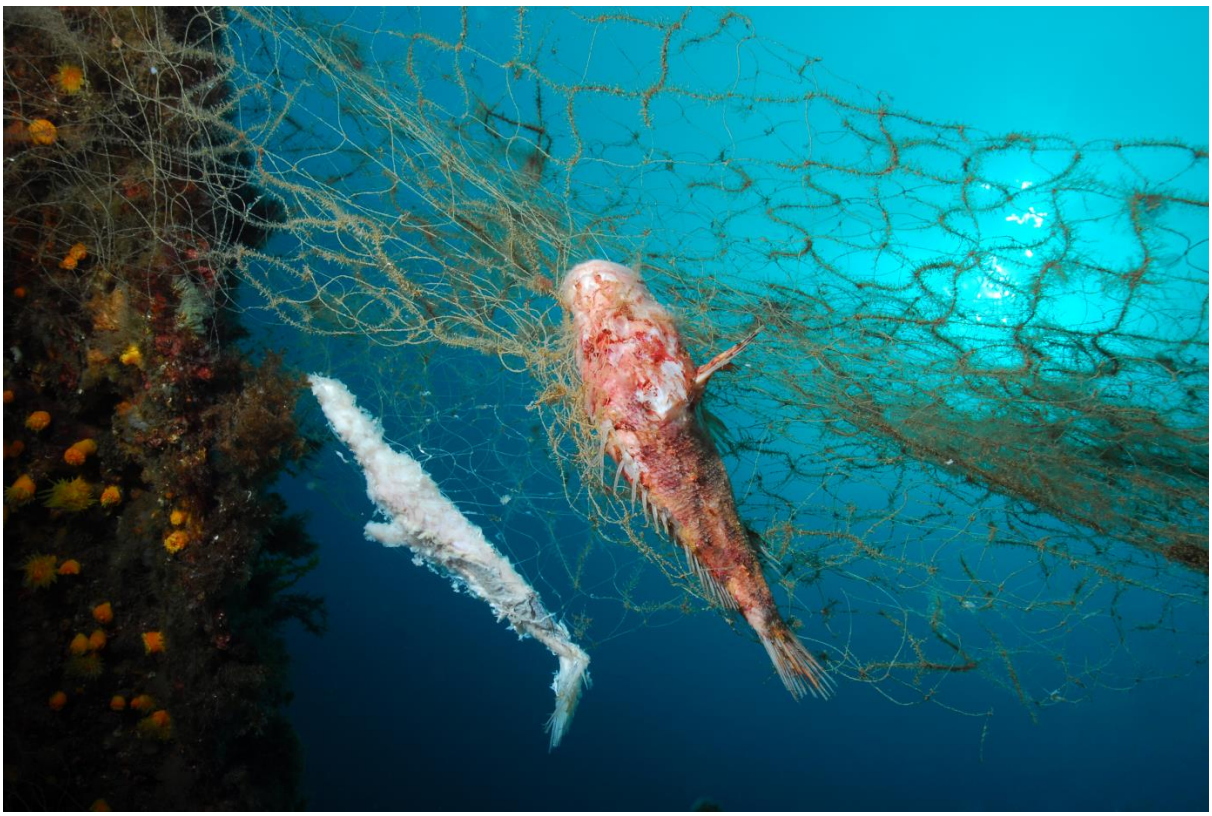


Figure 9 : filet de pêche abandonné sur une paroi de coralligène verticale. De tels engins continuent de "pêcher" plus ou moins longtemps et efficacement après leur abandon ou leur perte. © F. Zuberer

Ce milieu offre un cas particulièrement complexe de gestion des données, non seulement à cause de la richesse spécifique qu'il abrite, mais aussi par la diversité de milieux qu'il représente. Par ailleurs, la mer Méditerranée est une mer extrêmement soumise aux différentes pressions anthropiques (Figure 9) et à leurs évolutions rapides dans des contextes souvent localement très diversifiés (Coll et al, 2011). Il est admis qu'elle "sera soumise à l'horizon d'une génération à une pression de pollution d'origine anthropique de plus en plus forte, dont les conséquences seront démultipliées par les effets attendus du changement climatique" (Courteau, 2011).

Observation des habitats coralligènes, quels challenges à large échelle ?

Les habitats coralligènes sont difficiles à étudier car ils sont complexes, peu accessibles (entre 20 et 120 m de profondeur) et très variables dans les contextes locaux (Ballesteros, 2006), ce qui en fait un des écosystèmes parmi les plus diversifiés et difficiles à caractériser. En raison de ces difficultés et de la complexité intrinsèque de ce type d'habitat, les études approfondies ont été rares jusqu'aux années 2000 (Laborel, 1961 ; Laubier, 1966 ; Hong, 1982 ; Sartoretto, 1994). Depuis 2010, différentes approches visant à évaluer l'état de santé des assemblages coralligènes ont commencé à être développées (Cecchi et Piazzi, 2010 ; Deter *et al*, 2010). La plupart des protocoles ou indicateurs de surveillance proposés pour suivre sa santé écologique sont développés localement ou régionalement (Cecchi et Piazzi, 2010 ; Deter *et al*, 2010 ; Sartoretto *et al.*, 2016) sur un seul type de cet habitat (Pergent-Martini *et al.* 2014 ; Sini *et al.* 2015) et utilisent des techniques d'évaluation rapide (Bianchi *et al.*, 2007; Kipson, 2011; Gatti *et al.*, 2015) ou reposant uniquement sur la photo (Deter *et al.*, 2012b) selon les conditions environnementales. Différents faciès existants du coralligène ont été définis par la convention de Barcelone. Ils sont repris et décrits en français dans les référentiels de l'I.N.P.N.[49]. Ces définitions résistent mal à une disparité géographique très forte dans le cadre même de la définition de chaque faciès, et ce qui est considéré comme un faciès coralligène à l'est de la Méditerranée ne l'est pas forcément à l'ouest de celle-ci. Pour ces différentes raisons (typologies de cet habitat peu définies ou difficiles à appliquer dans tous les cas, données hétérogènes et multi-sources, observations sur le court terme, à faible fréquence et peu normalisée) aucun système d'observation à large échelle ouvert et offrant une information réutilisable n'a encore été proposé pour ce type de milieu. Pour comprendre l'état de cet habitat, il faut établir les bases d'une analyse intégrative, en proposant une structuration de l'observation puis de l'analyse et de la restitution de l'information produite pour que celle-ci soit compréhensible et utilisable par le plus grand nombre d'usagers potentiels (comme l'impose la convention d'Aarhus[50] dont la plupart des pays européens dont la France sont signataires).

Il faut aussi s'intéresser à tous les prismes d'observation de la biodiversité car comme l'état français le définit, reprenant les termes de la Convention de Rio (1992)[51] : « La biodiversité

[49] https://inpn.mnhn.fr/habitat/recherche/libelle/corallig%C3%A8ne/

[50] La convention d'Aarhus sur l'accès à l'information, la participation du public au processus décisionnel et l'accès à la justice en matière d'environnement, signée le 25 juin 1998 par trente-neuf États, est un accord international visant la « démocratie environnementale »

[51] https://www.cbd.int/convention/text/

est l'ensemble des milieux naturels et des formes de vie (plantes, animaux, êtres humains, champignons, bactéries, virus…) ainsi que toutes les relations et les interactions qui existent, d'une part, entre les organismes vivants eux-mêmes et, d'autre part, entre ces organismes et leurs milieux de vie. » (Source : Ministère en charge de l'Ecologie). Dans un habitat complexe où les interactions selon les contextes sont si mal connues et décrites, une amélioration de la connaissance de la structure de cet habitat, mais aussi des aspects fonctionnels entre communautés d'espèces est incontournable.

Néanmoins, la construction d'un tel système d'observation associé à son système d'information doit s'appuyer sur des moyens limités et reproductibles, demandant de fixer des objectifs réalistes. Malgré les études qui se multiplient sur ce type d'habitat tout autour de la Méditerranée, et malgré les besoins impérieux découlant de l'adoption des directives environnementales nationales et internationales (dont sur le plan européen la D.C.E., la D.C.S.M.M., D.H.F.F., et sur le plan international la convention sur la diversité biologique[52] et plus largement pour l'environnement la convention d'Aarhus[53]) aucune initiative ne tente de répondre à ce besoin de système pérenne et ouvert, utilisable localement comme à large échelle.

En expérimentant sur ce modèle complexe et peu connu d'"habitats coralligènes", ce travail tente d'apporter une contribution et un ensemble de recommandations méthodologiques qui pourront faciliter le développement de systèmes d'observations et d'informations pérennes, ouverts et plus efficaces que les systèmes actuels.

[52] https://www.cbd.int/doc/legal/cbd-fr.pdf

[53] Convention sur l'accès à l'information, la participation du public au processus décisionnel et l'accès à la justice en matière d'environnement. (Convention d'Aarhus). Celle ci est conclue à Aarhus le 25 juin 1998 et ratifiée par la France en 2002.
https://www.legifrance.gouv.fr/eli/decret/2002/9/12/MAEJ0230045D/jo/texte

5. DEVOTES, Deuxième programme cadre de cette étude

Dans le cadre du programme DEVOTES (DEVelopment Of innovative Tools for understanding marine biodiversity and assessing good Environmental Status), deux dispositifs innovants de surveillance de la biodiversité (ARMS : Autonomous Reef Monitoring System et ASU : Artificial Substrate Unit) ont été déployés dans différentes régions (Baltique, Atlantique, Méditerranée, Mer Noire et Mer Rouge).

Dans chaque région, des triplicats d'ARMS et d'ASU ont été installés dans trois sites différents. Une fois récupérés après au moins un an d'immersion, ces dispositifs ont été utilisés pour explorer la colonisation benthique sur des surfaces normalisées *via* des analyses photo. Cette mesure de la biodiversité sera aussi effectuée par séquençage génomique (en cours). Les objectifs de cette étude étaient de comparer les méthodes de suivi de la biodiversité et d'évaluer la connectivité entre les mers régionales.

Un autre objectif des programmes DEVOTES et CIGESMED auquel participe ce travail concernant l'approche photographique est de déterminer le rapport coût-bénéfice de ces dispositifs et de les comparer avec les autres méthodes traditionnelles d'évaluation de la biodiversité. Les analyses photos des faces de plaques issues des ARMS ont servi de base pour étudier les freins et proposer des recommandations concernant 2 aspects de ce travail de thèse : la construction de réseaux de suivi et d'observation pérennes et utiles pour différents types d'usages et le partage efficace des connaissances et à long terme avec ses différents utilisateurs potentiels (scientifiques, gestionnaires, élus, amateurs, grand public…) grâce à l'inter-opération des systèmes d'informations.

Pourquoi ces programmes de recherche pour construire des suivis à large échelle ?

Ces habitats ne sont pas seulement des « hotspots » de la biodiversité, ils représentent aussi de nombreux enjeux socio-économiques. Les activités telles que la pêche artisanale et la plongée sous marine dépendent fortement d'eux. Les pêcheurs y recherchent des espèces à haute valeur commerciale comme le corail rouge, les crustacés, les poissons de roche et d'autres produits de la mer. Les plongeurs recherchent la beauté des paysages offerts par les espèces colorées associées et dressées telles que les gorgones, les coraux et les bryozoaires. Au-delà de ces intérêts, d'autres services fournis par les habitats coralligènes sont discutés tels que la séquestration du CO_2 (Martin *et al*, 2013 ; Noisette, 2013) ou la stabilisation des fonds marins (Pedel *et al*, 2013), même s'ils ne seraient pas effectifs sur des échelles de temps comparables (Sartoretto et al., 1994). Ils sont menacés par les changements globaux et les pressions anthropiques. De nombreuses études montrent que l'intensification de certains usages comme la plongée ne sont pas sans conséquence sur l'intégrité du milieu (Teixidó *et al*, 2013). Du fait de sa faible vitesse de croissance, une destruction mécanique engendre tout de suite des dégâts qui peuvent être considérés comme irréversibles. D'autres perturbateurs comme certaines espèces invasives[54] (*Asparagopsis armata, Caulerpa cylindracea*, *Womersleyella setacea,* etc*.)* peuvent recouvrir entièrement les récifs coralligènes et empêcher ou au moins gêner la photosynthèse des corallines et donc la bio-construction. Selon les nouvelles directives européennes, seul le G.E.S. (Good Environmental Status) peut garantir le maintien de tous ces services fournis par les habitats coralligènes (D.C.S.M.M., D.H.F.F.).

[54] Espèce colonisant une région non connectée à sa région d'origine et dont l'aire de répartition s'agrandit rapidement et durablement (souvent par le fait de l'Homme), capable de se reproduire sans l'aide de l'homme, et qui pose des problèmes écologiques. Les phénomènes d'invasion biologique sont aujourd'hui considérés par l'ONU comme une des grandes causes de régression de la biodiversité, avec la pollution, la fragmentation écologique des écosystèmes

6. Questionnements, hypothèses et objectifs concernant l'observation à large échelle du benthos de substrat dur en milieu côtier et les systèmes d'information associés

La conservation de la biodiversité marine est une préoccupation mondiale illustrée par le programme de travail proposé par la Convention sur la Diversité Biologique (C.D.B.) visant à "promouvoir des actions politiques pour réduire la biodiversité et la dégradation des écosystèmes et des services écosystémiques, ainsi que leurs conséquences pour le bien-être humain."[55] La C.D.B. exhorte les parties prenantes "à promouvoir la production et l'utilisation d'informations scientifiques, à développer des méthodologies et des initiatives pour surveiller l'état et les tendances de la biodiversité et des services écosystémiques, partager des données, développer des indicateurs et des mesures, et effectuer des évaluations régulières et en temps opportun pour chaque usager de ces données".

Dans le cadre de ce travail de thèse, nous tentons d'apporter des solutions qui permettraient de répondre positivement et efficacement à ces recommandations en nous focalisant :

- sur la construction de réseaux de suivi et d'observation écologiques et environnementales pérennes et utiles pour différents types d'usages,
- sur le partage efficace des connaissances à long terme entre les producteurs de données et ses différents utilisateurs potentiels (scientifiques, gestionnaires, élus, O.N.G., amateurs, grand public…) et sur l'inter-opération des systèmes d'information,
- sur les méthodes, outils et interfaces d'analyses de la pléthore de données exploitant les nouvelles avancées dans le domaine du Big Data, de la gestion des données hétérogènes et de leur analyse sous forme de graphes.

En se basant sur deux dispositifs « cas d'étude » : les habitats coralligènes à l'échelle de la Méditerranée (programme CIGESMED) et la colonisation de récifs artificiels (ARMS) dans différentes mers régionales et dans le cadre des questions posées concernant l'observation à large échelle, cette thèse a pour objectif de proposer des méthodes et des protocoles, puis de tester leur applicabilité à un réseau multi-observateurs dans plusieurs pays.

Les premiers résultats de ces suivis et l'évaluation de l'efficacité des méthodes testées ont été analysés en vue de produire de premières recommandations pour construire et/ou soutenir la mise en place de réseaux de suivis utiles et pérennes de la biodiversité à l'échelle

[55] www.cbd.int/doc/decisions/cop-10/cop-10-dec-11-en.pdf La **Convention sur la diversité biologique** (**C.D.B.**) est un traité international adopté lors du sommet de la Terre à Rio de Janeiro en 1992, avec trois buts principaux : la conservation de la biodiversité ; l'utilisation durable de ses éléments ; le partage juste et équitable des avantages découlant de l'exploitation des ressources.

d'une zone biogéographique ou sur le plan international, s'appuyant sur les acteurs locaux (dispositifs allant des suivis de gènes aux suivis d'espèces et d'habitats).

- Il en résulte les hypothèses formulées dans le cadre de la thèse. L'une des hypothèses majeures de cette thèse est qu'en créant des modules avec différents types de protocoles, il est possible de favoriser les utilisations multiples et novatrices des données et leur mise à disposition pour des études ayant des périmètres thématiques, géographiques et temporels différents.
- La deuxième hypothèse consiste à poser le principe qu'il est possible de s'affranchir des technologies utilisées en construisant un système réparti, utilisant une standardisation minimale :
    - Qui permet non seulement une agrégation d'une partie des données qu'il faut identifier selon l'objectif initial du protocole,
    - Mais qui favorise aussi des usages possibles, en tenant compte des moyens potentiels de nouveaux usagers et des perspectives à court ou long terme de chacun des usages de la donnée.

Il s'agit donc de décrire les principes d'une méthode d'augmentation du "potentiel de la donnée" basée sur une architecture, des concepts d'utilisation et des services autour de cette donnée "répartie" et "hétérogène" sachant que la création de la donnée précédera nécessairement la conception de ces "nouveaux" usages. Une des conséquences sera, par exemple, que dans le cas d'une réutilisation dans un nouvel objectif scientifique, cette donnée sera produite avant l'écriture des hypothèses. La conséquence de cet état de fait est que les hypothèses scientifiques devront être bâties en fonction du contenu et de la qualité des données pré-existantes. Afin d'assurer la pertinence de ces hypothèses, il sera alors nécessaire de mettre en place des processus d'enrichissement et d'amélioration de la donnée (la "curation de données").

De plus, certains freins doivent être mis en lumière en tenant compte notamment des expériences et des leçons tirées des travaux sur les protocoles et les données produites dans le cadre de CIGESMED et de DEVOTES. Ce travail a ainsi pour objectif majeur d'identifier les verrous concernant la réutilisation des données produites, et notamment de décrire comment sur ces deux modèles d'études, il est possible de proposer des processus de traitement et de mise à disposition selon les modèles FAIR (*Findable, Accessible, Interoperable, Reusable*) et les nouveaux concepts émergeants autour du concept de *Data Management Plan* (D.M.P.) alors que l'information est produite puis stockée et utilisée dans différents pays, avec des technologies et des moyens qui diffèrent.

En particulier, des questionnements jalonnent ce travail de thèse et concernent :

- Les niveaux d'interopérabilité atteignables en tenant compte des moyens disponibles des travaux existant en la matière (ne concernant pas forcément les disciplines écologiques/environnementales) et de l'évolution des standards sur lesquels ils s'appuieront,
- La préservation de la donnée à long terme et notamment les problématiques d'obsolescence plus ou moins prévisible,
- La préservation des droits de l'auteur/inventeur des dispositifs, tout en augmentant et améliorant les différents accès aux données (brutes, traitées et de synthèse),
- La préservation de la véracité[56] de la donnée, notamment lorsque celle-ci est prétraitée par des systèmes intégratifs d'indication et d'aide à la décision.

Un deuxième objectif de cette thèse est d'apporter une part des réponses à ces questionnements : cette partie plus opérationnelle consiste donc à concevoir et tester le potentiel des données et de lever les verrous qui empêchent son amélioration grâce à la conception et au développement d'un prototype permettant de manipuler et d'agréger ces données sous forme de flux multi-formats et de représentations visuelles basées sur la théorie des graphes.

Ce prototype et les résultats de tests effectués sur les jeux de donnée des programmes CIGESMED et DEVOTES permettent de proposer de premières recommandations pour une mise en oeuvre opérationnelle et réussie d'un système d'accès multi-usagers aux données de la biodiversité marine. Dans le chapitre suivant, nous détaillons différents protocoles et méthodes de production de données produites pour répondre aux différentes hypothèses énoncées dans le cadre des programmes CIGESMED et DEVOTES, avant d'en détailler les différents résultats puis de les discuter.

[56] Le groupe d'analystes Gartner (https://blogs.gartner.com/) et IBM utilisent six lettres V fondamentales pour décrire le *Big Data* : Volume, Variété, Vélocité, Visibilité, Valeur et Véracité. La véracité est la capacité d'un grand ensemble de données contenant certaines données incertaines à donner les mêmes résultats lorsqu'il est soumis à l'analyse que le même ensemble comportant uniquement des données « certifiées ».

Chapitre 2 : Les travaux concernant les protocoles d'observation dans le cadre des programmes CIGESMED et DEVOTES

Ce chapitre utilise des éléments des publications David *et al.*, 2014a, David *et al.*, 2014b, David *et al.*, 2014c, David *et al.*, 2014d, Féral *et al.*, 2014, Féral *et al.*, 2016, Guillemain, 2014 et Thierry De Ville d'Avray, 2014, David *et al.*, 2018.

1. Questionnements et hypothèses concernant l'efficacité des outils, méthodes et protocoles

1.1 Questionnements et hypothèses concernant la mise en oeuvre du protocole "Intercalibration"

Les suivis de longues séries temporelles et/ou à larges emprises spatiales sont difficiles à mener, dès lors qu'il faut les réaliser sur une longue durée impliquant parfois plusieurs équipes d'observateurs successives et/ou dans des zones géographiques différentes. La robustesse et la reproductibilité de l'observation sont plus difficiles à obtenir, voire impossibles pour certaines mesures, et cela même si les méthodes de modélisation se développent (Gimenez *et al.*, 2014).

Dans un cadre de production de données multi-sources, dont les programmes DEVOTES et CIGESMED sont de bons exemples, la recherche de l'équivalence des systèmes d'observation et l'inter-calibration d'observateurs deviennent cruciales. Des approches intégratives, pluri- ou transdisciplinaires, sont nécessaires à l'étude de systèmes où la production de données dans chaque discipline est discontinue, plus ou moins précise et mal répartie. Pourtant, toutes les variables de ces systèmes (e.g. caractérisation des activités économiques, des installations humaines, études des productions, caractéristiques des objets reconstitués ou découverts, données biotiques et abiotiques, cartographies des pressions anthropiques et naturelles, services rendus et ressentis, image sociétale) interagissent dans le temps et à chaque échelle spatiale.

D'une manière générale, les données, lorsqu'elles sont rassemblées, sont souvent au mieux « empilées » et n'utilisent pas les mêmes standards. Les typologies de champs ne sont la

plupart du temps pas uniformisées lorsque ceux-ci contiennent le même type d'information (géographiques, temporelles, nom d'auteurs, objets, constructions humaines…) même si certains référentiels sont petit à petit institutionnalisés. *De facto*, les correspondances entre les données générées par ces études de différentes disciplines, portant cependant sur les mêmes territoires, sont encore peu aisées à réaliser, surtout sur un temps long. Améliorer le potentiel de ces données (et donc ainsi leur valeur) nécessite de mettre en place une stratégie de curation des données, d'organiser la gestion de leur cycle de vie et leur accès ("data management plan" notamment).

A fortiori, dans le cadre d'un même protocole, mesurer la variabilité induite par le choix de telle ou telle méthode ou matériel permet de mieux cerner l'efficacité de chaque combinaison expérimentale. Cette étape est la phase d'inter-calibration des méthodes. L'objectif auquel ce travail doit contribuer est d'aboutir à une base de données standardisée de facteurs et de mesures à une échelle nécessitant le recours à de nombreux observateurs formés et équipés de manière différente. Le présent travail propose une méthode, des tests et une première itération dans les développements et la mise en oeuvre des protocoles pour se rapprocher de cet objectif.

Au préalable, une phase d'inter-calibration des méthodes / matériel / opérateurs a été mise en œuvre dès le début du programme CIGESMED. Cette inter-calibration a pour objectif d'évaluer la variabilité liée aux paramètres expérimentaux, la participation relative de chaque facteur ayant une influence soupçonnée sur la variabilité des données. Elle permet de savoir dans quelles situations les résultats obtenus par différents protocoles sous-marins sont comparables, et dans quels cas les résultats doivent être interprétés avec précaution. En outre, cette phase de test permet de sélectionner le meilleur protocole à appliquer (le plus facile et le plus fiable) en fonction des types d'habitats et des moyens disponibles. Le questionnement lié à cette phase est de savoir comment isoler la variabilité naturelle inter-sites ou intra-sites, une fois pris en compte les différents facteurs inhérents au système d'observation et ayant une influence sur les résultats. Cela demande de vérifier l'influence de chaque facteur en en faisant varier, autant que cela est possible, qu'un seul à la fois : lorsque l'on change de méthode de transect, de taille de transect, cela a-t-il un impact sur les résultats de l'analyse ? Lorsque l'on modifie la qualité du matériel vidéographique ou photographique (et donc son coût), cela influence-t-il le résultat des analyses sur photos ? Quel en est l'impact sur le résultat ? Enfin, lorsque l'on change d'opérateurs (sur les photos) ou d'observateur (en plongée), quelle est l'importance de la variabilité due à ce changement (qui est inévitable lorsque le protocole est appliqué à large échelle et sur le long terme, d'autant plus que les opérateurs et observateurs ont souvent un statut précaire) ?

1.2. Questionnements et hypothèses concernant le protocole "Cartographie des profils et peuplements" (Module 1 du protocole CIGESMED)

Les cartographies des habitats marins sont en général peu fournies en données (Claudet & Fraschetti, 2010) et les incertitudes sur les cartes publiées souvent peu documentées voire non décrites (*cf.* résultats du programme Cartham : CARtographie des Habitats Marins mis en œuvre par l'Agence des Aires Marines Protégées[57]). C'est d'autant plus vrai pour le milieu coralligène qui est un assemblage complexe réparti en trois dimensions à petite échelle et dont on dispose actuellement de très peu de cartographies précises. La cartographie réalisée dans le cadre du programme CIGESMED étayera la connaissance de la distribution spatiale de l'habitat coralligène sur les sites échantillonnés.

La cartographie des sites devrait le plus possibl être basée sur la notion de faciès. Elle est encore difficile à mettre en œuvre dans un programme à visée d'aide à la gestion car les avis divergent concernant les définitions de faciès dans les habitats coralligènes (Michez *et al.,* 2011, Vassallo *et al,* 2018) et les différences entre eux peuvent être difficiles à cerner par des plongeurs de niveaux et de région d'opération différents.

Dans le cadre de CIGESMED, il a été décidé de caractériser les sites *via* un ensemble de catégories simples et très tranchées de paramètres physiques et/ou biologiques, qui peuvent regrouper et se partager différents faciès (le plongeur ne doit pas se poser trop de questions sous l'eau et son estimation doit être la plus aisée possible quel que soit son niveau). Cette caractérisation, constituée d'une description de l'inclinaison, de l'orientation, de la rugosité et des peuplements principaux pour des segments de 5 m de l'habitat, est appelé profil. Les critères proposés ont été définis après différents tests auprès de plongeurs chevronnés comme débutants, avec comme objectif de minimiser l'incertitude des descriptions. Identifier des profils typiques d'un contexte et analyser leur concomitance constitue la phase de contextualisation. L'influence de certains paramètres relevés peut être difficile à mesurer tous contextes confondus, mais se montrer plus significative dans un contexte précis. Par exemple, la cartographie nous permet d'étudier séparément les données issues du traitement des photographies selon le type de profil relevé par les plongeurs. Cette approche permet ensuite de mieux cerner la variabilité naturelle selon des catégories de contextes (i.e. variabilité inter-sites).

Les paramètres issus des métadonnées ou de l'analyse des photos peuvent aussi constituer des éléments de contexte pour l'approche moléculaire.

[57] http://cartographie.aires-marines.fr/?q=node/43

La conception et le test de ce protocole ont permis de proposer des éléments de réponse sur la précision relative de différentes méthodes de mesure de distances, sur la manière optimale de recueillir des informations sur les assemblages d'espèces présents sur chaque portion de transect, et enfin de proposer des principes de relevés simplifiés de paramètres physiques et morphologiques du substrat. L'utilisation d'une typologie commune aux différentes régions utilisable par les pays participant aux travaux de CIGESMED a été un des verrous à lever pour mieux cerner la diversité des assemblages d'espèces en fonction des profils structurels et physiques relevés par les plongeurs (profondeur, pente, rugosité, etc.). La question sous-jacente est : quels sont les paramètres et les valeurs de paramètres utiles et accessibles aux observateurs de niveaux différents qui peuvent être mis en oeuvre sur une large échelle (à l'est et à l'ouest de la Méditerranée) en minimisant les biais liés à l'hétérogénéité des milieux, et à l'hétérogénéité des observateurs en terme de connaissances et d'expérience. L'hypothèse testée pour chacun de ces paramètres est qu'en réduisant le nombre de valeur que peut prendre chaque paramètre, il est possible d'en relever un plus grand nombre, et ainsi mieux décrire les contextes correspondant aux quadrats photo.

1.3. Questionnements et hypothèses concernant la mise en oeuvre du protocole "Analyse d'images" dans le cadre de CIGESMED

Les réseaux de surveillance des habitats marins sont actuellement établis, au mieux, à l'échelle nationale, et les indices varient souvent d'un endroit à l'autre (Borja *et al.*, 2009). Pour autant, les variations observables sur une gamme d'espèces des peuplements des habitats coralligènes peuvent être relativement ténue à l'échelle de deux décennies, et les évènements portant sur d'autres taxons peuvent se traduire par des variations naturelles interannuelles importantes des variables prises en compte sans que cela ne soit le signe d'une dégradation de l'habitat (Teixidó *et al.*, 2011a).

Les protocoles récents de suivi de peuplements du benthos de type coralligène sont basés essentiellement sur des suivis, par des plongeurs, d'espèces dressées comme les gorgones rouges (*Paramuricea clavata*) (e.g. Linares *et al*., 2008, Deter *et al*., 2012a, Kipson *et al.*, 2015), et des analyses considérant séparément les strates basales et élevées (e.g. Gatti *et al*., 2015) et parfois une strate intermédiaire (Sartoretto *et al*., 2017). La plupart sont basés sur de la photographie des reliefs le long de transects[58] pré-installés par des plongeurs (e.g. Deter *et al*., 2010).

[58] Le transect est la trajectoire du plongeur sur laquelle des relevés (occurrence, abondance, estimation de la biomasse, etc.) peuvent être faits de manière systématique ou aléatoire.

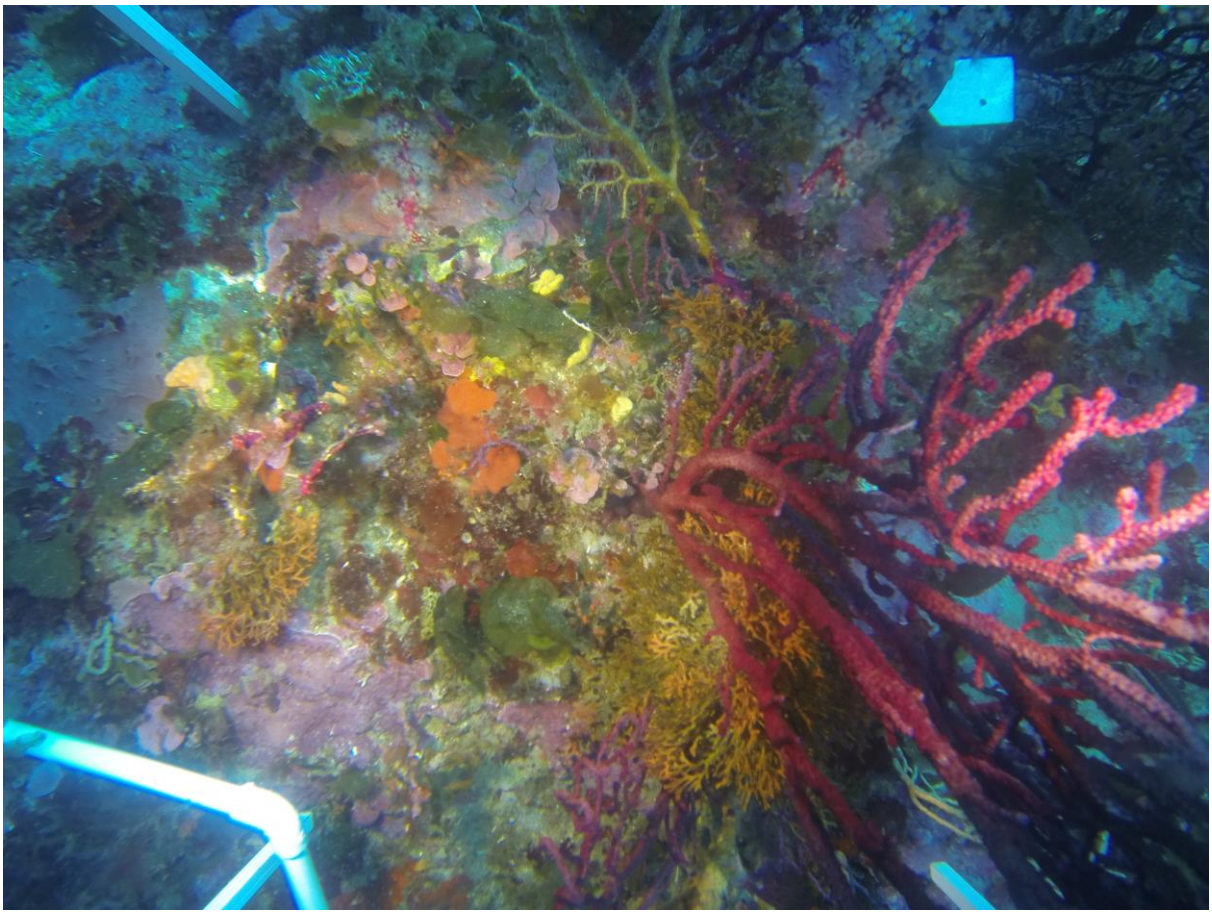


Figure 10 : Exemple de photo prise avec le quadrat-photo. Certaines espèces dressées comme ici Paramuricea clavata (à droite de l'image) peuvent couvrir une partie importante du quadrat-photo. Le cadre doit occuper le maximum de place dans l'image, et la profondeur de champs doit être suffisante pour que toute la surface du quadrat soit nette, malgré le relief et les espèces dressées. La lumière doit éclairer efficacement toute la zone délimitée par le quadrat (et sans ombre portée des espèces les plus grandes), d'où l'importance d'avoir au moins deux lampes éclairant d'un axe différent.

Ces méthodes de suivi le long d'un transect sont inspirées de suivis en récifs coralliens (depuis les années 1990, les programmes communautaires visant à étudier un plus grand nombre de récifs coralliens, tels que Reef Check (Hodgson, 2000), sont de plus en plus utilisés). Elles permettent au plongeur de réaliser une série de photographies le long du transect en glissant un cadre le long d'un cordon qui a été préalablement déroulé et qui matérialise ce transect. Le cadre délimite la zone de la photo considérée. Idéalement, et pour

optimiser la reconnaissance des taxons, ce cadre doit occuper le maximum de place dans le champ de la photographie (Figure 10).

Les photographies prises à travers ce cadre s'appellent les photos-quadrats. Ceux-ci peuvent être indépendants ou solidaires du dispositif de prise de vue (photo ou film). Ils doivent permettre de mesurer un ensemble de paramètres biotiques plus finement que les relevés de profils décrits grâce au protocole de cartographie (Module 1 du protocole CIGESMED[59]). Ces paramètres sont des mesures telles que : occurrences, abondances, taux de recouvrement, dominances, taille des peuplements et des individus, type de limites entre peuplement, fractionnement.
L'exploitation des photographies nécessite l'usage d'un logiciel d'analyse d'images permettant d'attribuer à des points ou des zones distribués sur la photographie. Il en existe différents actuellement comme PhotoQuad (Trygonis et Sini, 2012), Seascape (Teixidó et al., 2011b), CPCe (Kohler et Gill, 2006), ImageJ (une application de traitement d'images du domaine public développée par l'Institut National de santé des États-Unis[60]) ou PhotoGrid[61] pour ne citer qu'eux.

Un des objectifs du projet CIGESMED était de réaliser une étude de l'efficacité méthodologique de suivi écologique de l'habitat coralligène (qui permettront notamment le développement et le test d'indicateurs) en s'appuyant sur une caractérisation des sites.
La caractérisation des sites repose sur l'analyse de photographies réalisées sur le milieu coralligène étudié, via un logiciel de traitement d'images.

Le questionnement principal est de savoir s'il possible de caractériser un site avec un ou plusieurs des paramètres mesurables sur une série de quadrats photo. L'hypothèse à tester est que la simplicité d'une analyse portant sur les fréquences relatives de taxons identifiables sur photo (meilleur rapport coût-avantage) serait suffisante (1) pour caractériser les sites, et (2) utilisable pour détecter des modifications comme une détérioration de l'état écologique du site.

[59] http://www.cigesmed.eu/-Protocol-1-Profile-s-
[60] http://rsb.info.nih.gov/ij/
[61] http://www2.hawaii.edu/~cbird/PhotoGrid/frames.htm

1.4. Questionnements et hypothèses concernant la mise en oeuvre du protocole “Analyse d’images” dans le cadre de DEVOTES

L'évaluation de la biodiversité est un grand défi à toute échelle spatio-temporelle (Selig *et al.*, 2013 ; Borja *et al.*, 2016). Outre les directives européennes antérieures (directive-cadre sur l'eau, WFD 2000/60 / EC, DHFF), de nouvelles exigences de surveillance de la qualité de l'environnement sont apparues avec la mise en œuvre de la directive-cadre sur la stratégie marine (DCSMM 2008/56 / CE).

Afin d'uniformiser la surveillance des substrats durs benthiques, les écologues utilisent souvent des plaques ou d'autres unités d'échantillonnage artificiel d’une surface prédéfinie. Une fois que ces unités sont colonisées par des organismes marins, elles peuvent être utilisées pour surveiller ou manipuler expérimentalement les communautés benthiques (par exemple, Judge *et al.*, 1997, Bowden *et al*., 2006, Altman et Whitlach 2007, Piola et Johnston 2008, Sorte *et al*., 2010).

Pour uniformiser davantage l'échantillonnage des habitats benthiques, en particulier pour fabriquer un habitat tridimensionnel considéré proche de ceux rencontrés dans le milieu, la Division des récifs coralliens (CRED) de l'Administration nationale océanique et atmosphérique des États-Unis (N.O.A.A.) a développé des structures de surveillance autonome des récifs (ARMS : Artificial Reef Monitoring System). Les ARMS sont constitués de plaques de fixation de PVC empilées et sont conçus pour imiter la complexité structurelle des habitats des récifs coralliens (Knowlton *et al*., 2010). Les ARMS ont été utilisés pour comparer la composition de la communauté des invertébrés entre deux sites le long d'un gradient latitudinal à l'aide de métabarcodes (Leray et Knowlton 2015), pour surveiller les récifs coralliens dans les Caraïbes et l'Indo-Pacifique (Knowlton *et al*., 2010) et comparer les variabilités de paramètres via des approches morphologiques et l’étude des métabarcodes dans la Mer Rouge (Pearman *et al.*, 2016).

Depuis 2013, le projet européen DEVOTES (Développement d'outils innovants pour la compréhension de la biodiversité marine et l'évaluation du bon état environnemental) a utilisé les ARMS pour normaliser le suivi des communautés benthiques dans les substrats durs. Ce programme traite principalement du descripteur 1 de la D. C. S. M. M. (M. S. F. D. en anglais), qui concerne la diversité biologique.

Deux principaux problèmes associés à la surveillance de l'environnement marin sont le manque de temps et les coûts de ces suivis. Alors que Leray et Knowlton (2015) ont évalué les ARMS entre les régions tempérées et subtropicales, ils ont utilisé une approche de metabarcoding relativement coûteuse. Le questionnement principal est de savoir comment, grâce à une approche rapide d’analyse photographique, il est possible de détecter des

changements de conditions environnementales, et si cette sensibilité à ces changements est possible dans différentes mers et différentes conditions.

Dans cette étude, nous avons testé le potentiel d'une évaluation photographique des organismes sessiles se fixant sur les ARMS en tant qu'outil de criblage rapide de la structure des peuplements dans une gamme de conditions environnementales. À cette fin, nous avons analysé la colonisation d'un sous-ensemble de plaques après plus d'un an d'immersion dans différentes conditions environnementales dans deux mers régionales d'Europe (Atlantique Nord-Est et Méditerranée) et dans la Mer Rouge. Nous avons d'abord testé si la composition de la communauté, déduite des photographies, était significativement différente entre les surfaces de plaques distinctes de l'ARMS, entre les sites (dans chaque mer) et entre les mers. Deuxièmement, nous avons également étudié l'effet de divers facteurs environnementaux reflétant à la fois le niveau de pression anthropique et la diversité locale des habitats sur les modèles de biodiversité. Troisièmement, puisque les protocoles de surveillance doivent être aussi simples et rentables que possible, nous avons effectué des analyses en considérant chaque groupe taxonomique seul (plutôt que la composition de la communauté entière) et chaque plaque séparément pour déterminer si ces analyses plus détaillées donnaient des résultats complémentaires.

2. Méthodes d'intercalibration, de cartographie et d'analyses d'images

Les sites présentés en figure 11 ont été utilisés pour tester les différents protocoles du programme CIGESMED et sont listés dans le tableau 2. Tous les sites ont été utilisés pour l'approche photographique en *patch* (c'est à dire en plusieurs morceaux) dont la méthode est détaillée dans la partie intercalibration de ce chapitre. Parmi eux, grâce à leur proximité, le fait que ces sites soient moins exposés aux vents dominants comme le Mistral (qui peuvent contraindre à annuler une plongée) et leur typicité, trois sites ont été choisis comme sites principaux d'étude (l'îlot Tiboulen du Frioul en pleine baie de Marseille, l'ilôt de Moyade en plein coeur du Parc National des Calanques, et le Phare de Cassidaigne). Ces trois sites offrent aussi des situations contrastées en terme de fréquentation humaine, et proposent différents types d'orientation des habitats coralligènes. L'îlot Tiboulen du Frioul étant le plus proche, il a été choisi pour y installer le transect permanent. Trois sites choisis pour la possibilité d'y trouver des habitats coralligènes à faible profondeur et bien protégés des vents dominants ont servis aux entraînements et à l'inter-calibration des plongeurs (Figuier, la Grotte à Perés et le nord de l'îlot d'Elevine - aussi appelé Erevine). Tous les sites ont été utilisés pour l'étude photographique par patch, les études de certains bio-constructeurs

(*Lithophylum spp.* et *Myriapora truncata*), et des relevés cartographiques pour connaître les contextes des échantillonnages.

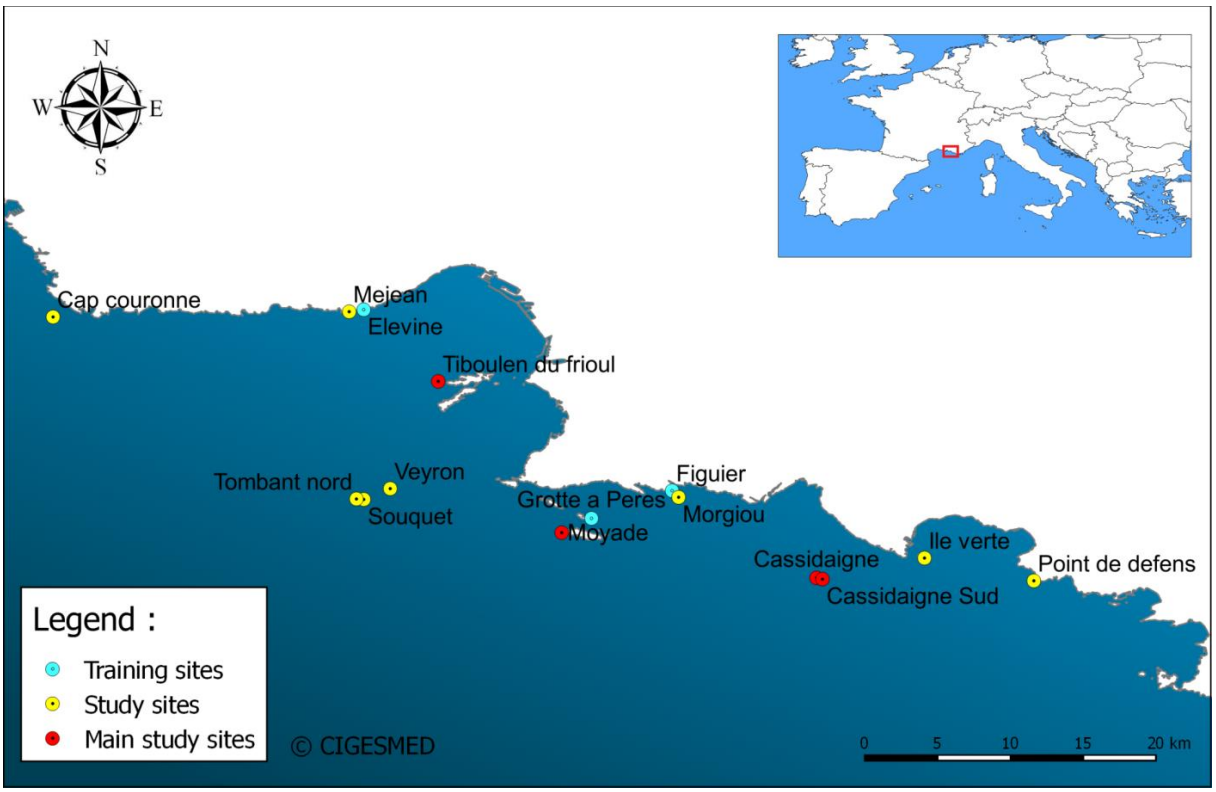


Figure 11 : L'îlot Tiboulen du Frioul en pleine baie de Marseille, l'ilôt de Moyade en plein coeur du Parc National des Calanques, et le Phare de Cassidaigne sont les trois sites d'études principaux. Figuier, la Grotte à Perés et le nord de l'îlot d'Elevine sont les sites d'entraînement. Tous les sites ont été utilisés pour l'étude photographique par patch, les études de certains bio-constructeurs (Lithophylum spp. et Myriapora truncata), et des relevés cartographiques pour connaître les contextes des échantillonnages (Tableau 2).

Tableau 2 : Sites d'échantillonnage en plongée autour de Marseille[62] utilisés pour CIGESMED dans le cadre de cette thèse (en gras les sites principaux, et pour les types de sites, C pour cartographie, E pour entrainement, P pour prélèvements contextualisés, Q pour Quadrats photo, I pour inter-calibration).

Zone	Nom du lieu	Site	Sigle	Latitude	Longitude	Type de site
Cote bleu	Cap couronne	Cap couronne	COU	43°19.550'N	5° 00.568'E	C, P,
Cote bleu	Elvine	Elvine	ELV	43° 19.780'N	5° 14.210'E	E
Cote bleu	Méjean	Méjean	MEJ	43° 19.700'N	5° 13.480'E	C, P, Q
Rade	Frioul	Tiboulen du frioul*	FTF	43° 16.820'N	5° 17.160'E	C, E**, I**, P ,Q
Rade	Planier	Tombant nord	PTN	43° 11.950'N	5° 13.780'E	C, P, Q
Rade	Planier	Souquet	PSO	43° 11.959'N	5° 14.073'E	C, P, Q
Rade	Planier	Veyron	VEY	43° 12.414'N	5° 15.176'E	C, P, Q
Les Calanques	Morgiou	Morgiou	MOR	43° 12.060'N	5° 27.100'E	C, P, Q
Les Calanques	Plane	Grotte à Peres	PGP	43° 11.190'N	5° 23.470'E	C, P, Q
Les calanques	Riou	Moyade	RMO	43° 10.600'N	5° 22.240'E	C, **,I** P, Q
Les Calanques	Riou	Riou sud	RRS	43° 10.370'N	5° 23.420'E	C, P, Q
Les Calanques	Riou	Impérial de terre	RIT	43° 10.370'N	5° 23.580'E	C, P, Q
Les Calanques	Sormiou	Figuier	SFI	43° 12.330'N	5° 26.790'E	E
À l'est des Calanques	Cassidaigne	Phare de Cassidaigne	CCA	43**°** 14.575'N	5**°** 54717°E	C, P, Q
À l'est des Calanques	La Ciotat	Ile verte	CIV	43° 09.371"N	5° 37'01.9"E	C, P, Q
À l'est des Calanques	La Ciotat	Pointe de défens	LPD	43° 08.365"N	5° 41'48.1"E	C, P, Q

[62] D'autres sites non utilisés pour cette thèse ont été échantillonnés en utilisant les protocoles de CIGESMED : https://www.imbe.fr/les-sites-de-la-baie-de-marseille.html

2.1 Méthodes d'inter-calibration

Plusieurs méthodes d'échantillonnage photographique et de traitement d'images ont été testées afin de sélectionner selon les conditions la méthode la plus efficace pour mieux appréhender la variabilité d'un certain nombre de mesures ou de facteurs contextuels. Mesurer la variabilité induite par le choix de telle ou telle méthode ou matériel dans un contexte mieux défini permet de mieux cerner l'efficacité de chaque combinaison expérimentale, à condition de pouvoir faire des comparaisons en ne modifiant qu'un seul de ces facteurs.

Les quadrats photo[63] sont des photographies délimitées par un cadre (le quadrat) dont la dimension est fixée. Ils doivent permettre de mesurer un ensemble de paramètres biotiques plus finement que les relevés de profils qui sont présentés dans le protocole de cartographie (Chapitre 2 partie 2.2). Ces paramètres sont les occurrences, les abondances, les taux de recouvrement, les fréquences relatives ou dominances, la taille ou la surface de chaque individu ou de la totalité des individus d'un taxon, le type de limites entre les taxons (par exemple, régulier ou pas), l'homogénéité ou l'hétérogénéité de la mosaïque créée par le recouvrement de différents taxons.

La variabilité de ces paramètres dépend :

- Des conditions naturelles, variant dans le temps (saisonnières, annuelles ou plus longues : les espèces saisonnières[64] doivent notamment être identifiées),
- D'une possible prédation / broutage des organismes fixés
- Des conditions physiques lors de la mesure (qui sont des variables de contexte)
- De l'observateur (pratique de la plongée, photographie sous-marine, condition et capacités physiques),
- Des conditions et du matériel utilisé pour ces observations (lampes et flash, appareils photos et objectifs, type et taille des cadres),
- Des connaissances et du niveau de pratique de l'opérateur,
- Des qualités des logiciels et techniques que celui-ci utilisera,
- Et des pressions anthropiques (qui sont aussi des variables de contexte).

Cette phase d'inter-calibration des méthodes / du matériel / des opérateurs est testée de manière itérative, permettant de corriger ou d'améliorer les process afin de les rendre moins coûteux ou plus efficaces. L'inter-calibration a été réalisée pour évaluer la variabilité due à

[63] Photographie prise à travers le quadrat, délimitant une aire précise. Unité d'échantillonnage photographique.

[64] Espèce qui ne perdure pas pendant toutes les saisons et donc dont la présence sur un site ou un transect peut être ignorée, si la fréquence d'observation pour la prendre en compte n'est pas assez forte.

chacun des paramètres expérimentaux précités, en comparant autant que possible la sensibilité de chaque paramètre (avec la difficulté de n'en faire varier qu'un seul à la fois, ce que n'est jamais exactement le cas). Il permet de savoir dans quelles situations les résultats obtenus par différents protocoles sous-marins sont comparables. En outre, cette phase de test aide à sélectionner le meilleur protocole à appliquer (le plus simple et le plus fiable) en fonction d'autres paramètres (il y a donc une efficacité contextuelle), comme par exemple les types d'habitats ou de faciès coralligènes considérés. Cette étape est nécessaire avant l'étude de la variabilité naturelle inter-sites ou intra-sites, surtout lorsque, et c'est inévitable à large échelle, les moyens matériels et humains diffèrent. À ce jour, trois variables ont été étudiées : la méthode d'échantillonnage, la qualité de la caméra et le niveau de connaissance des opérateurs chargés d'identifier les espèces.

Méthode de choix des variables mesurables et les modalités qu'elles peuvent prendre

Dans le cadre du programme CIGESMED, le choix et la définition des variables mesurables et à tester lors des protocoles ont été élaborés lors des différents séminaires qui ont jalonné son déroulement. Après avoir listé l'ensemble des descripteurs possibles, plusieurs ateliers ont été organisés à cet effet avec les partenaires des différents pays. Il a été demandé en premier lieu à chaque participant de proposer des modalités possibles pour chacun des descripteurs avec un formulaire en ligne[65], puis d'effectuer un choix entre les propositions lorsqu'il y avait divergence d'opinion. Pour chaque terme, la définition a ensuite été améliorée pour tenir compte des modalités fixées.

Méthode d'étude de la variabilité due à l'échantillonnage

Les sites étudiés sont situés dans la baie de Marseille. Ce sont des transects de 10 mètres de long à 28 mètres de profondeur. Les habitats coralligènes choisis sont des parois dominées par *Paramuricea clavata* – la gorgone rouge (Figure 12_1), comportant différentes densités, des anfractuosités de petites à larges, et une grande richesse d'espèces.
Le protocole demande aux plongeurs de faire des quadrats photo à l'aide d'un cadre de 50 cm sur 50 cm (Figure 12_2).

[65] http://www.cigesmed.eu/Bottom-up-initiative-on-a

Figure12_1 : Paramuricea clavata – la gorgone rouge © F. Zuberer

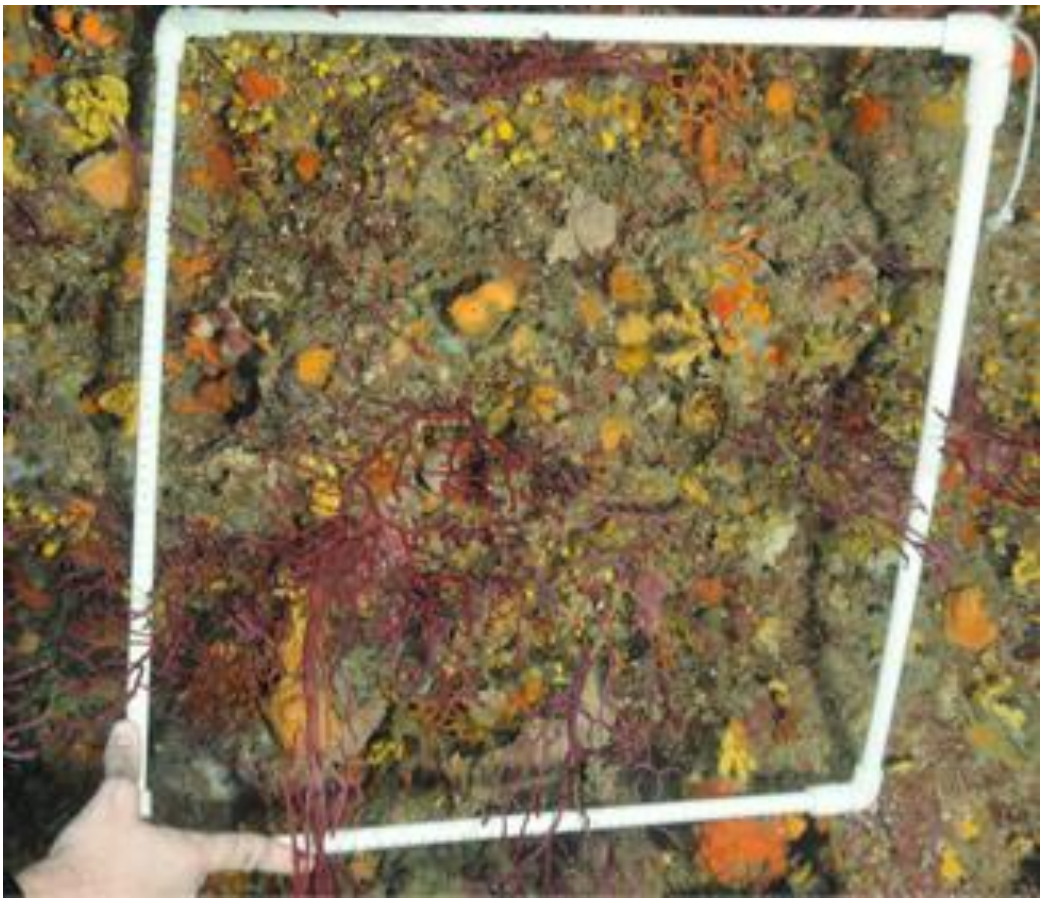


Figure 12_2 : cadre de 50 cm sur 50 cm sur une paroi coralligène verticale

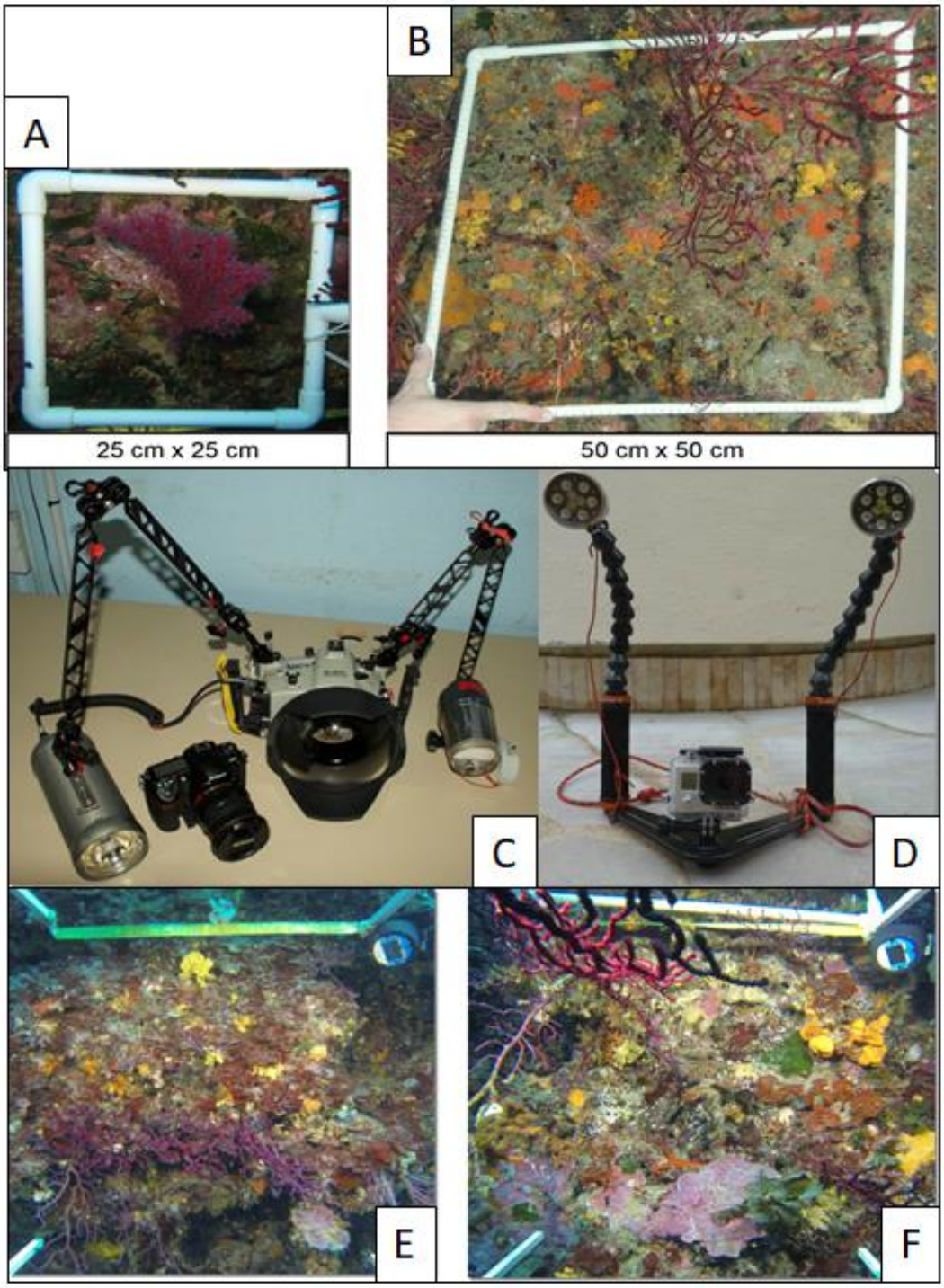


Figure 13 : Des comparaisons ont été faites avec les deux cadres 25 cm sur 25 cm (A) et 50 cm sur 50 cm (B), l'appareil photo professionnel Nikon D300s, avec l'appui de 2 phares de 600 lumens sur bras articulés (C), et la caméra "GoPro hero 3 black" deux bras articulés équipés de phares sola 600 light and motion (D). D'autres tests ont été faits avec deux intensités lumineuses différentes (50% en E et 100% en F avec la GoPro).

Les photos sont prises avec une caméra “GoPro hero 3 black” fixe sur un cadre comportant des pieds espacés de 50 cm, et deux bras articulés équipés de phares sola 600 light and motion (Lumens : Haute-600, Med-450, 300 faible).
Des tests ont aussi été effectués (voir figure 13) soit avec des quadrats plus petits de 25 cm sur 25 cm, soit avec un appareil photo professionnel Nikon D300s, avec l’appui de deux phares puis d’un seul, en lumière maximale (600 lumens) puis de plus faible intensité (300 lumens).

Deux méthodes d'échantillonnage ont été comparées : (i) transect linéaire permanent et (ii) transect de patchs aléatoires. Pour comparer les méthodes d'échantillonnage, 20 quadrats photo réalisés sur un transect permanent ont été comparés à 18 (ou 27) quadrats photo (2 patchs minimum, 3 patchs si possible) effectués avec la méthode des patchs aléatoires. La mise en œuvre de la “méthode (i)” a consisté à ce que le plongeur commence à partir d'un point permanent et réalise 20 photos-quadrats de manière continue sur le transect de 10 mètres de long, suivant une ligne horizontale virtuelle à profondeur “constante” (Figure 14).

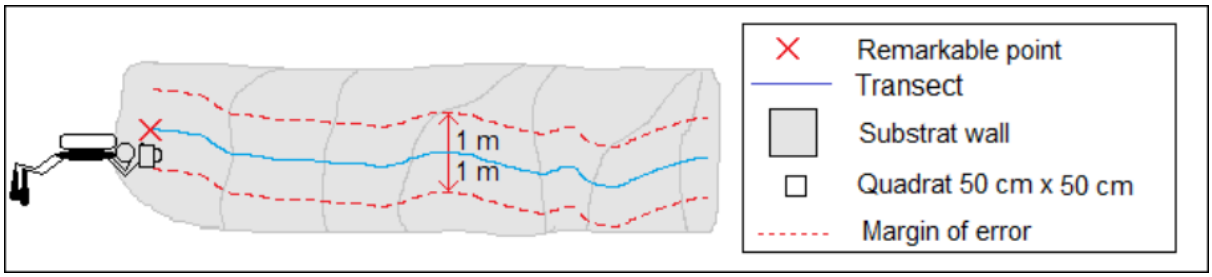


Figure 14 : marge d'erreur de trajectoire des plongeurs sur les transects linéaires

La mise en oeuvre de la “méthode (ii)” consistait à créer des patchs de 9 quadrats photo placés de façon aléatoire, à une profondeur constante. Pour faire un patch, le plongeur place un cadre marquant le centre du patch. Ensuite, il fait les quadrats photo autour du cadre en commençant par le coin inférieur gauche, et en finissant par le coin supérieur droit. Il devrait dessiner un patch de 3 par quadrats photo, comme l'illustre la figure 15.

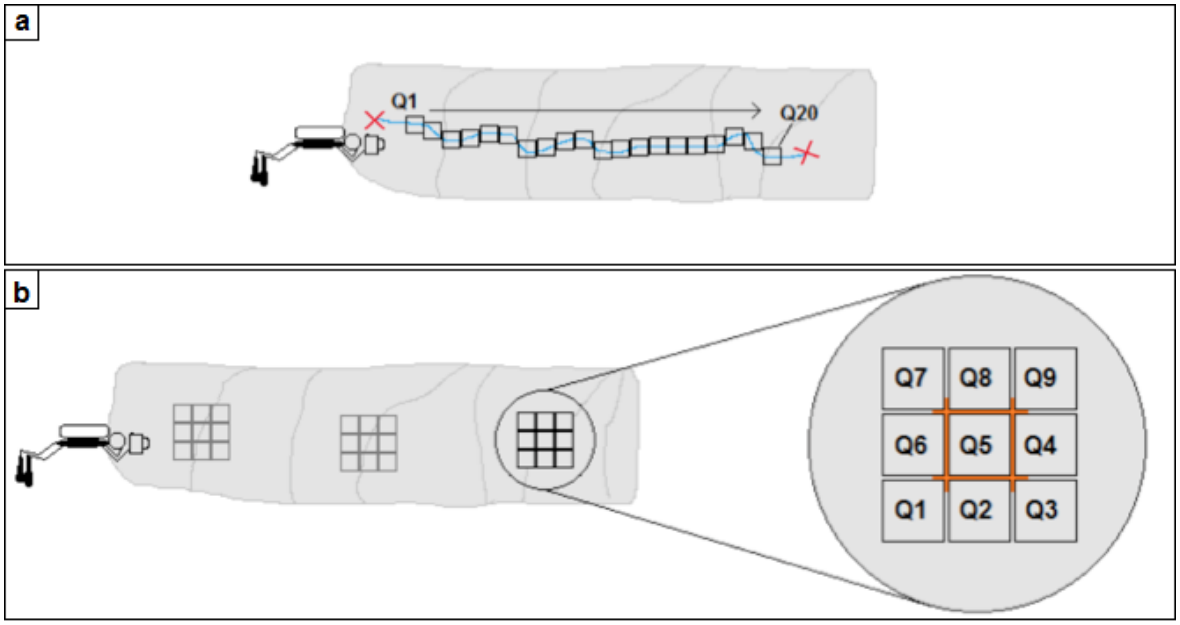


Figure 15 - Les deux types de transects testés. Transect linéaire (a) : 20 quadrats photo sont effectués à une profondeur donnée, localisés par des marques permanentes. Transect de patchs aléatoires (b) : 3 groupes de 9 quadrats photo sont pris, suivant le schéma des nombres indiqués de 1 à 9, plusieurs fois à la même profondeur.

Méthode d'étude de la variabilité due à l'observateur

La variabilité observateur a été étudiée sur plusieurs opérations en plongée. Les méthodes n'étant pas prédéfinies dans nos protocoles, nous avons défini des objectifs de mesures de différents paramètres puis testé différentes méthodes de mises en oeuvre pour chacun de ces objectifs pour estimer l'impact de la variabilité entre observateurs. Sur l'ensemble des variables choisies et définies en séminaire, un sous-ensemble a été testé par plusieurs observateurs en plongée. Les modalités de ces variables ont été adoptées lorsque, sur les mêmes transects, les observateurs donnaient une même valeur à la variable. Les contraintes de temps et la difficulté d'une mesure nécessitant un entraînement ont aussi été intégrées pour permettre une appropriation plus rapide du protocole.

Pour chacun de ces paramètres, plusieurs observateurs faisaient donc la même mesure au même endroit, afin de tenir compte de l'influence des contextes des mesures sur cette variabilité inter-opérateurs (température et donc froid pour le plongeur, profondeur et donc possible narcose, pente de la paroi et donc difficulté ou non à se stabiliser, courant et donc fatigue et moindre efficacité de l'observateur). L'entraînement et l'expérience du plongeur ont aussi été pris en compte : l'objectif était de savoir si la variabilité d'un paramètre détectée entre deux plongées et/ou deux plongeurs devenait négligeable ou restait importante avec

l'entraînement. Si celle-ci devenait négligeable, l'entraînement à recommander pouvait être pris en compte dans le choix de la méthode de mesure du paramètre.
Les paramètres testés sont les parcours de distance, les relevés de recouvrement (fréquence des taxons les plus présents), la pente du substrat, l'orientation de la paroi, la rugosité, la profondeur, la lumière, et la qualité des quadrats photographiques.
Pour chacun des paramètres, la variabilité entre observateurs était qualifiée comme meilleure ou moins bonne que pour les autres méthodes de mesure de ce paramètre (ratio d'images exploitables ou non exploitables par exemple), et a permis de choisir les méthodes à appliquer dans le protocole. Le temps nécessaire pour chaque action a également été pris en compte pour définir chaque action à mettre en œuvre dans le cadre des protocoles CIGESMED. Lorsque pour certains paramètres, des biais dus à la méthode mise en œuvre ont été découverts (estimation d'abondance mal reproduite entre plongeurs par exemple), une autre itération de tests a été effectuée.

Méthode d'étude de la variabilité due à l'opérateur

Les images ont été analysées par les opérateurs utilisant le logiciel Photoquad (Trygonis et Sini, 2012). Cent points[66] ont été distribués selon la méthode de la randomisation stratifiée (l'image est divisée en 100 carrés, puis les points sont placés aléatoirement dans chaque carré). Ensuite, l'opérateur a attribué chaque point à une catégorie et une sous-catégorie parmi ces trois : (i) taxons supérieurs (tels que le phylum), (ii) abiotiques, (iii) indéterminés. Dans la première catégorie (i), les sous-catégories sont des taxons inférieurs (comme le gène ou l'espèce). La deuxième catégorie (ii) est subdivisée en quatre sous-catégories : sédiments, roches nues, débris organiques ou débris. Dans le troisième (iii), il existe trois sous-catégories : image floue, ombre / trou et taxon non identifié.

[66] Ce nombre de points a été fixé après comparaison pour le même set de photo avec 250 points

Focus sur PhotoQuad

PhotoQuad est un logiciel d'analyse d'image qui permet de superposer un calque aussi appelé « layer » à la photo (Figure 16). Ce calque permet d'attribuer des valeurs (aussi appelés attributs) à des points, segments ou zones de la photo à la manière d'un logiciel SIG. Les zones peuvent être créées manuellement ou par différents types de détection des formes, des couleurs et des textures contenues dans la photo. Les points, dont le nombre est fixé par le protocole d'échantillonnage, sont générés et répartis soit de manière uniforme, soit de manière totalement aléatoire soit aléatoires stratifiées sur un calque placé sur une image.

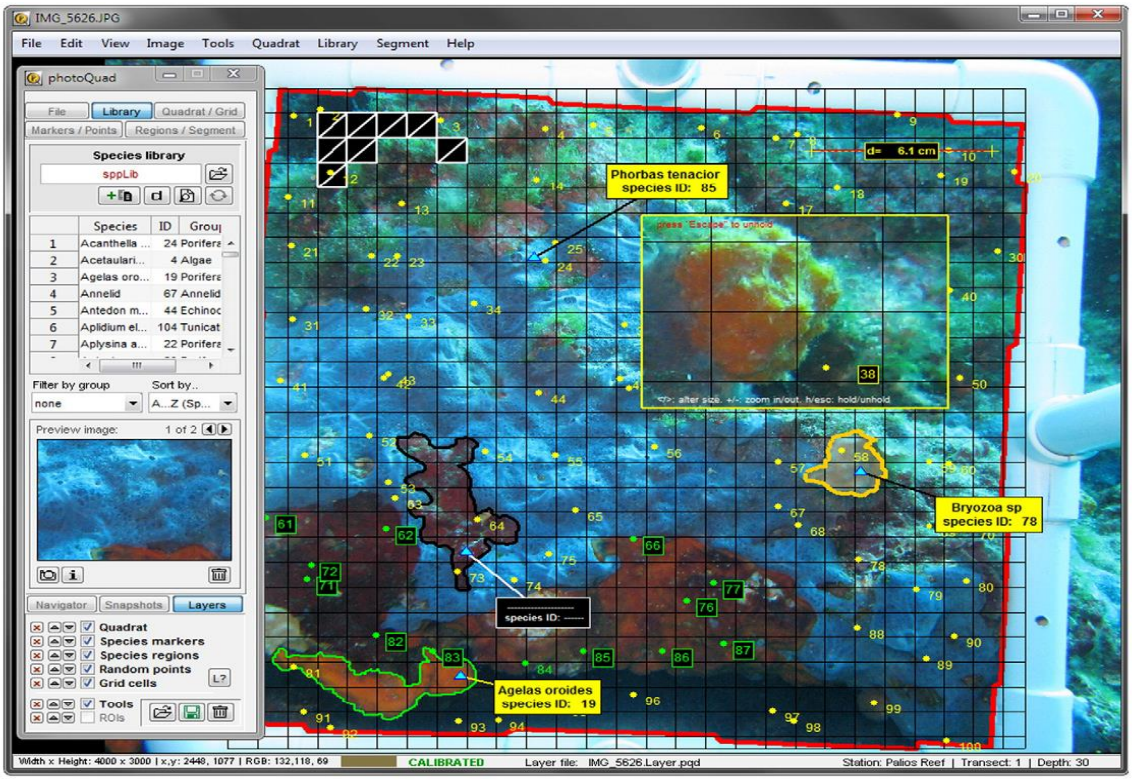


Figure 16 : Logiciel Photoquad (Trygonis et Sini, 2012) : celui-ci permet de superposer une couche d'information sur la photo, en y détourant les formes ou en répartissant des points et/ou des cellules. L'opérateur choisit alors d'attribuer à ces différents objets une valeur (le plus souvent un nom de taxon) prédéfinie dans une "librairie". Au cours de l'analyse, l'opérateur peut enrichir sa librairie avec noms de taxons supplémentaires. Ce logiciel a été choisi car il posséde pour les objectifs des programmes CIGESMED et DEVOTES plus de qualités que ces concurents (souplesse d'utilisation, gratuité, nombre d'options d'analyse)

Le logiciel permet ensuite d'assigner des noms (espèces, genres, familles, phylums dans le cadre de nos travaux) à ces zones ou à ces points. Pour les protocoles utilisés dans le cadre de cette thèse, c'est la méthode des points aléatoires stratifiés qui est utilisée.

Les quadrats photo doivent bien entendu présenter une échelle, calibrée à partir d'un objet de longueur connue lors de la prise de la photo. L'objectif de ces protocoles est d'estimer la densité de points / la surface relative des différents taxons.

La détermination taxonomique se fait uniquement de manière visuelle, sans manipulation de spécimen. Les critères permettant l'identification au niveau de l'espèce ou au moins du genre étant la plupart du temps invisibles sur les photos, l'identification se fait au niveau taxonomique le plus bas possible avec les critères visuels uniquement extérieurs (Figure 17).

En fonction des connaissances sur les présences ou absences d'espèces dans les différentes mers régionales, ces niveaux taxonomiques peuvent donc être différents d'une mer à l'autre.

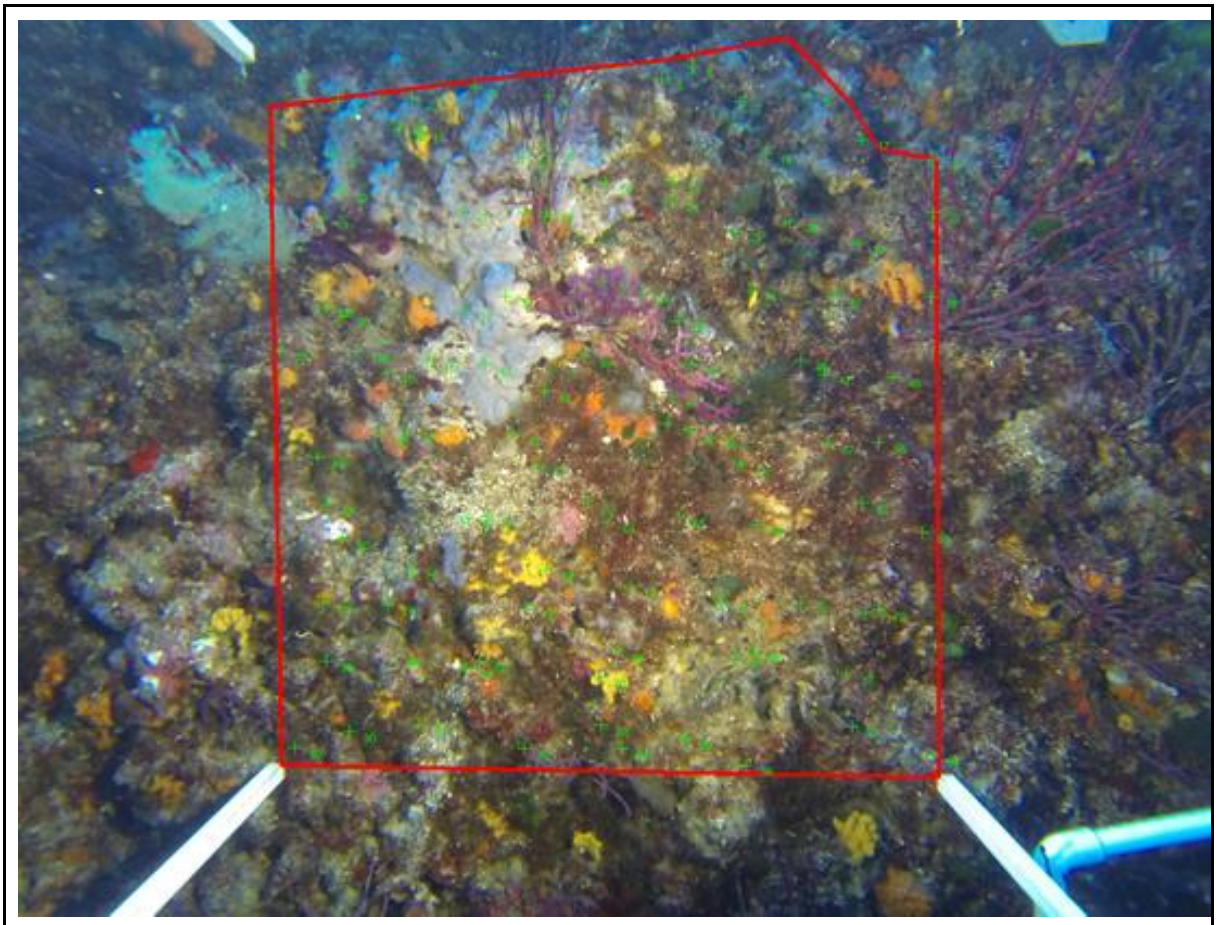


Figure 17 : Traitement d'une photo sur Photoquad : délimitation en rouge de la zone à analyser avec les points répartis de manière "aléatoire stratifiée" et matérialisés par des croix vertes. L'opérateur clique droit sur chaque croix et lui assigne le nom du taxon placé sous le centre de la croix ou bien une catégorie comme "indéterminé" ou "non vivant" choisie dans la "Librairie". La zone maximale étudiée (figurée par le trait rouge) est limitée par les 4 coins du quadrats, dont sont retirées les espaces non exploitables (zones mal éclairées, trous, etc.). En bas à droite on aperçoit le cadre de positionnement[67] autour duquel seront positionnées 8 photos (la neuvième étant effectuée dans le cadre).

[67] Cadre permettant de positionner 9 photographies avec une variabilité minimisée et à partir d'un point fixe. Les coins du cadre sont en fait des croix autour desquelles se calent les quadrats photo.

Méthode d'étude de la variabilité due au système d'observation

La technologie évolue et les instruments, qui ont une durée de vie beaucoup plus faible en mer, sont amenés à être régulièrement changés. Leur inter-calibration à chaque remplacement est indispensable. Les données montrent parfois uniquement les variations dues simplement aux changements de technologie.

Pour comparer les deux appareils photo, deux ensembles de 8 quadrats photo effectués sur un transect permanent au même endroit ont été utilisés. L'un a été fait avec la caméra de qualité moyenne, et l'autre avec la caméra de haute qualité. Mais comme les images n'ont pas été prises au même moment, l'espèce *Paramuricea clavata* a perturbé les observations car elle avait ses polypes étalés ou non, selon l'ensemble. Afin de nous affranchir de cette perturbation, toutes les observations de *Paramuricea clavata* ont été retirées des deux ensembles.

2.2 Méthode de cartographie

La cartographie est réalisée sur différents types de sites, qui peuvent être soit de petites îles, soit des secs (tombants) sans repères externes précis, soit le long de traits de côte présentant un maximum d'orientations (Nord, Sud, Est, Ouest et intermédiaires comme par exemple Nord-Est). Pour chaque site, deux profondeurs sont étudiées, 28 m (± 1m) (CIGESMED) et 42 m (35 à 45 m dans les faits) (IndexCor). Ces cartographies peuvent être complètes[68] ou cartographie partielle[69] selon l'usage qui en est fait dans le cadre des différents "*work packages*" du programme CIGESMED.

Le relevé des profils est inscrit sur une tablette légendée avec le nom du ou des observateur(s), la date, le site, le transect et la profondeur étudiée.

La réalisation de cette cartographie sur l'ensemble du pourtour des sites est impossible (trop de distance, surtout à grande profondeur, par rapport au temps de plongée possible, et présence de toutes les orientations possibles peu évidente pour chacun des sites). Les données ont été récoltées le long de transects (Figure 18) subdivisés en segments de 5 m de longueur (distance parcourue par le plongeur en comptant ses coups de palme) sur 1 m de large.

[68] Relevé complet des profils en détaillant tous les champs selon les recommandations du protocole.

[69] Relevé limité aux profils intéressants à prélever (i.e. profils prédéterminés pour lesquels la fréquence du profil est attestée et la fréquence des populations de *Myriapora truncata* et *Lithophylum cabioche* sont suffisantes pour subir un prélèvement). Cette cartographie sera majoritairement mise en œuvre pour les prélèvements sur sites secondaires, lointains ou pour lesquels il n'y a pas de continuum de coralligène suffisant pour effectuer une cartographie complète. Cela peut aussi être le cas pour la deuxième profondeur, moins accessible, afin de raccourcir le temps d'intervention.

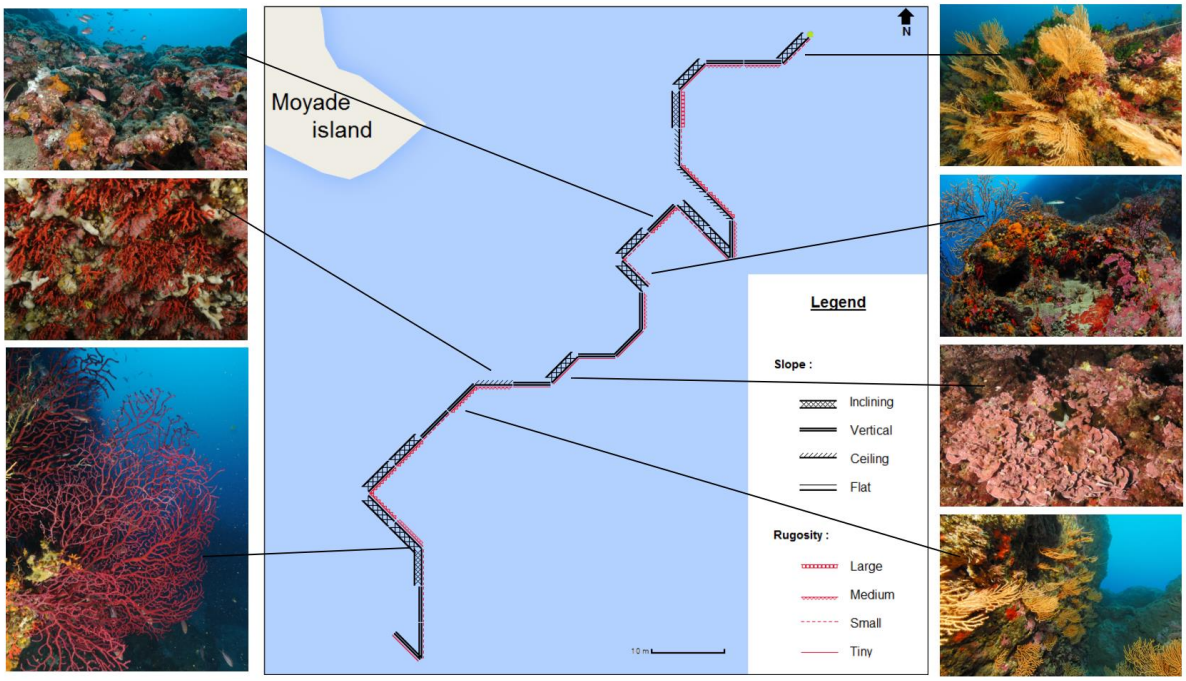


Figure 18 : exemple de relevé d'une plongée au Sud-Est de l'îlot de Moyade lors d'un entraînement : les plongeurs relèvent les profils tous les 5 mètres à 28 m de profondeur. Cette figure illustre la variété des profils qui peuvent coexister sur la même paroi autant en termes d'habitats que de rugosité ou qu'en terme d'inclinaison (décrits à la suite de ce paragraphe). Les orientations relevées face à la paroi pendant cette plongée sont SE, E, NE, N, NO, ce qui se traduit par des expositions opposées NO, O, SO, S, SE. Chaque plongeur participant compare ensuite ses relevés avec le reste de la palanquée.

Le plongeur doit se maintenir à la profondeur fixée au départ. La pratique montre qu'avec de l'entraînement, l'erreur maximale est de 1 mètre au-dessus et en-dessous de cette profondeur objectif. La calibration du trajet de 5 mètres en comptant le nombre de coups de palmes nécessaires demande un peu d'entraînement. Des tests nous ont permis de déterminer que l'erreur sur la distance de 5 mètres était inférieure à 0.7 mètre au bout de quelques entraînements, mais permettait de parcourir des distances plus grandes (parfois jusqu'à deux fois plus qu'en déroulant puis enroulant une cordelette) tout en permettant au plongeur d'effectuer le relevé des facteurs contextes. La calibration est réalisée en début de plongée en installant au fond un étalon (un cordon de 5 mètres avec deux plombs), et s'avère plus efficace et en fin de compte aussi précise que l'usage d'un pentadécamètre (50 mètres) : son installation monopolise un des plongeurs pendant toute la plongée et contre toute attente, la précision de la mesure, liée à la présence d'aspérités, au relief et à la rectitude du cordon n'est pas meilleure qu'avec la méthode des coups de palmes. Cette calibration doit

cependant se faire à chaque début de plongée, en palmant calmement et régulièrement. Le nombre de coups de palme dépend du sens et de la force du courant. Le plongeur scientifique doit aussi apprécier à partir de quel niveau de courant cette calibration n'est plus valable et donc quand les conditions ne permettent plus de réaliser cette cartographie.

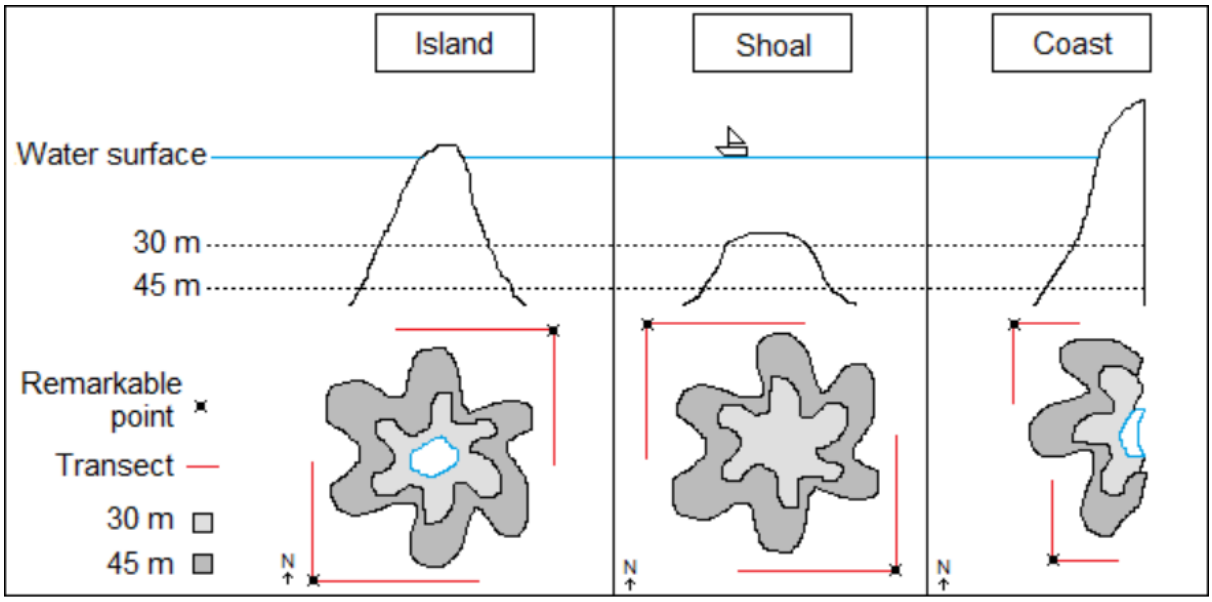


Figure 19 : choix des sites et placement des transects en fonction de la configuration du substrat dur [Island (Île), Shoal (haut fond) ou Coast (côte). L'objectif étant d'effectuer des transects dans des situations les plus contrastées possibles, le positionnement du transect était établi à partir d'un point remarquable (rocher, cap, faille…) aux deux profondeurs fixées et sur deux côtés opposés. Idéalement, les transects s'effectuent en coin, de manière à pouvoir collecter des données pour toutes les orientations disponibles en un minimum de distance. Dans le cas d'une petite île ou d'un haut-fond, les points de départ se situent au Nord-Est et au Sud-Ouest (ou au Nord-Ouest et au Sud-Est selon la présence plus évidente de points remarquables). Dans le cas d'un site sur une côte où ces points complètement opposés sont impossibles, les points remarquables étaient choisis de part et d'autre des caps permettant la meilleure opposition et le maximum d'orientations dans l'échantillonnage.

Pour placer le transect, un point remarquable servant de départ a été choisi par les plongeurs, de préférence dans un angle, afin de pouvoir couvrir plusieurs orientations. Ce point remarquable a été décrit et pris en photo pour chaque site et les coordonnées G.P.S. (Global Positioning System) ont été notées le plus précisément possible (relevé de la position des bulles des plongeurs ou d'un objet flottant lâché par eux, par le surveillant en

surface)[70] afin de pouvoir le retrouver sans difficulté. A partir de ce point, le transect peut être décrit de deux manières :

- En suivant tous les segments dans un seul sens à partir du point remarquable,
- En suivant une partie des segments dans un sens, puis en revenant au point remarquable et en poursuivant dans le sens opposé au premier (à privilégier pour la sécurité et la précision du placement des segments). Si le point remarquable est situé à peu près au milieu environ de la zone d'intérêt, deux palanquées peuvent travailler en même temps (Figure 19).

Pour définir le profil de chaque segment, la typologie suivante a été utilisée :

- Orientation de la paroi : Nord, Sud, Est, Ouest et les quatre intermédiaires (Nord-Est, Nord-Ouest, Sud-Est, Sud-Ouest). Pour déterminer l'orientation, l'observateur se met face à la paroi au milieu du segment, et relève la direction qui va vers la paroi.

Si le segment observé présente plusieurs orientations, l'observateur relève la direction la plus intermédiaire (cas de creux ou des pointes de récif). La nomenclature utilisée sur les plaquettes est la suivante N, NE, E, SE, S, SO, O et NO. L'orientation relevée est donc l'opposée de l'exposition de la paroi (Figure 20).

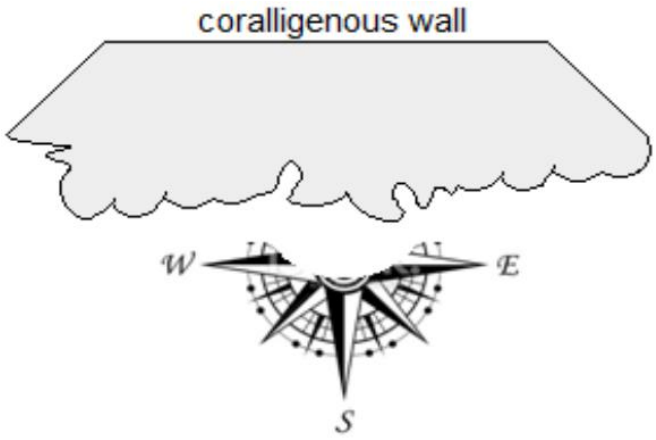


Figure 20 : Le plongeur se retrouve face à la paroi et relève N (pour Nord). L'exposition de la paroi est donc S (pour Sud). Ce choix de transformation du relevé après coup a été effectué car il est difficile de relever les orientations en tournant le

[70] Avec un report sur Modèle Numérique de Terrain (M.N.T.) c'est-à-dire une cartographie en 3D, format World Geodetic System 1984 (W.G.S.84) qui est le système géodésique standard mondial, notamment utilisé par le système GPS.

dos à la paroi, et que la transformation en exposition était rendue plus difficile à cause de la narcose.

NB : certains plongeurs ont préféré relever les directions en degrés sur les compas, la transformation en direction demie cardinales leur semblant moins instinctive avec les effets de la narcose.

- Inclinaison de la paroi : à la différence de Glasby (2000), qui utilisait les inclinaisons « plate » et « verticale sous surplomb », le protocole CIGESMED décline ce paramètre en 4 valeurs : [V, I, F, C] = Vertical, incliné (Inclined), plat (Flat), en surplomb (Ceiling) (Figure 21).

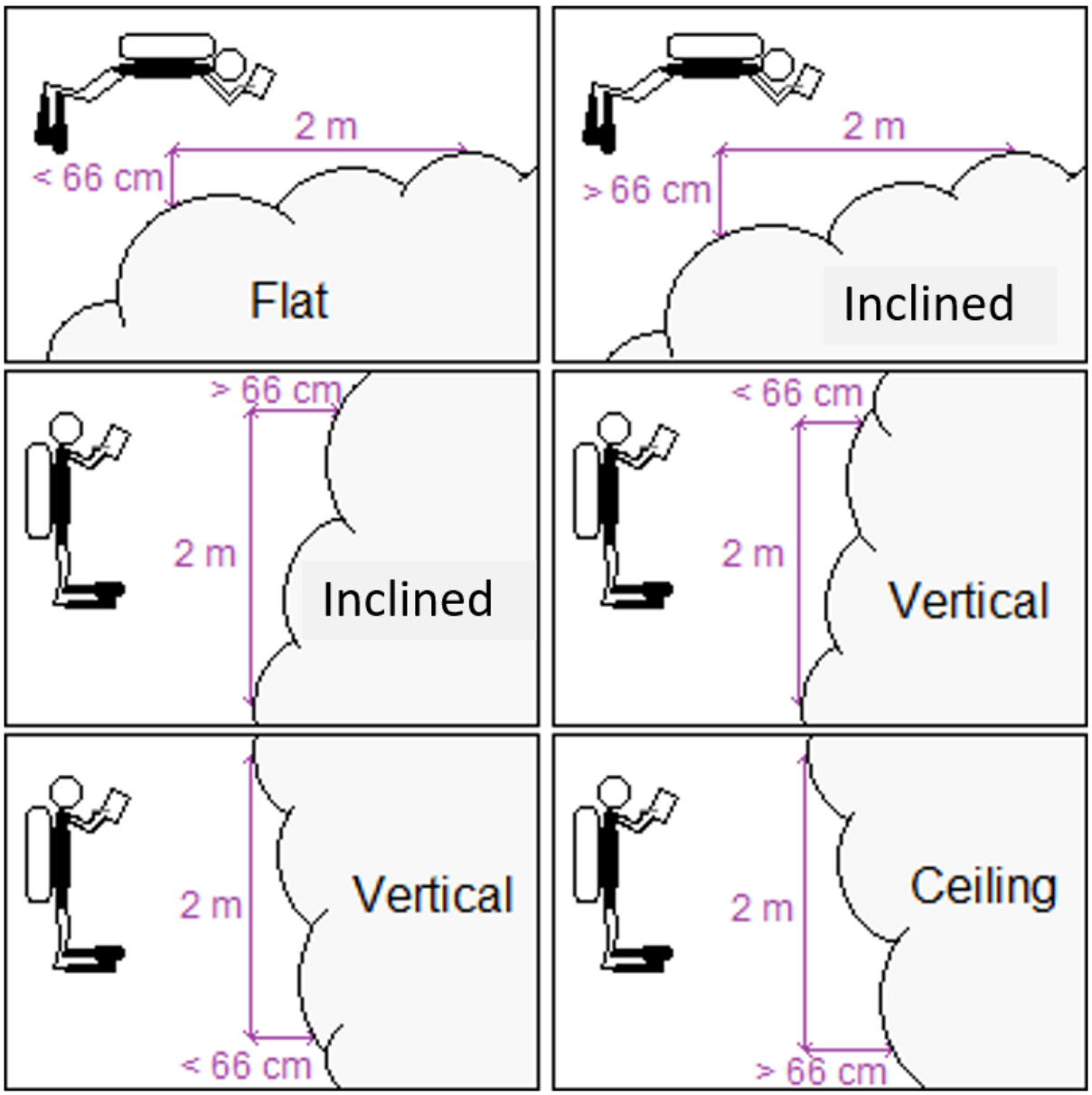


Figure 21 : Les catégories d'inclinaison en trois tiers se déterminent au juger, en s'appuyant mentalement sur une image de triangle rectangle (en s'entraînant), avec une vérification en utilisant des repères corporels que chaque plongeur définit pour lui-même : en théorie, pour une mesure (par exemple de la tête au bout des palmes), pour le côté opposé, il choisit ⅓ de cette mesure pour le côté adjacent (par exemple de l'épaule à la main). Par exemple, pour une mesure de 2 m pour le côté opposé, il choisit ⅓ de cette mesure c'est à dire "66 cm" pour le côté adjacent. En cas de doute, on choisit la catégorie la plus "à l'ombre" entre Vertical et Ceiling, et plutôt "inclined" pour les deux autres limites entre catégories.

La catégorie « Vertical » correspond à une bande de substrat (de 2 m de large autour de la même isobathe) ou une suite d'encorbellements qui se superposent : c'est à dire en se décalant légèrement de la paroi, elle donne l'impression d'être à peu près à la même distance du plongeur un mètre au-dessus et un mètre en-dessous, on doit l'apercevoir en entier (de 66% à 100% d'inclinaison).

La catégorie « Flat » correspond à une bande de substrat (de 2 m de large autour de la même isobathe) ou à un ensemble de concrétions et de rochers qui semble être à peu près au même niveau et très exposé à la lumière. Le plongeur peut s'en assurer lorsque la même distance du plongeur au substrat est à peu près la même lorsque celui-ci se place au-dessus (de 0% à 33% d'inclinaison). Ce profil peut être particulièrement soumis aux apports de particules et aux impacts mécaniques.

La catégorie « Ceiling » correspond à la présence d'un surplomb au-dessus de l'observateur de la taille du plongeur-observateur au moins, sur la majeure partie du segment. C'est-à-dire qu'il couvre le plongeur vu du dessus.

La catégorie « Inclined » correspond à la catégorie intermédiaire entre « Vertical » et « Flat » (de 33% à 66% d'inclinaison). Ceci signifie que sur la largeur de la bande considérée (2 m), le plongeur a l'impression qu'il y a plus de 66 cm (bien sûr, approximée) de dénivelé lorsque ce profil se rapproche de la catégorie « flat », et moins de 66 cm d'avancée du bas de la bande lorsque le profil se rapproche de la catégorie « Vertical ». L'exactitude de l'estimation de l'inclinaison, particulièrement concernant la catégorie « Inclined », dépend plus ou moins de l'appréciation du plongeur, qui ne peut pas passer trop de temps à estimer le type d'inclinaison pour chaque segment. La catégorie « Inclined » est à privilégier dès qu'un doute subsiste, et on admettra que cette inclinaison comporte la plus grande variété d'exposition à la lumière, aux apports de particules et aux courants.

2.3 Méthode d'analyse d'image dans le cadre de CIGESMED

Le principe général de ce protocole de caractérisation des sites repose sur l'analyse de photographies réalisées sur le milieu coralligène étudié. L'analyse des peuplements par transects et quadrats photo est réalisée sur les profils majoritaires.
Plusieurs méthodes d'échantillonnage photographique et de traitement d'images ont été testées afin de sélectionner selon les conditions la méthode la plus efficace pour mieux appréhender la variabilité d'un certain nombre de paramètres.

Liste des taxons utilisés

Une librairie de taxons reconnaissables est constituée sur les conseils d'experts. Cette librairie correspond aux taxons qui seront considérés prioritairement dans l'analyse des quadrats photographiques, et pour lesquels les opérateurs doivent avoir une connaissance minimale. Cette liste permet d'établir quels sont les taxons communs entre les parties orientales et occidentales de la Méditerranée. Elle permet aussi de comparer les profils des habitats sur lesquels on les rencontre, et de découvrir les valeurs de paramètres qui semblent favoriser leur installation. Cette liste sera aussi une base pour les guides méthodologiques qui seront élaborés à l'intention des réseaux de science participative. Certains taxons sont considérés à un niveau supra-spécifique (genre, famille…).

Quelques principes ont été suivis pour construire cette librairie de taxons :
- se baser sur une liste de taxons choisis et validés par des experts,
- sélectionner les espèces reconnaissables en plongée,
- sélectionner les espèces qui ne sont pas rares.

La liste choisie et validée lors du coup d'envoi du programme CIGESMED est celle proposée par le CAR/ASP (UNEP – MAP – RAC/SPA, 2009), complétée par les écologues français. Selon la même logique, les écologues grecs et turcs l'ont adapté pour leurs eaux. Les synonymies éventuelles ont été traitées en se basant sur les sites WoRMS[71] et Algaebase[72] début février 2014. La validation de la bonne dénomination et des descriptions sommaires des espèces a été faite par les experts disponibles en France.

Traitement des photos

Le traitement des photos a été réalisé grâce au logiciel Photoquad (Trygonis et Sini *et al.*, 2012) <http://www.mar.aegean.gr/sonarlab/photoquad/index.php>. Ce logiciel d'analyse d'images 2D est équipé d'outils permettant notamment de mesurer le recouvrement par espèce, de faire des dénombrements et de l'analyse spatiale. La technique utilisée pour comparer les opérateurs (voir Méthode d'étude de la variabilité due à l'opérateur dans la partie inter-calibration dans ce chapitre - page 89) a été ensuite utilisée pour traiter tous les quadrats photo (page 91)

[71] http://www.marinespecies.org/
[72] http://www.algaebase.org/

Contextualisation des quadrats photo lors de la mise en œuvre des relevés photographiques

Dans le cadre de CIGESMED, une méthode de contextualisation (c'est-à-dire de description du contexte) a permis de replacer les photos et vidéos prises et les échantillonnages faits sur deux espèces dans leur contexte.

La première contextualisation des données issues des photos a été faite à l'aide des profils. (Rappel : un profil est constitué des facteurs : orientation, inclinaison, rugosité, recouvrements majoritaires et éléments remarquables tels que la taille des individus ou la taille des peuplements. *cf.* partie Cartographie). On a limité le recueil de quadrats photo dans un premier temps sur deux types de profils d'inclinaison, car plus propices à la présence de coralligène à 30 mètres : incliné et vertical. Ces profils ont été choisis à cause de leur fréquence, et leur disponibilité géographique dans les zones d'applications du programme CIGESMED (France, Grèce et Turquie), et leur orientation. Une contextualisation supplémentaire a été faite en fonction d'éléments déterminés à partir de la photographie (creux, zones exposées, zones à sédimentation…) et/ou à partir des éléments de contexte de l'acquisition (météo, profondeur, visibilité, remarques du plongeur…). Dans le cadre du WP4 de CIGESMED[73], la contextualisation des sites basée sur la cartographie a aussi été utilisée pour l'étude des peuplements de certains bio constructeurs. Aussi, a-t-on privilégié un profil commun comportant suffisamment de possibilités d'échantillonnage de *Myriapora truncata* et de *Lithophyllum cabiochae*.

Nomenclature et archivage des photos

Chaque photographie (quadra-photo) est étiquetée selon une nomenclature précise prédéfinie :

[programme]_[site]_[date] _D[profondeur]_T[n°transect]_Q[n°quadra] _[auteur]

Exemple : CIGESMED_CAS _20140123_ D1_T02_Q08_AA01

Les numéros des photos sont reportés sur la plaquette d'observations complémentaires utilisée lors de la plongée. Chaque plaquette est archivée en mode papier (photocopie) et numérique (scan) et porte son nom avec la même nomenclature que les photos. Toutes les données sont ensuite saisies et portent les mêmes codes. Les formats scannés et saisis des données sont ensuite stockés sur un serveur dupliqué, et mis en accès sur le site internet du programme. Nous verrons par la suite que différents procédés de curation permettront de corriger d'éventuelles erreurs.

[73] http://www.cigesmed.eu/-Module-de-travail-4-outils-

2.4 Méthode d'analyse d'image dans le cadre de DEVOTES

Sites échantillonnés

Les sites échantillonnés dans cette étude se situent dans la partie méridionale du golfe de Gascogne, qui correspond à l'Atlantique Nord-Est, au Nord-Ouest de la côte méditerranéenne autour de Marseille, en Mer Adriatique et en Mer Rouge (figure 22). Trois unités ARMS (trois réplicas) ont été installées sur trois sites dans chaque mer pour un total de neuf unités par mer, à une profondeur comprise entre 7 et 17 m (tableau 3, figure 23). Les informations sur les temps de déploiement et les sites sont données dans le tableau 3. Les sites ont été choisis en fonction des pressions naturelles et humaines données par l'expertise des auteurs dans leurs régions respectives.

Tableau 3 : Informations sur le déploiement des ARMS et les sites suivis par quatre partenaires : AZTI (sites du golfe de Gascogne), CNRS-IMBE (sites du Nord-Ouest de la Méditerranée), CoNISMa (sites de la Mer Adriatique) et KAUST (sites de la Mer Rouge).

SEA REGION (code)	SITE (code)	DEPLOYMENT DATE	RECOVERY DATE	SITE ID [Replicates]	LATITUDE	LONGITUDE	DEPTH (m)
Adriatic Sea (AdS)	Grotta Azzurra (Azz)	Jul-14	Jul-15	CONI_S1	N43 37.313	E13 31.691	7
	Due Sorelle (Sor)	Jun-14	Jul-15	CONI_S2	N43 32.953	E13 37.699	8.7
	La Scalaccia (Sca)	Jun-14	Jul-15	CONI_S3	N43 36.291	E13 33.102	8.8
NW Mediterranean (NWM)	Ile de l'Erevine (ELV)	Jun-13	Dec-14	CNRS_S1	N43 19.780	E05 14.210	17
	Ile Riou (RRS)	Jun-13	Dec-14	CNRS_S2	N43 10.370	E05 23.420	17
	Phare de Cassidaigne (CCA)	Jun-13	Dec-14	CNRS_S3	N43 08.740	E05 32.740	17
Bay of Biscay (BoB)	Lekeitio (Lek)	Jun-13	Jul-14	AZTI_S1	N43 22.311	W2 30.258	12.5
	Zumaia (Zum)	May-13	Jul-14	AZTI_S2	N43 18.748	W2 13.641	11
	Pasaia (Pas)	May-13	May-14	AZTI_S3	N43 20.230	W1 55.639	11
Red Sea_Jeddah (ReS)	Janib Sa'ara reef (JSR)	Apr-13	Jun-14	KAUS_S1	N21 27.253	E39 06.661	10
	South of Jeddah (SOJ)	Apr-13	Jun-14	KAUS_S2	N21 13.508	E39 07.237	10
	Qaham reef (QAR)	Apr-13	Jun-14	KAUS_S3	N21 04.921	E39 12.063	10

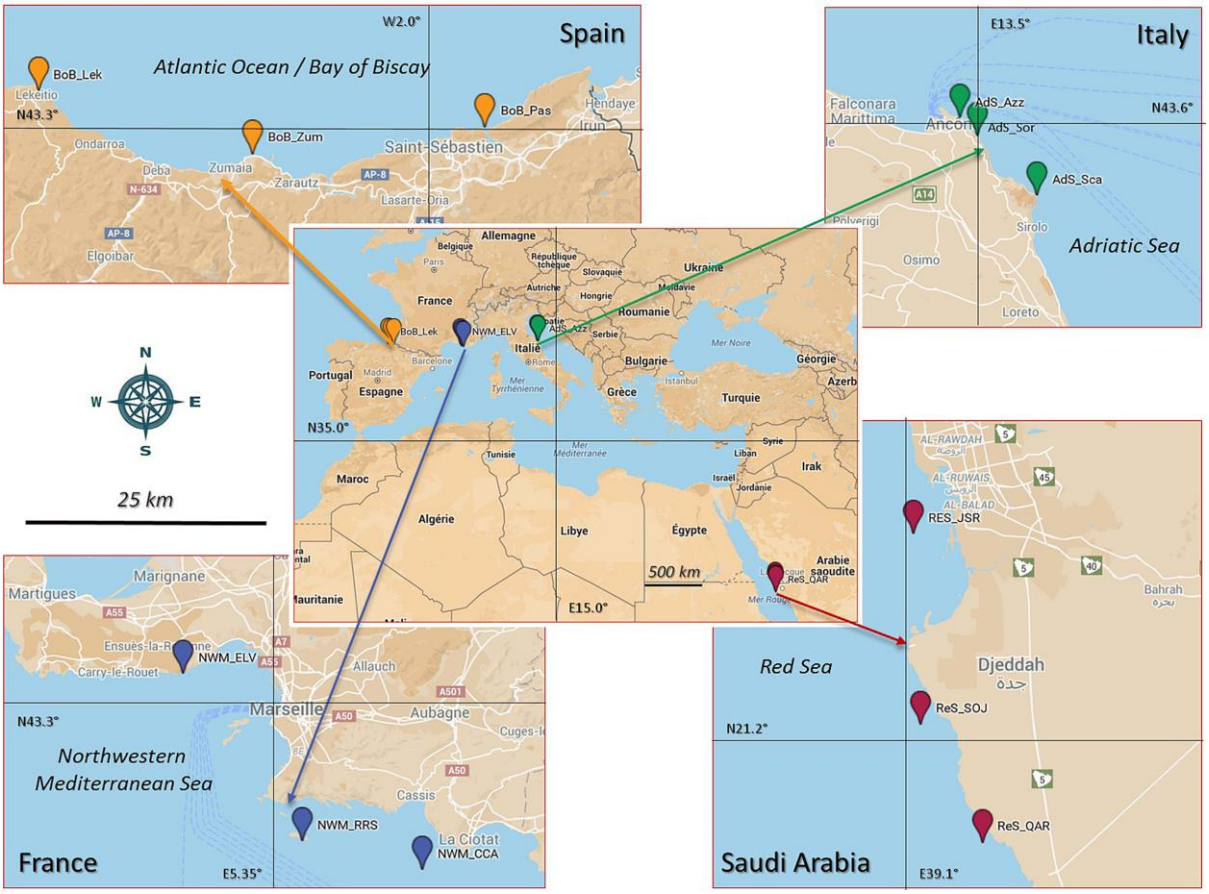


Figure 22 : Position géographique des sites DEVOTES ARMS. Leurs positions géographiques précises sont données au tableau 3.

*Tableau 4 : Contexte anthropique et environnemental des sites d'étude. Chaque site (colonne) est désigné par deux codes à trois lettres, un pour la région marine et un pour le site, séparés par un trait de soulignement (noms complets dans le tableau 3). Les facteurs sont booléens : la présence est indiquée par Y, l'absence par N. Dans la ligne 2 : pour chacune des quatre régions marines, l'effet du site a été testé pour les trois contrastes possibles opposant un site aux deux autres sites (PERMANOVA simple par mer : Face x Site); les nombres (1, 2 et 3) se rapportent respectivement aux contrastes les plus significatifs, les seconds et les moins significatifs et les significations sont représentées par des symboles usuels (NS: non significatif, **: p <0,01, ***: p <0,001, ****: p <0,0001). Pour chaque facteur environnemental (ligne 3 à 12), nous avons mis en surbrillance en vert lorsque la configuration du site la plus contrastée correspondait au site qui différait le plus des deux autres pour le facteur environnemental, en rouge, quand ce n'était pas le cas. Si les facteurs environnementaux n'influencent pas la composition de la communauté (donc les contrastes du site), 1/3 devrait correspondre à chaque contraste possible, donc 1/3 au cas favorable (en vert). Nos résultats ne s'écartent pas des attentes aléatoires avec 6 vertes, 14 rouges. Les abréviations (abr.) sont expliquées dans Tab. 1.*

	BoB_Lek (1)	BoB_Zum (3)	BoB_Pas (2)	NWM_ELV	NWM_RRS	NWM_CCA	AdS_Azz	AdS_Sor	AdS_Sca	ReS_JSR	ReS_SOJ	ReS_QAR
Most contrasted site vs. 2 other ones for each Sea region	1****	3***	2***	1***	3**	2***	1****	2**	3 NS	3 NS	1****	2**
Protection status	N	N	N	N	Y	N	N	N	N	N	N	N
General anthropization	N	N	Y	Y	N	N	Y	Y	Y	Y	N	N
Marine debris	N	N	N	N	N	N	Y	Y	Y	Y	N	Y
Sewage output	Y	Y	Y	N	N	N	Y	N	Y	Y	N	N
Chemical pollution	N	N	Y	Y	Y	Y	Y	Y	Y	Y	N	N
Urbanization	N	N	N	Y	N	N	Y	N	Y	Y	N	N
Harbor	Y	Y	Y	Y	N	N	N	N	N	Y	N	N
Nearby Seagrass meadows	N	N	N	Y	N	Y	N	N	N	N	N	N
Nearby sand	Y	N	N	Y	N	Y	N	N	Y	Y	Y	Y
Nearby mud	N	N	N	Y	N	N	N	N	Y	N	N	N

Installation et récupération des ARMS

Chaque unité ARMS est composée de neuf plaques et entretoises en PVC de 22,5 cm x 22,5 cm empilées dans une série alternée d'espaces d'inter plaques ouverts et fermés, attachés à une plaque de base de 35 cm x 45 cm (figure 23). De plus amples détails sur l'assemblage standard, le déploiement et la récupération de l'ARMS sont disponibles sur le site Web de la N.O.A.A.[74] et dans González-Goñi *et al.*, 2003.

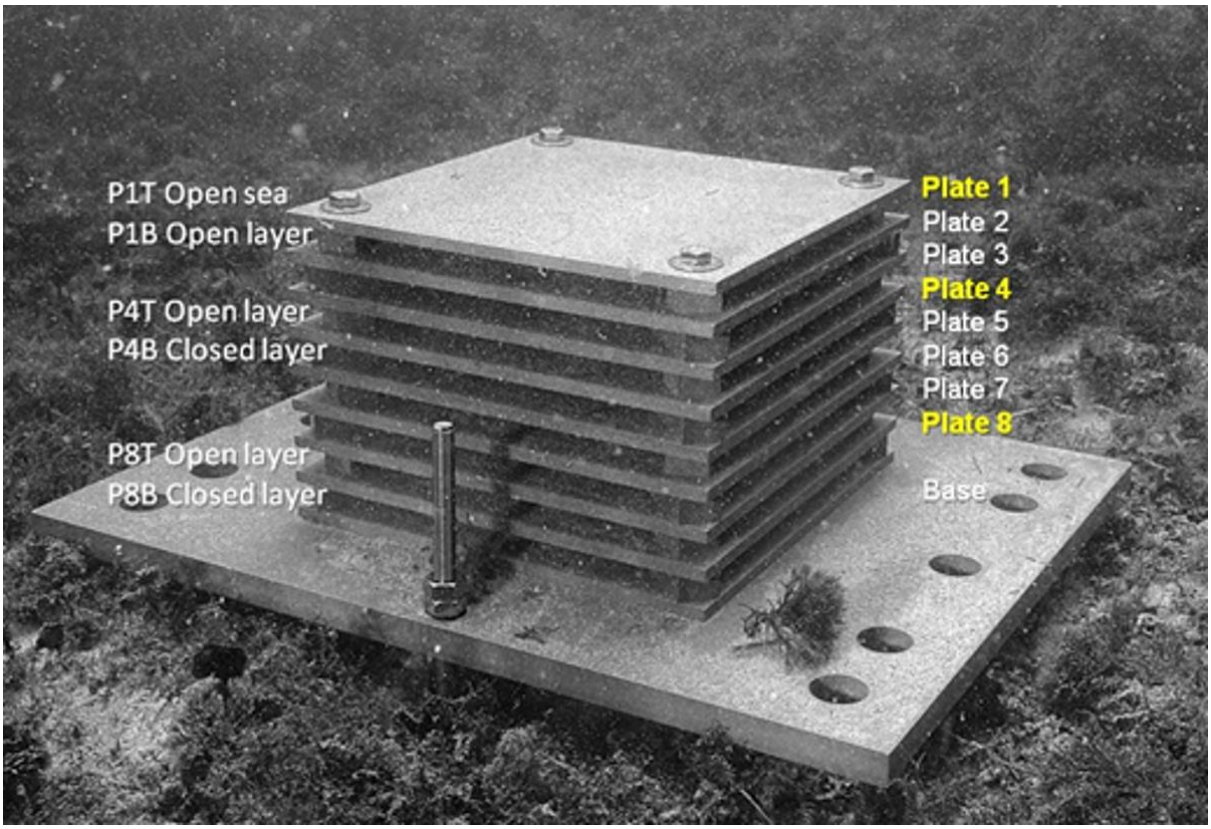


Figure 23 : Un ARMS nouvellement déployé (Île de l'Erevine, NW Méditerranée). L'utilisation alternative des espaceurs courts et longs en PVC donne une tour de quatre couches ouvertes et de quatre couches fermées. © photo CNRS / F. Zuberer

Les unités ARMS ont été installées par des plongeurs et submergées pendant 12 à 16 mois, selon la mer (Tableau 2). Par la suite, les ARMS ont été récupérés et renvoyés au laboratoire, où ils ont été démontés et traités. Chaque surface de la plaque a été brossée doucement pour enlever la faune mobile sans détacher les organismes sessiles. Les plaques ont été

[74] https://www.pifsc.noaa.gov/cred/survey_methods/arms/overview.php

conservées dans de l'eau de mer aérée avec des bulleurs jusqu'à ce que des photographies soient prises (figure 24).

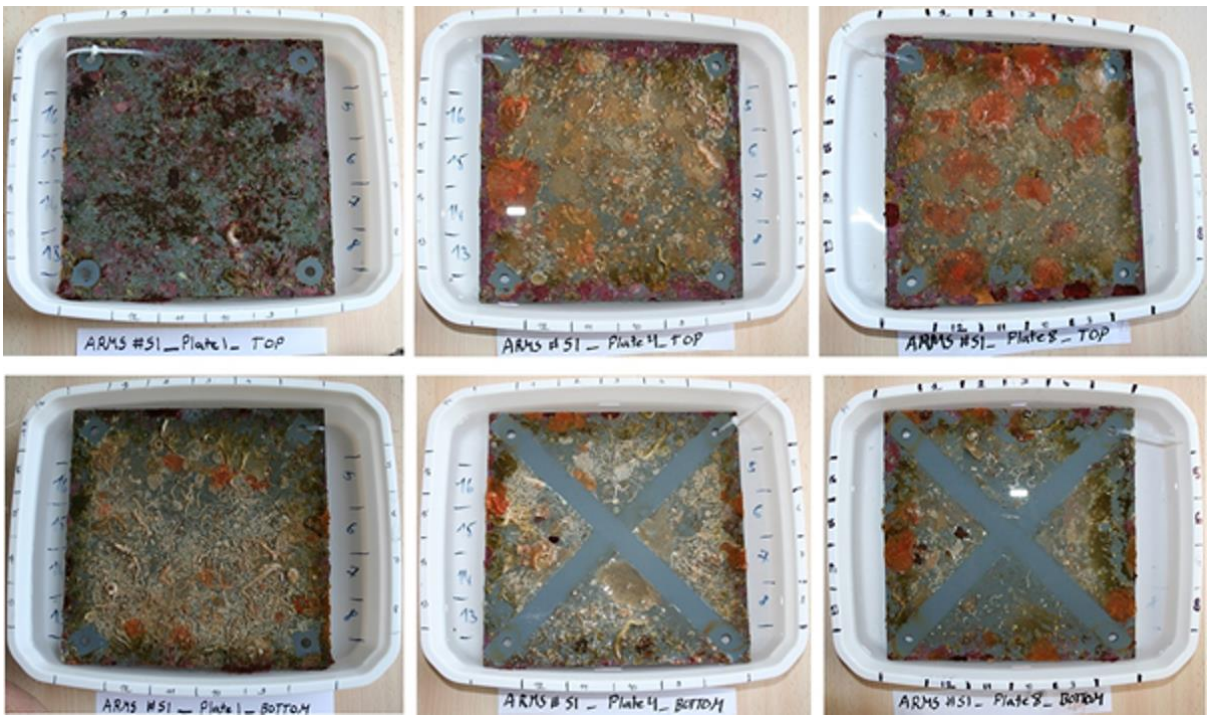


Figure 24 : Faces de Plaques des ARMS échantillonnées (de gauche à droite : 1, 4 et 8, images supérieures : faces supérieures P1T, P4T et P8T, images inférieures : faces inférieures P1B, P4B et P8B) après avoir été récupérées en Mer et avoir collecté la fraction mobile [NW Méditerranée, Ile Riou] . Les marques croisées dans deux des plaques de fond résultent des entretoises transversales complètes ou partielles qui sont placées de manière alternée entre certaines plaques.

Analyses de photos

Trois plaques (P) (plaques 1, 4 et 8) (figure 24) ont été sélectionnées pour l'analyse et pour chacune, les surfaces supérieures (T) et inférieures (B) ont été analysées individuellement (c.-à-d. 6 faces de plaques analysées par ARMS). Ces faces ont été sélectionnées pour pouvoir considérer les différents habitats trouvés dans un ARMS, qui représentent différentes conditions pour les organismes s'y installant. La surface supérieure de la plaque 1 est exposée à la lumière directe et sans aucune protection contre les prédateurs ou les brouteurs, tandis que les cinq autres faces ne le sont pas. Parmi ces faces à l'ombre, P1B, P4T, P8T sont ouvertes au courant, tandis que les faces P4B et P8B n'ont pas d'espaces fermés dûs au compartimentage, ce qui fait que le courant circule peu dans l'espace entre ces plaques (Figures 23 et 24).

Les photographies ont été analysées à l'aide du logiciel Photoquad® (Trygonis et Sini, 2012). Chaque photographie était divisée en 64 carrés et un point était choisi au hasard dans chaque carré. L'organisme présent sous chaque point a été identifié au niveau taxonomique le plus bas possible par des scientifiques des quatre mers (les scientifiques ont analysé les ARMS de leur propre région). Avant l'analyse globale, certaines catégories taxonomiques ont été fusionnées afin d'être compatibles entre les quatre régions et de minimiser l'éventuel effet généré par des observateurs différents. Les catégories initiales sont fournies dans des documents supplémentaires (fichier S2, Annexe 4.4). Les catégories finales (fusionnées) sont : Annélides, Bryozoaires, Mollusques, Cnidaires, Porifères, Crustacés, Tuniciers coloniaux, Ascidies, algues corallines calcaires (ci-après ACC), autres Rhodophytes, Chlorophytes, Phaeophytes, Autres algues, Foraminifères, "indéterminé", et "non vivant". Les points qui tombaient sur des parties non colonisables de la plaque en raison de la présence de croix ou d'entretoises compartimentantes (figure 24) étaient considérés comme « non vivants » (tous les partenaires n'avaient pas créé de catégorie «non colonisable» en analysant leurs photos). Les analyses communautaires ont été effectuées avec et sans inclusion des catégories « non vivant » et « indéterminé » pour l'ensemble des données.

Facteurs environnementaux

La simplicité et l'opérabilité requises pour la surveillance explique également notre utilisation de facteurs de contexte « définis de manière générale ». Sur les côtes européennes, les trois sites ont été choisis pour refléter des situations environnementales contrastées et respecter les conditions d'être sur des fonds durs et à une distance raisonnable pour rendre le travail de terrain faisable et s'assurer que le pool potentiel d'espèces colonisatrices était partagé entre les sites dans une mer. Dix facteurs environnementaux binaires ont été évalués à dire d'experts de manière à permettre des tests d'hypothèse : présence d'un statut de protection (tel qu'un parc national ou une aire marine protégée), anthropisation générale, débris marins (présence déclarée d'objets en plastique visibles, déchets, et filets abandonnés), rejet d'eaux usées, pollution chimique (influence évaluée selon dire d'expert), urbanisation (influence évaluée selon dire d'expert), port (influence évaluée selon dire d'expert), herbiers à proximité, sable proche et case proche [à proximité immédiate avec une influence probable sur ARMS (avis d'expert plongeur)].

Analyses statistiques

Nous avons utilisé le package PRIMER (v.7) (Clarke *et al.*, 2014 et Clarke & Gorley, 2015) pour toutes les analyses de la communauté qui ont été effectuées sur l'ensemble des données (avec des échantillons de toutes les faces, sites et mers). Pour toutes les analyses, nous avons utilisé la mesure de dissimilarité de Bray-Curtis. Toutes les analyses ont été

réalisées deux fois, avec ou sans transformation “fourth root” des données d'abondance (les transformations sont fortement recommandées pour réduire l'effet des taxons abondants sur les matrices de dissimilarité, mais nous voulions vérifier que nos résultats n'étaient pas biaisés par une transformation particulière).

L'analyse multivariée a été entreprise en utilisant à la fois PERMANOVA (effets fixes, sommes de carrés de type III sauf pour les plans imbriqués où le type I était utilisé) et, pour permettre des comparaisons avec des études antérieures, ANOSIM (tous les facteurs étaient non ordonnés). Nous avons fait 9999 permutations pour chaque test. Nous avons d'abord testé l'effet de la mer, des sites (imbriqués dans la mer) et de la face de la plaque. Nous avons ensuite testé l'effet de chaque facteur environnemental dans un “schéma croisé en trois dimensions” contenant également les facteurs mer et plaque (à partir d'ensembles de données partielles incluant uniquement les régions maritimes pour lesquelles le facteur environnemental variait d'un site à l'autre). Nous avons effectué une PERMANOVA imbriquée avec les facteurs mer et site (à l'intérieur de la mer) mais sans le fac-similé pour montrer les conséquences de la non-distinction des faces des plaques, comme cela a été entrepris dans une récente étude de metabarcoding (Pearman *et al.*, 2016). Nous avons comparé la dispersion des compositions de la communauté entre les mers, les sites et les plaques (et testé l'hypothèse nulle qu'elle ne variait pas) en utilisant PERMDISP. Des analyses supplémentaires ont également été effectuées sur des jeux de données partiels. Les abondances de taxons ont été transformées logarithmiquement avant d'effectuer l'analyse de variance pour déterminer si les effets de la mer et du site étaient significatifs sur l'abondance relative de chaque catégorie taxonomique. Les ANOVA ont été réalisées dans R version 3.2.4 (R Core Development Team, 2016). De même, nous avons testé si les effets de la mer et du site sur la composition de la communauté étaient significatifs pour chaque face de plaque en utilisant individuellement une PERMANOVA imbriquée (six analyses distinctes). Nous avons également testé séparément pour chaque mer et chacune des deux faces de la plaque (P1T, P4B) si l'effet du site sur la composition de la communauté était significatif (huit analyses distinctes ont donc été effectuées sur de petits ensembles de données pour quatre mers). Nous avons calculé, pour chaque mer, l'effet du site dans chacune des trois conceptions contrastées possibles opposant un site par rapport aux deux autres sites, dans une PERMANOVA croisée à deux facteurs avec des facteurs de site et de face de plaque. Cela permet de vérifier si la configuration des sites les plus contrastés correspond à des facteurs environnementaux contrastés eux aussi (Tableau 4).

3. Résultats concernant l'efficacité et la mise en pratique des outils, méthodes et protocoles

3.1 Résultats concernant l'inter-calibration

Résultats concernant les choix de variables et les modalités qu'elles peuvent prendre

Afin de décrire de manière commune les composantes des habitats coralligènes, un ensemble de variables a été communément adoptée et des modalités ont été fixées pour chacune d'entre elles. Ces variables constituent un vocabulaire contrôlé (que je considère comme un résultat clef) qui pourra dans le futur être réorganisé sous la forme d'un micro-thésaurus. Le résultat qui en découle est un dictionnaire de données rassemblant différents types de descripteurs. Ces descripteurs peuvent parfois prendre une valeur et une seule, ou plusieurs valeurs, et caractérisent les liens entre une espèce ou un groupe d'espèces et sont plus ou moins bien décrits dans la littérature pour les différents taxons. 20 descripteurs (appelés aussi traits biologiques[75]) de structures des taxons traitent de leur morphologie (Tableau 5) et de leur apparence (Tableau 6), 36 descripteurs décrivent les relations fonctionnelles entre taxons et notamment les relations de type trophique (tableau 7_1 et 7_2), de type reproduction (tableau 8), rôle structurant dans les habitats coralligènes (tableau 9), et autres traits fonctionnels comportementaux (tableau 10). Dans le tableau 11, des descripteurs appelés "anthropocentriques" concernent la perception sociale des taxons, le tableau 12 présentant un exemple de descripteurs concernant la notion de services écosystémiques et une première proposition simplifiée de descripteurs concernant l'intérêt économique des taxons (tableau 13). Ce type de descripteur est encore peu abouti et nécessitera une recherche pluridisciplinaire et collective pour devenir utilisable à large échelle, en s'appuyant par exemple sur un premier travail réalisé dans ce sens (Thierry De Ville D'Avray 2018, Thierry de Ville d'Avray *et al.*, 2018).

Enfin, pour permettre d'élargir encore les possibilités d'analyse, un travail identique a été fait pour caractériser les sites, les transects ou des observations libres avec 37 descripteurs de contexte (tableau 14, 15, 16_1, 16_2 et 17).

Même si chaque descripteur ne s'applique pas forcément à tous les taxons et/ou observations (ou à une relation entre deux taxons) et que certains ne sont pas complètement

[75] Les « traits biologiques » appelés aussi « traits de vie » sont des descripteurs biologiques et comportementaux associés à un taxon, à une communauté ou à un habitat. Ils peuvent être quantitatifs (respiration, croissance, mode/rythme/stratégie de reproduction et alimentation) ou écologiques (*preferendum* de température, salinité, pH, etc.).

indépendants (par exemple, la taille et la relation proie-prédateur), il existe un très grand nombre de possibilités de combinaisons de valeurs de modalités à explorer pour trouver des inférences entre contextes, fonctions écologiques, traits d'espèces et structures de communautés. Une illustration de ces explorations utilisant les méthodes de fouille de données sera présentée par la suite (voir chapitre 4 partie 1.4.).

Les différents descripteurs et les modalités proposés par la communauté CIGESMED sont accessibles en ligne sur le web[76].

Tableau 5 : Descripteurs de structure des taxons traitant de morphologie avec leurs modalités (mod.).

Trait [en]	**Trait [fr]**	**nb mod.**	**Modalities [en]**	**Modalités [fr]**
growth form	stature	2	encrusting ; erected	encroûtant ; dressé
class of size	classe de taille	7	nail; finger; fist; foot, palm; human size ; more than human size	ongle ; doigt ; poing ; pied , palme ; plongeur ; plus que plongeur
average width	largeur moyenne	inf	in mm	en mm
maximum width	largeur maximale	inf	in mm	en mm
covering	recouvrant	2	yes ; no	oui ; non
thickness	épaisseur	inf	in mm	en mm
transparency	transparence	3	null ; low ; hight	nulle ; faible ; forte
biomass	biomasse	inf	in g / individu	en g / individu
consistency	consistance	2	soft ; solid	mou ; solide
shock resistance	résistance aux impacts	3	low ; medium ; hight	faible ; moyen ; fort
substrate adesion	adhésion au substrat	3	low ; medium ; hight	faible ; moyen ; fort
shape variability	variabilité de la forme	3	very homogeneous among individuals; variable among individuals, but resembling; very heterogeneous among individuals	très homogène entre individus ; variable entre individus, mais se ressemblant ; très hétérogènes entre individus

[76] http://www.cigesmed.eu/Bottom-up-initiative-on-a

Tableau 6 : Descripteurs d'apparence pour les individus / colonies avec leurs modalités (mod.).

Trait [en]	**Trait [fr]**	**nb mod.**	**Modalities [en]**	**Modalités [fr]**
maximum coverage	couverture max	inf	in cm2	en cm2
maximum encrusting	encroutement max	inf	in cm2	en cm2
surface type	type de surface	8	pierced ; alveolate ; smooth ; veiny ; granular ; striped ; rough ; irregular	percée ; alvéolée ; lisse ; veineuse ; granuleuse ; striée ; rugueuse ; irrégulière
main color	couleur de rattachement	6	red ; yellow ; blue ; green ; brown ; none	rouge ; jaune ; bleue ; vert ; marron ; sans
colors	couleurs	12	violet ; blue ; green ; yellow ; orange ; pink ; red ; grey ; beige ; brown ; black ; white	violet ; bleue ; verte ; jaune ; orange ; rose ; rouge ; grise ; beige ; marron ; noir ; blanche
height	hauteur	inf	in mm	en mm
colonial	colonial	2	simple ; colonial	simple ; colonial
shape	forme	6	long and simple; long and complex; Round and simple; Round and complex, plain and simple, flat and complex	long et simple ; long et complexe ; ronde et simple ; ronde et complexe, plat et simple, plat et complexe

Tableau 7_1 et 7_2 : Descripteurs de traits fonctionnels concernant les taxons traitant de relations entre espèces et relations trophiques avec leurs modalités (mod.).

Trait [en]	**Trait [fr]**	**nb mod.**	**Modalities [en]**	**Modalités [fr]**
prey of	proie de	inf	specie[X]; gender[X]; familly[X]	espèce[X]; genre[X]; famille[X]
parasite of	parasite de	inf	specie[X]; gender[X]; familly[X]	espèce[X]; genre[X]; famille[X]
parasitized by	parasité par	inf	specie[X]; gender[X]; familly[X]	espèce[X]; genre[X]; famille[X]
symbiont of	symbiote de	inf	specie[X]; gender[X]; familly[X]	espèce[X]; genre[X]; famille[X]
commensal of	commensal de	inf	specie[X]; gender[X]; familly[X]	espèce[X]; genre[X]; famille[X]
epiphyte[77] of	épiphyte de	inf	specie[X]; gender[X]; familly[X]	espèce[X]; genre[X]; famille[X]
epibionte[78] of	épibionte	inf	specie[X]; gender[X]; familly[X]	espèce[X]; genre[X]; famille[X]
habitat of	espèce habitat	inf	specie[X]; gender[X]; familly[X]	espèce[X]; genre[X]; famille[X]

Trait [en]	**Trait [fr]**	**nb mod.**	**Modalities [en]**	**Modalités [fr]**
predator of	prédateur de	inf	specie[X]; gender[X]; familly[X]	espèce[X]; genre[X]; famille[X]
trophic category	catégorie trophique	5	top carnivorous; lower carnivorous; herbivore; planktonophagous	carnivore supérieur ; carnivore inférieur ; herbivore ; plantonophage

[77] organisme autotrophe (capable de photosynthèse) se servant d'autres plantes comme support.
[78] organisme non parasite utilisant les surfaces externes d'animaux plus grands que lui comme support.

omnivory index	indice d'omnivorie	3	1 ; 2 ; 3 and more	1 ; 2 ; 3 et supérieur
feeding habit	mode d'alimentation	4	predator; active filter; passive filter; photosynthetic; grazer ; crusher	chasseur ; filtreur actif ; filtreur passif ; photosynthétiseur ; brouteur ; broyeur
diet	régime alimentaire	10	algae ; plankton ; corals ; fixed invertebrates ; small fishes ; large fishes; carrion ; wastes; crustaceans ; mollusks ; echinoderms	algues ; plancton ; coraux ; invertébrés fixés ; petits poissons ; gros poissons ; déchets ; crustacés ; mollusques ; échinodermes
condition of food	condition d'alimentation	3	inorganic ; organic alive ; organic dead	inorganique ; organique vivant ; organique mort
feeding period	période d'alimentation	2	diurnal ; nocturnal	diurne ; nocturne
prey size	taille des proies	3	nutriment ; microphagous ; macrophagous	nutriment ; microphage ; macrophage

Tableau 8 : Descripteurs de traits fonctionnels concernant les taxons traitant de reproduction avec leurs modalités (mod.).

Trait [en]	Trait [fr]	nb mod.	Modalities [en]	Modalités [fr]
type of reproduction	type de reproduction	4	asexual; sexual; sexual and asexual; indeterminate	asexué ; sexué ; asexué et sexué ; indéterminé
reproductive strategy	mode/stratégie de reproduction	3	oviparous; viviparous; ovoviviparous	ovipare ; vivipare ; ovovivipare
reproductive period	période de reproduction	12	1 ; 2 ; 3 ; 4 ; 5 ; 6 ; 7 ; 8 ; 9 ; 10 ; 11 ; 12	1 ; 2 ; 3 ; 4 ; 5 ; 6 ; 7 ; 8 ; 9 ; 10 ; 11 ; 12
migrant	migrateur	2	yes ; no	oui ; non
hermaphrodism	hermaphrodisme	5	protogynous ; protandrous ; simultaneous ; none : non documented	protogyne ; protandre ; simultané ; aucun ; non documenté

Tableau 9 : Descripteurs de traits fonctionnels concernant les taxons traitant de fonction structurante avec leurs modalités (mod.).

Trait [en]	Trait [fr]	nb mod.	Modalities [en]	Modalités [fr]
coralligenous builder	constructeur du coralligène	2	yes / no	oui/non
competition for the upper layer	compétition pour l'espace "strate haute"	4	null; weak; proven; strong	nulle ; faible ; avérée ; forte
competition for the bottom layer	compétition pour l'espace "strate basse"	4	null; weak; proven; strong	nulle ; faible ; avérée ; forte
keystone species	espèce clef	4	no ; possible ; yes ; unknown	non ; possible ; oui ; inconnu
engineer species	espèce ingénieur	2	yes / no	oui / non
bio-eroder	bio-erodeur	2	yes / no	oui / non

Tableau 10 : Descripteurs de traits fonctionnels concernant les traits de vie des taxons avec leurs modalités (mod.).

Trait [en]	**Trait [fr]**	**nb mod.**	**Modalities [en]**	**Modalités [fr]**
life span	longévité	3	infra-annual; pluriannual; decennal and more	infra-annuelle ; pluriannuelle ; décenale et plus
seasonality	saisonnalité	12	1 ; 2 ; 3 ; 4 ; 5 ; 6 ; 7 ; 8 ; 9 ; 10 ; 11 ; 12	1 ; 2 ; 3 ; 4 ; 5 ; 6 ; 7 ; 8 ; 9 ; 10 ; 11 ; 12
ecological sensitivity	sensibilité? écologique	4	sensitive; opportunistic; ubiquitous; invasive	sensible ; opportuniste ; ubiquiste ; invasive
gregarious	grégarité	3	solitary ; couple ; small groups ; big groups	solitaire ; couple ; petits groupes ; grands groupes
behaviour	comportement	5	curious ; indifferent ; fearful ; threatening ; aggressive	curieux ; indifférent ; peureux ; menaçant ; agressif
defenses	défenses	9	chameleon ; camouflage ; bite ; urticant ; spines ; grip ; electricity ; poison ; allopathy	caméléon ; camouflage ; morsure ; urticant ; épines ; pinces ; electricité ; poison ; allopathie
mobility	mobilité	4	sessile; low mobility and territorial; Mobile and territorial; very mobile	immobile ; peu mobile et territorial ; mobile et territorial ; très mobile
habitat function	Fonction de l'habitat pour l'espèce considérée	3	refuge ; feeding ; nursery	refuge, nourriture, nurserie
habitat dependency level	Niveau de dépendance à l'habitat pour l'espèce considérée	3	essential habitat ; secondary habitat ; non-used	habitat essentiel ; habitat secondaire ; non utilisé

Tableau 11 : "Exemples de descripteurs utiles à la gestion" des taxons concernant la perception humaine avec leurs modalités (mod.).

Trait [en]	Trait [fr]	nb mod.	Modalities [en]	Modalités [fr]
flagship species	espèce phare	2	yes / no	oui/non
status	statut UICN	3	threatened ; endangered ; critically endangered	menacé ; en danger ; en voie d'extinction
protection list	liste de protections	3	regional red list ; national red list ; european annexe list	liste rouge régionale ; liste rouge nationale ; annexe européenne
regulated fishery	pêche réglementée	2	yes / no	oui / non
protected species	espèce protégée	2	yes / no	oui / non
conspicuous	commun et facilement identifiable	2	yes / no	oui/non

Tableau 12 : exemple de "descripteurs utiles à la gestion" des taxons concernant les services écosystémiques avec leurs modalités (mod.). D'autres descripteurs de ce type sont détaillés et discutés dans la thèse de Laure Thierry de ville d'avray, 2018.

Trait [en]	Trait [fr]	nb mod.	Modalities [en]	Modalités [fr]
carbon sequestration	séquestration du carbone	3	null, possible, certain	null, possible, avéré

Tableau 13 : exemple de "descripteur utiles à la gestion" des taxons concernant l'intérêt économique des taxons avec leurs modalités (mod.).

Trait [en]	Trait [fr]	nb mod.	Modalities [en]	Modalités [fr]
edibility	comestibilité	4	poisoned; no culinary interest; appreciated; sought	empoisonnée ; sans intérêt culinaire ; appréciée ; recherchée
commercial species	espèce commerciale	3	no value; low commercial value; high commercial value	sans valeur commerciale ; faible valeur commerciale ; forte valeur commerciale

photogenic species	espèce photogéniqu e	3	no; sometimes photographed; searched for photography	non ; parfois photographiée ; recherchée pour la photographie
means of collection	type de récolte	inf	longline fishing ; fixed gillnet ; trammel nets ; trammel-gillnet ; bottom fishing ; rock fishing ;	pêche à la palangre ; pêche au filet fixe maillant ; pêche au filet fixe trémail ; pêche au filet fixe combiné ; pêche à la palangrotte ; pêche à soutenir ;
collectible	objet de collection	3	collectible	sans intérêt ; taxon apprécié ; taxons très recherché
economical interest	intérêt économique s	3	chemical ; food ; cosmetics ; pharmaceutical ; touristic ; craft industry ; aquariums	chimique ; alimentaire ; cosmétique ; pharmaceutique ; touristique ; artisanat ; aquariums

Tableau 14 : Descripteurs des préférendums des taxons avec leurs modalités (mod.).

Trait [en]	**Trait [fr]**	**nb mod.**	**Modalities [en]**	**Modalités [fr]**
brightness/radiance	luminosité	4	subsurface ; infralittoral[79] ; circalittoral[80] ; bathyal[81] ;	subsurface ; infralittoral ; circalittoral ; bathyal ;
minimum brightness tolerated	luminosité minimale tolérée	4	subsurface ; infralittoral ; circalittoral ; bathyal ;	subsurface ; infralittoral ; circalittoral ; bathyal ;
optimal maximum temperature	température optimale max	inf	in degrees Celsius	en degrés Celsius
optimal minimum temperature	température optimale min	inf	in degrees Celsius	en degrés Celsius
sedimention tolerance	tolérance à la sédimentation	3	no sedimentation tolerance ; low sedimentation tolerance ; good sedimentation tolerance ;	pas de tolérance à la sédimentation ; faible tolérance à la sédimentation ; bonne tolérance à la sédimentation
preferential substratum	substrat préférentiel	4	mud ; sand ; biogenic substrate ; rocky shore	vase ; sable ; substrat biogénique ; substrat rocheux
sensitivity to organic pollutants	sensibilité aux polluants organiques	3	weak; strong; undocumented	faible ; forte ; non documentée
sensitivity to mechanic actions	sensibilité aux actions mécaniques	3	weak; strong; undocumented	faible ; forte ; non documentée
salinity preferendum	préférences salinité	3	halophilic; halophobic; euryhaline	halophile ; halophobe ; euryhaline
minimal depth	profondeur minimale	inf	in m for one specific region	en m pour une région donnée
maximal depth	profondeur maximale	inf	in m for one specific region	en m pour une région donnée

[79] En méditerranée, l'étage infralittoral est la partie du littoral constamment immergée dont la frange supérieure correspond à la ligne de base (le zéro des cartes). Sa limite inférieure est celle qui est compatible avec la vie des algues photophiles et des phanérogames marines.

[80] L'étage circalittoral correspond à la partie basse de la zone photique, la partie du littoral la plus profonde, presque totalement sombre. En Méditerranée, cet étage commence à la limite inférieure des herbiers de posidonies, jusqu'à la profondeur où les algues sciaphiles disparaissent.

[81] L'étage bathyal s'étend des marges du plateau continental (-200 mètres ; étage circalittoral) et le début des plaines abyssales (-2.000 à -4.000 mètres ; étage abyssal).

orientation considering current	orientation par rapport au courant	3		limite son exposition ; indifférent ; maximize son exposition

Tableau 15 : Descripteurs des répartitions des populations de taxons avec leurs modalités (mod.).

Trait [en]	**Trait [fr]**	**nb mod.**	**Modalities [en]**	**Modalités [fr]**
population density	densité peuplement	3	rare individuals or isolated population; individual or dispersed sparsely populated, dense and / or frequent population	individus rares ou population isolée ; individu dispersé ou population clairsemée, population dense et/ou fréquentes
frequency	fréquence	3	1 or 10 or 100	1 ou 10 ou 100
coverage	recouvrement	3	very partial; frequent but discontinuous; continuous for several meters	très partiel ; fréquent mais discontinu ; continu sur plusieurs mètres
distribution area	aire de répartition	inf	[region lists]	[liste régions]
abundance (under 100 points)	abondance (sur 100 points)	inf	% individus specie[X]; gender[X]; familly[X]	% individus espèce[X]; genre[X]; famille[X]
pourcentage coverage	pourcentage de couverture	inf	% coverage specie[X]; gender[X]; familly[X]	% coverage espèce[X]; genre[X]; famille[X]
total perimeter for a taxa	périmètre total du taxon	inf	total perimeter of specie[X]; gender[X]; familly[X] in cm	total des périmètres de l'espèce[X]; genre[X]; famille[X] en cm
nb different individus	nb different individus	inf	total number of individuals for specie[X]; gender[X]; family[X] on the photo	nombre total des individus espèce[X]; genre[X]; famille[X] sur la photo

Tableau 16_1 : Descripteurs de répartition des populations de taxons avec leurs modalités (mod.).

Trait [en]	Trait [fr]	nb mod.	Modalities [en]	Modalités [fr]
relative abundance	abondance relative	inf	% individus for specie[X]; gender[X]; family[X] / % individus for specie[Y]; gender[Y]; family[Y]	% individus espèce[X]; genre[X]; famille[X]/ % individus espèce[Y]; genre[Y]; famille[Y]
relative percentage coverage	pourcentages de couverture relatifs	inf	% coverage specie[X]; gender[X]; family[X] / % coverage for specie[Y]; gender[Y]; family[Y]	% couverture espèce[X]; genre[X]; famille[X]/ % couverture espèce[Y]; genre[Y]; famille[Y]
relative percentage between two taxa	pourcentages de périmètre entre deux taxons	inf	% perimeter specie[X]; gender[X]; family[X] / % perimeter for specie[Y]; gender[Y]; familly[Y]	% périmètre espèce[X]; genre[X]; famille[X]/ % périmètres espèce[Y]; genre[Y]; famille[Y]
relative perimeter between two taxa	importance relative du périmètre	inf	sum of perimeter specie[X]; gender[X]; family[X] / sum of perimeter for specie[Y]; gender[Y]; family[Y]	somme périmètre espèce[X]; genre[X]; famille[X]/ somme des périmètres espèce[Y]; genre[Y]; famille[Y]

Tableau 16_2 : Exemples de descripteurs de classification des populations de taxons avec leurs modalités (mod.).

Trait [en]	Trait [fr]	nb mod.	Modalities [en]	Modalités [fr]
	typologie d'associations Eunis présentes		[référentiel eunis]	[référentiel eunis]
	typologie faciès SPN majoritaire présent		[référentiel MNHN]	[référentiel MNHN]
	typologie faciès SPN présents		[référentiel MNHN]	[référentiel MNHN]

Tableau 17 : Descripteurs des contextes sous forme de "facteurs abiotiques" des sites étudiés et/ou des transects et/ou des observations avec leurs modalités (mod.).

Trait [en]	**Trait [fr]**	**nb mod.**	**Modalities [en]**	**Modalités [fr]**
slope	pente	4	flat; inclined; vertical; under overhanging??	plat ; incliné ; vertical ; sous surplomb
drop shadow	ombre portée	2	yes / no	oui ; non
orientation	orientation	8	N, NE, E, SE, S, SW, W, NW	N, NE, E, SE, S, SO, O, NO
current exposure	exposition au courant	3	current always weak; large intermittent currents; often or always important currents	courant toujours faibles ; courants importants intermittents ; courants souvent ou toujours importants
exposure to the feather of a watercourse	exposition à la plume d'un cours d'eau	3		eutrophisation négligeable ; eutrophisation importante intermittente; eutrophisation souvent ou toujours importante
temperature	température	inf	T° in C°	T° en C°
salinity	salinité	inf	in g / m3	en g / m3
pH	pH	inf	numeric [0 - 14]	numérique [0 - 14]

Résultats concernant l'étude de la variabilité due à l'observateur

Tableau 18 : certaines des variables présentées dans les paragraphes précédents sont plus ou moins sensibles aux changements d'observateurs, sans entraînement ou après entraînement. Les principales variables relevées par les observateurs sont indiquées dans le tableau ci-dessous (issues de David et al., 2014d), ainsi que les méthodes rejetées car entraînant plus de variabilité entre les observateurs pour une même observation avec entraînement (E) ou idéalement sans entraînement (les méthodes étaient comparées avec au moins 3 essais par observateurs et 4 observateurs différents, et lorsque des entraînements étaient nécessaires, l'observateur pouvait s'entraîner pendant une plongée) :

Metric	Possible values Accepted	Possible values Rejected	Signification
Time	Digital positive integer	Time before and after the measurement	Time in minutes on the timer
Orientation	Front of the wall N, NE, E, SE, S, SW, W, NW	N, S, E, W (not enough precise) / orientation in degrees / exposition	North, Northeast, East, Southeast, South, Southwest, West, Northwest
Inclination (E)	V, I, F, C	Slope in degrees	Vertical, Inclined, Flat, Ceiling
Roughness	0, +, ++, +++	None, Not plane, Small infractuosity, Boulders, cave	No roughness, Low roughness (fist), average roughness (head), large roughness (shoulders)
Upper stratum (E)	CR, Esp, EC, ES, PC (3 max)	*More species were too much difficult to Write on a simple table and took too much time*	*Corallium rubrum, Erect sponge, Eunicella cavolinii, Eunicella singularis, Paramuricea clavata*
Basal stratum (E)	EnRA, EnGA, ErRA, ErGA, Bryo, Cod, Sp, Pey, Turf (3 max)	*More species were too much difficult to Write on a simple table and took too much time*	Encrusting red algae, Encrusting green algae, Erect red algae, Erect green algae, Bryozoan, Codium spp, Sponge, Peysonnelia spp, Turf
	HT, LP, LC, MA, PA (3 max)	*More species were too much difficult to Write on a simple table and took too much time*	*Halimeda tuna, Leptosamia pruvotii, Lithophyllum cabiochae, Mesophyllum alternans, Parazooxanthus axinellae*
Remarkable species (E)	According to diver's knowledge	Only if time is not too much consumed	The diver must specify his knowledge fields on the form
Remarkable stands	According to diver's knowledge	Not more than 1 word	The diver must specify his knowledge fields on the form
Solid waste	Nb objects	Sizes of objects	The diver must precise the object's type and its size (50 cm, 1 m, several metres…)

Ces test ont donné le résultat présenté dans le tableau 18 et ont permi de proposer la première version des protocoles d'observation en plongée de CIGESMED. Par exemple, la technique préconisée pour l'orientation des parois est de se mettre face à la paroi (et donc de noter l'inverse de l'exposition que l'on transforme ensuite en exposition dans le jeu de données). Tourner le dos à la paroi engendre une grande variabilité entre plongeurs pour les relevés d'orientation de paroi et ceci même avec de l'entraînement. À l'usage, certains plongeurs préféraient la noter en degrés, face à la paroi, mais ils prenaient alors moins bien en compte la totalité du segment en compte, ce qui induit aussi plus de variabilité pour le même relevé entre deux plongeurs.

Ce sont les relevés de recouvrements pour les strates hautes et basses qui ont été les plus difficiles à fixer, et qui demandent le plus d'entraînement. Pour cela, il a été décidé de se limiter à chaque fois à 3 observations maximum pour les strates inférieures et supérieures

Concernant la prise de vue, les tests ont montré qu'il est possible avec de l'entrainement de réaliser en une seule plongée (de 15 min), l'échantillonnage de deux transects de 10 m chacun, soit 40 quadrats photographiés.

Résultats concernant l'étude de la variabilité due à l'échantillonnage

Les résultats sont présentés dans la figure 25. Il montre qu'au niveau du phylum, les deux méthodes donnent des résultats équivalents. Les différences d'effectifs ne sont significatives que dans le phylum de Porifera. Ainsi, les résultats sont comparables.

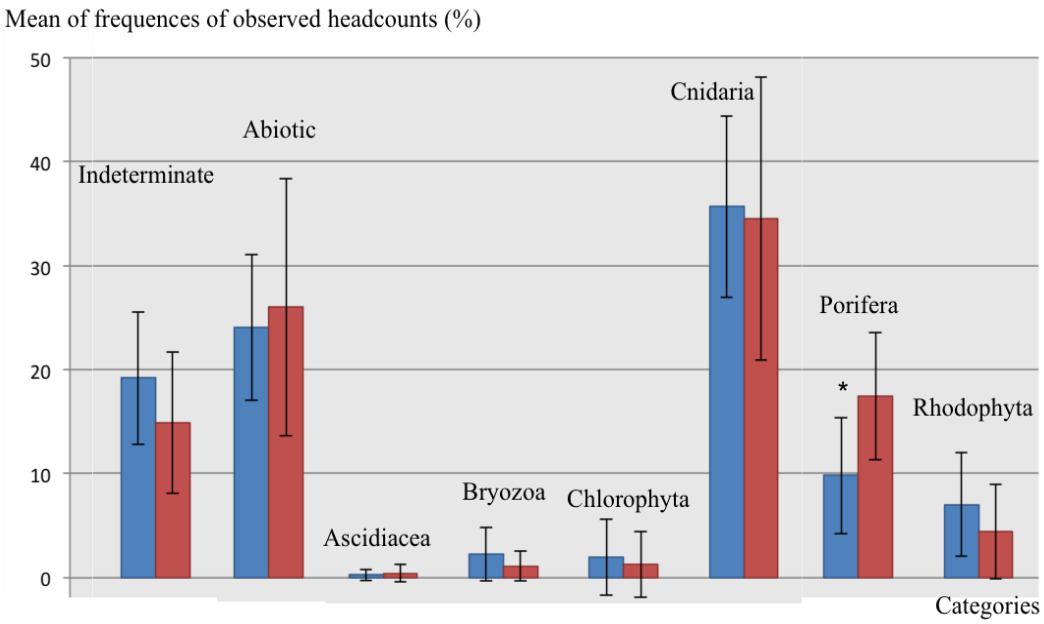


Figure 25 - Comparaison des résultats obtenus par la méthode du transect permanent (bleu) et la méthode des patchs aléatoires (rouge). * p<0.05 selon le test de Mann-Whitney-Wilcoxon avec un risque de 5%. (Thierry de Ville d'Avray, 2014)

Résultats concernant l'étude de la variabilité opérateur

Les résultats sur certaines catégories sont indiqués sur la figure 26. Ils montrent qu'après un seul échange entre les trois opérateurs, le novice améliore beaucoup sa capacité d'identification. Pour certaines catégories, le niveau de connaissance des trois opérateurs est rapidement homogénéisé : par exemple pour les catégories *Cnidaria* et *Porifera*. Dans la plupart des cas, ils sont très difficiles à identifier sur la photographie. De ce travail, il ressort que les espèces de l'embranchement des cnidaires sont les plus faciles à identifier sur la photographie par les débutants, alors que les espèces de l'embranchement des éponges sont pour la plupart très difficiles à identifier sans formation spécifique, *a fortiori* sur photographie.

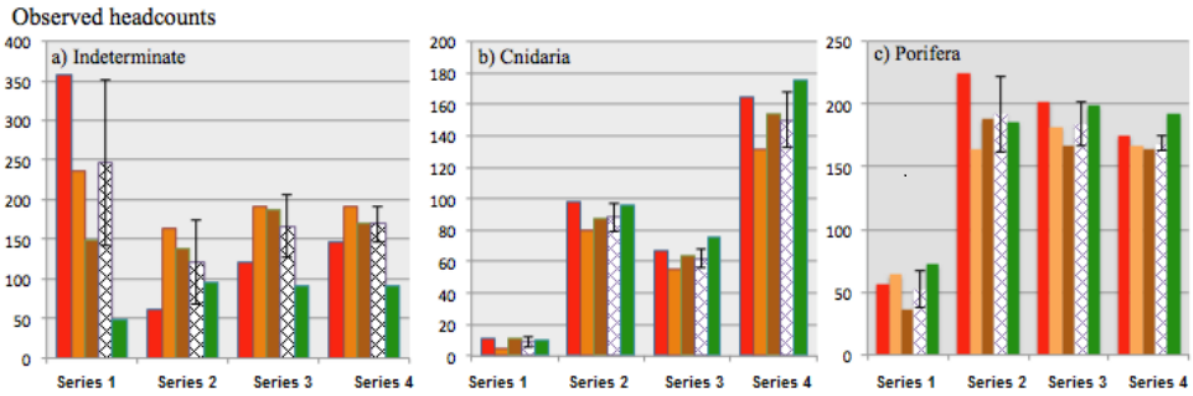


Figure 26 - Comparaison des résultats observés par différents opérateurs : un novice (en rouge) et deux expérimentés (orange et brun). La moyenne et l'écart type entre les trois opérateurs sont en blanc. Les identifications validées par les trois opérateurs sont en vert (Thierry de Ville d'Avray, 2014).

Résultats concernant la variabilité technique du système d'observation

La figure 27 montre qu'au niveau de l'espèce, les deux caméras donnent des résultats équivalents. Mais la caméra de haute qualité a permis de réduire le nombre de "indéterminé" et ces observations ont été attribuées à d'autres catégories, en majorité au niveau du genre.

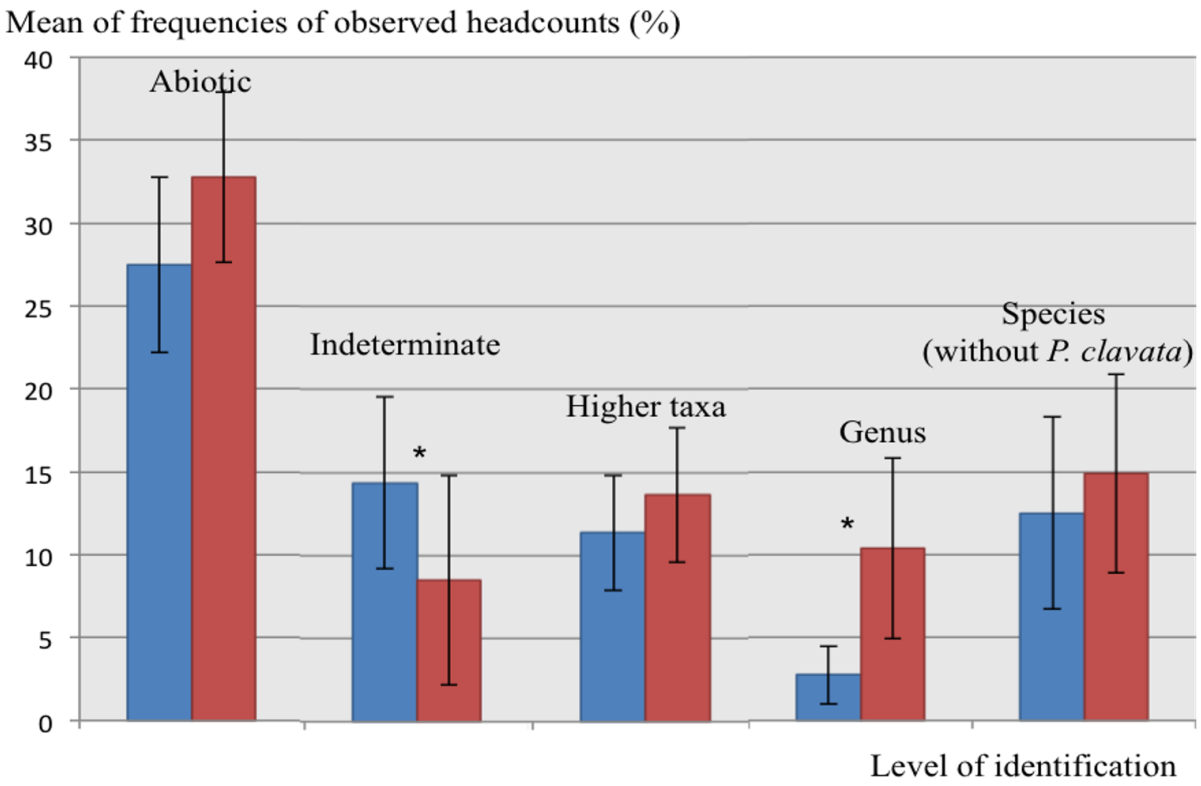


Figure 27 - Comparaison des résultats obtenus avec une caméra de qualité moyenne (bleu) et une caméra haute qualité (rouge). L'étoile marque une différence significative selon le test de Mann-Whitney-Wilcoxon avec un risque de 5%. (Thierry de Ville d'Avray, 2014)

3.2 Résultats concernant la contextualisation reposant sur la cartographie du coralligène

L'utilisation de la classification ascendante hiérarchique (CAH) sur toutes les données traitées a permis (i) de regrouper les espèces en fonction des valeurs d'orientation, de pente et de rugosité, et (ii) de regrouper les paramètres de profil selon les espèces par segment. Cette CAH montre quatre groupes de paramètres de profil (Figure 28).

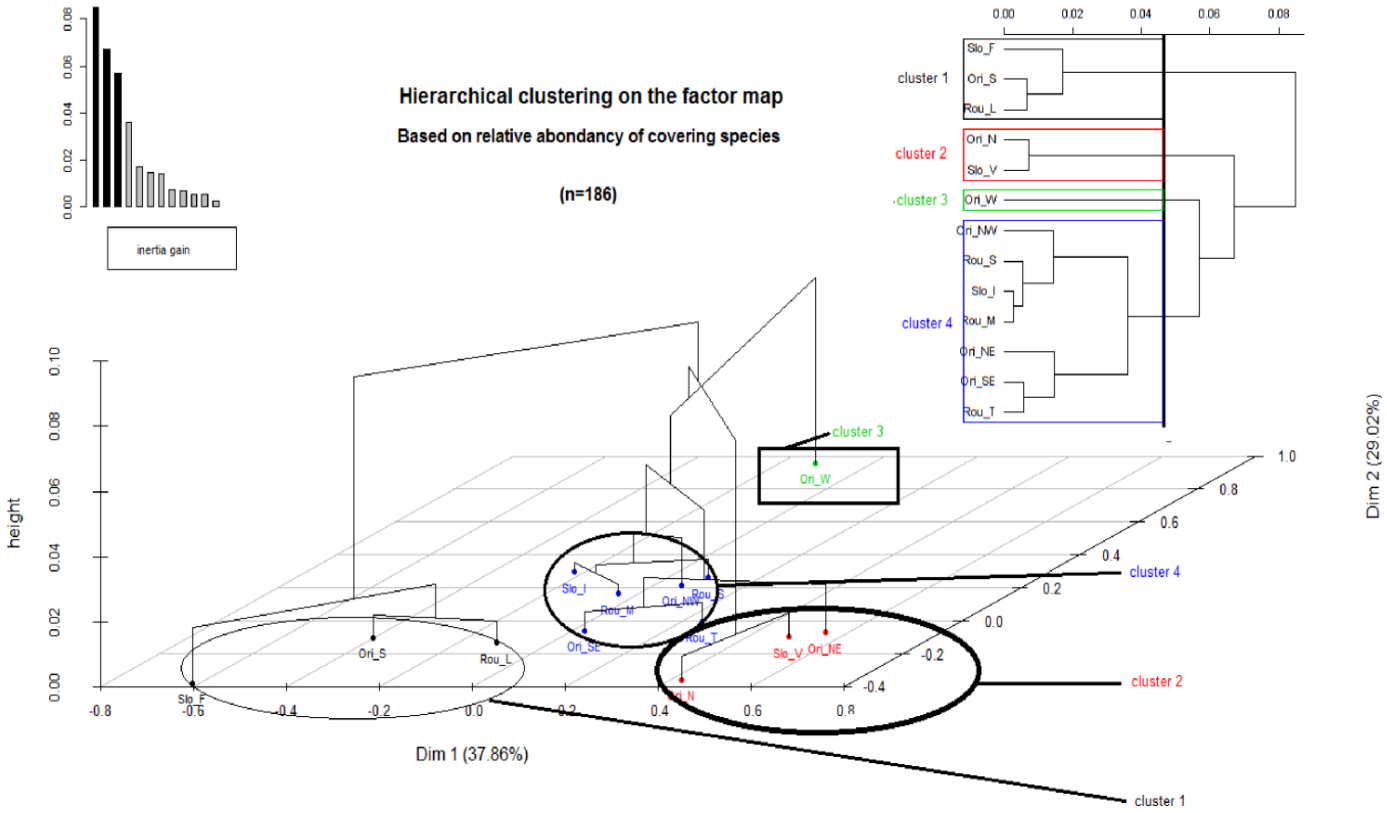


Figure 28 : La classification ascendante hiérarchique à deux dimensions montre une bonne répartition des assemblages d'espèces selon un premier gradient (37,86% de l'explication de la variabilité) avec deux clusters opposées (cluster 1 et cluster 2 sur les figures), et un deuxième gradient pour les valeurs de chaque catégorie.

3.3 Résultats concernant l'analyse d'image dans le cadre de CIGESMED

Les résultats présentés concernent uniquement les suivis effectués autour de Marseille dans le cadre de CIGESMED car nous avions accès à de nombreux paramètres de contextes pour les sites[82]. Les sites ayant les relevés les plus précis ont été utilisés pour illustrer ce travail de thèse. Les premiers sets d'analyse d'image ont permis de fixer la méthode la plus efficace en terme d'effort de prospection et de comparer les qualités et défauts de différentes options (pour rappel : 25 cm sur 25 cm et 50 cm sur 50 cm, l'appareil photo professionnel Nikon D300s, avec l'appui de 2 phares de 600 lumens sur bras articulés, et la caméra "GoPro hero 3 black" deux bras articulés équipés de phares sola 600 light and motion et les deux intensités lumineuses différentes). Nos collègues de Turquie étaient de prime abord plus favorable à l'utilisation d'un petit cadre, ceci permettant une photo plus précise et proche de la paroi et un moindre encombrement en plongée. En fin de compte, ce sont les quadrats de 50 cm sur 50 cm finalement qui ont été choisis. Il s'est avéré d'une part que les sites ayant de grandes espèces dressées représentaient une trop grande part des surfaces de 25 cm sur 25 cm, et que l'étagement lié au relief y était mal échantillonné. C'est aussi non seulement pour cette raison d'étagement mais aussi parce qu'il est parfois difficile de trouver un habitat coralligène continu sur 10 mètres sur tous les sites que les patchs carrés ont été préférés aux transects linéaires. Les transects linéaires non continus ont été abandonnés car ils impliquent que le plongeur choisisse s'il est encore dans un habitat coralligène ou pas.

Les images prises avec la GoPro s'avèrent exploitables au même niveau que les images de meilleure qualité de l'appareil professionnel Nikon à de rares exceptions près pour le niveau de détermination obtenu, même si la moindre netteté engendre un moindre confort pour l'opérateur chargé d'analyser les images. L'appareil professionnel Nikon a néanmoins permis de réduire les zones non exploitables (par mauvaise netteté ou éclairage insuffisant). Enfin, et sans surprise, un éclairage dégradé entame fortement l'exploitabilité de la surface de la photo (jusqu'à 60% de points indéterminés). La comparaison de deux sets de photos analysées avec 100 points d'une part, puis 250 points n'a pas montré de différences significatives entre les fréquences relatives d'espèces, à part pour quelques taxons de petite taille et de faible fréquence. Cette différence n'a pas montré un intérêt suffisant pour multiplier le temps d'analyse des photos par 2,5.

[82] L'exploitation globale des données et la comparaison entre régions marine, à la charge de l'équipe Turque responsable du WP2 de CIGESMED, est en cours de finalisation.

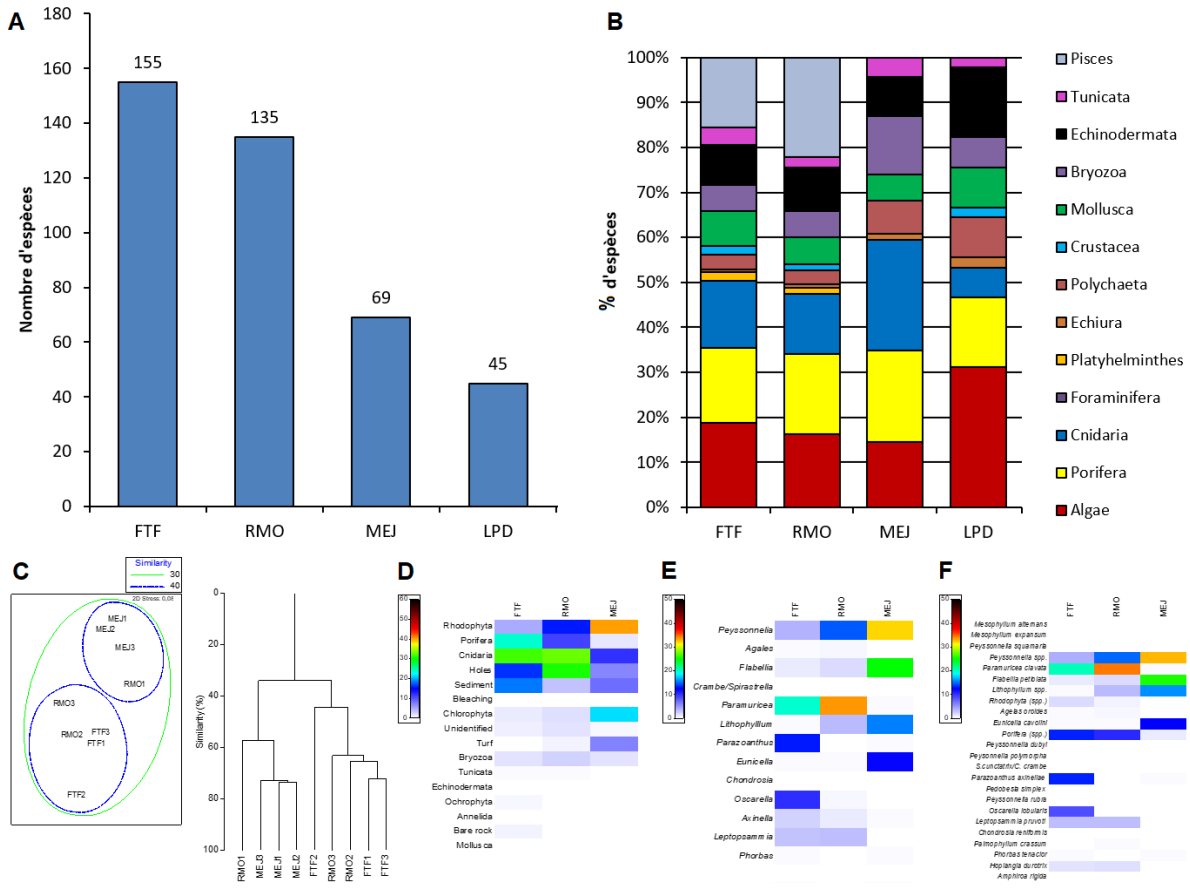


Figure 29 : nombre de taxons identifiés différents (A) et répartition par groupe taxonomique des taxons identifiés (B) sur les sites FTF, RMO, MEJ et LPD en plongée. Les sites FTF et RMO sont 3 fois plus riches que MEJ (site d'exercice), et LPD qui est en dehors du Parc National des Calanques. Le nombre d'espèces par groupes est comparable entre les sites, surtout pour les groupes les mieux représentés. En C, Une n.M.D.S. et une CAH ont été réalisées en comparant les fréquences relatives de taxons (distances de Bray-Curtis) à partir de l'analyse de 3 sets de 9 quadrats photo réalisés sur des profils verticaux concernant 3 stations de CIGESMED (FTF, MEJ et RMO). Les figures D, E et F représentent la structure des peuplements en fonction des groupes taxonomiques (D), Genres (E), ou espèces (F) entre les trois sites. (Féral et al., 2016)

Le nombre d'espèces par site mesuré par annotation en plongée varie de 45 à 155 (Figure 29 A). La proportion d'espèces pour chaque groupe taxonomique est comparable, surtout pour les groupes d'espèces majoritaires pour les trois stations (*Cnidaria*, *Porifera*, *Algae*) qui représentent à eux trois près de 50% des recouvrements sur tous les sites (Figure 29 B).
Les classifications (Figure 29 C) montrent qu'une seule des trois stations contient des sets photos qui sont proches (MEJ) alors que deux des sets photos de RMO sont plus proches de deux sets de photos de FTF que le troisième de FTF, et qu'un set de photos du site RMO est plus proche des stations MEJ. Les figures 29 D, 29 E et 29 F montrent que la structure

des peuplements en fonction des groupes taxonomiques (D), Genres (E), ou espèces (F) diffèrent significativement entre les trois sites, même si FTF et RMO ont des assemblages d'espèces plus proches pour ces trois niveaux de comparaison (Fi). Concernant de possibles détections d'impact, des dépôts de sédiments ont été trouvés dans toutes les stations, mais leur pourcentage de couverture était supérieur à 15% dans une station (FTF). Le blanchiment des algues n'a été enregistré dans aucune des trois stations.

3.4 Résultats concernant l'analyse d'image dans le cadre de DEVOTES

Tableau 19 : la proportion de taxons identifiés au niveau de l'espèce était généralement faible et les mollusques étaient les plus faciles à déterminer :

Number of identified taxa	Specie level	Other levels	Total
Annelida	5	15	20
Bryozoa	5	14	19
Cnidaria	1	6	7
Crustacea	2	2	4
Mollusca	8	2	10
Colonial *Tunicata*	2	0	2
Non-colonial Tunicata	1	3	4
CA	1	3	4
Rhodophyta	2	5	7
Chlorophyta	1	1	2

La liste de tous les taxons initialement identifiés dans chaque mer (c'est-à-dire avant de le remplacer par un rang taxonomique plus élevé pour avoir une comparabilité des mers) est disponible dans l'annexe 4. Les données brutes utilisées pour les analyses statistiques des communautés correspondent au nombre de points attribués à chacune des catégories taxonomiques remplacées parmi les six faces de plaque de tous les ARMS (annexe 4). Les résultats présentés ci-dessous ont été obtenus avec les données d'abondance en regroupant les catégories « non vivantes » et « indéterminées ». Une analyse les excluant (c'est-à-dire

les abondances non transformées et sans les catégories « non vivant » et « indéterminé ») a donné des résultats similaires.

En moyenne, le pourcentage de colonisation d'ARMS par une catégorie taxonomique identifiable variait selon les mers de 50% dans la Mer Adriatique et 60% dans la Mer Rouge à plus de 70-75% dans le golfe de Gascogne et la Mer Méditerranée. Il n'y avait pas de patron cohérent entre les faces supérieures et inférieures des mers pour la plaque 1 (la comparaison n'était pas possible pour les plaques 4 et 8 à cause de la compartimentation des faces inférieures, mais pas des faces supérieures). Dans l'Adriatique et la Mer Rouge, les faces inférieures (P1B) étaient plus colonisées que les suivantes (P1T), avec 72% et 52% pour la Mer Adriatique et 84% et 71% pour la Mer Rouge, mais les différences étaient petites (et inversées) dans les deux autres mers.

La colonisation biologique a été sous-estimée en raison des points indéterminés (en moyenne 2% des points étaient indéterminés dans AdS, 4% dans ReS et BoB, et 7% dans NWM). Les groupes les plus abondants représentés sur les plaques d'ARMS (Figure 30) étaient *Annelida*, *Bryozoa*, *Porifera* et *Mollusca* pour les animaux, ACC et autres *Rhodophyta* pour les algues (Tableau 19). Des groupes comme *Tunicata* (*Tunicata* coloniale et non coloniale) et Cnidaria pour les animaux et *Chlorophyta* et *Phaeophyta* pour les algues étaient beaucoup moins abondants ou répandus (et leur identification était difficile sur les photographies au-delà du phylum). Les animaux représentaient la plus grande partie de la colonisation avec jusqu'à 54% pour le Nord-Ouest de la Méditerranée, 48% pour le golfe de Gascogne, 38% pour la Mer Adriatique et 32% pour la Mer Rouge, alors que les algues représentaient plus de 20% colonisation totale dans le Nord-Ouest de la Mer Méditerranée et le golfe de Gascogne, 11% pour la Mer Adriatique et 27% pour la Mer Rouge. Dans toutes les mers, l'abondance relative des groupes les plus abondants différait entre les faces supérieures et inférieures des plaques : Annelida et Bryozoa étaient plus fréquents sur les faces inférieures alors que ACC et autres Rhodophyta préféraient les faces supérieures (figure 30).

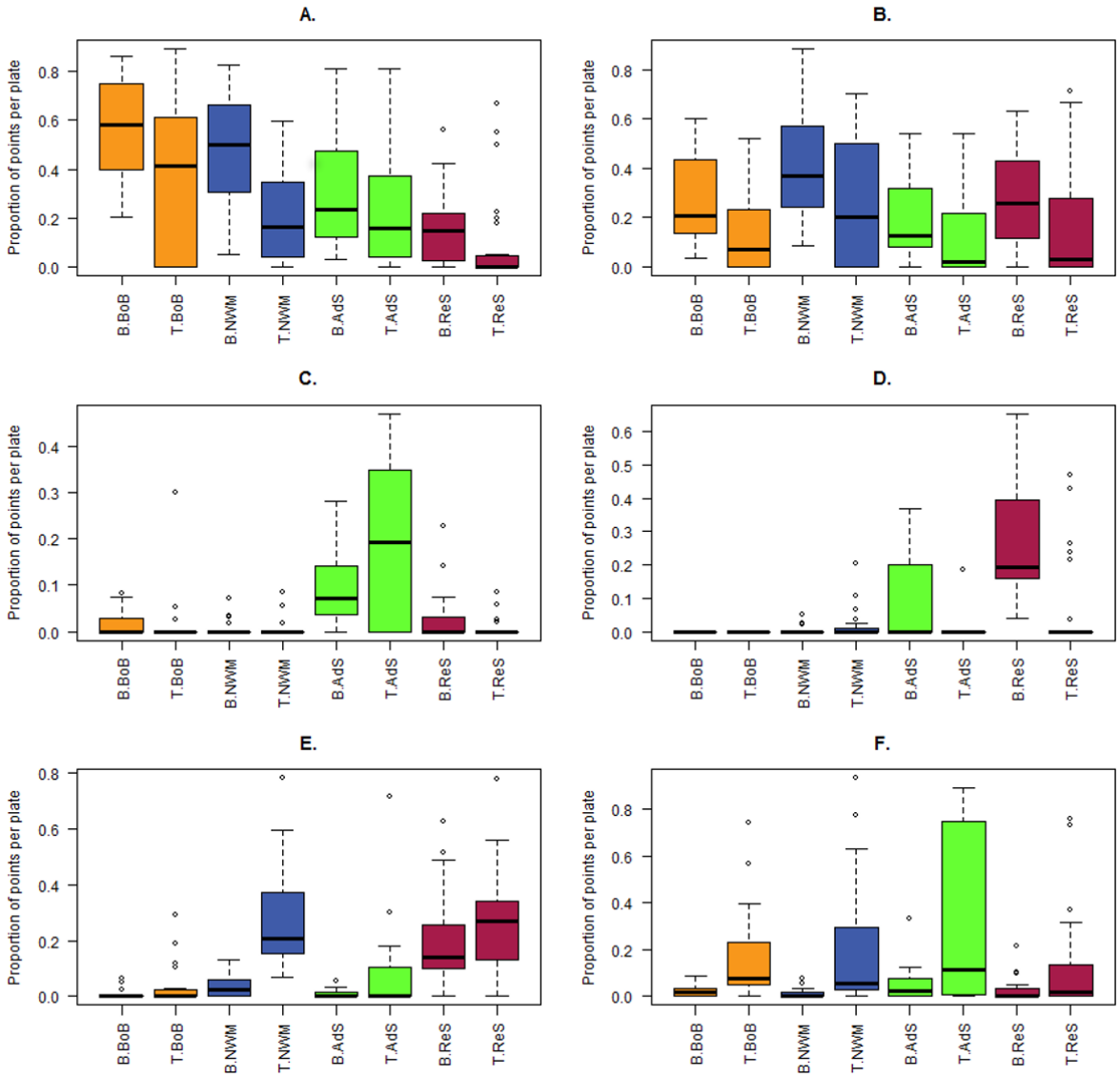


Figure 30 : Boxplots montrant l'abondance des principaux groupes d'animaux et d'algues dans les ARMS des quatre régions de la mer. Annelida (A), Bryozoa (B), Mollusca (C), Porifera (D), Algues corallines calcaires (E), autres Rhodophyta (F), sur ARMS de quatre mers. Les mers sont ordonnées d'Ouest en Est et nommées comme dans le Tableau 1. Les plaques supérieures (T) et inférieures (B) ont été analysées séparément.

En général, les fréquences des organismes installés (logs transformés en abondance) ont montré des différences significatives entre les mers, mais pas entre les sites de chaque mer (Tableau 20). En particulier, les annélides et les bryozoaires étaient plus abondants dans le golfe de Gascogne et dans le Nord-Ouest de la Méditerranée que dans la Mer Adriatique et la Mer Rouge. Les mollusques étaient plus abondants dans la Mer Adriatique et les Porifera étaient particulièrement abondants dans l'Adriatique et la Mer Rouge (en particulier sur les

faces inférieures des plaques), tandis que les Algues Corallines (ACC) étaient plus abondantes dans le Nord-Ouest de la Méditerranée et dans la Mer Rouge (Figure 30). Pour les faces supérieures et inférieures des plaques, les mers pourraient être différenciées sur la base des abondances de ACC, Mollusca et, malgré leur faible abondance, Phaeophyta. Seul Chlorophyta dans les plaques supérieures a révélé une variabilité intra-mer significative.

Tableau 20 : Valeurs-P pour chaque taxon des ANOVA imbriquées (4 régions, 3 sites dans chaque région) en fonction des abondances log-transformées et écrites en caractères gras lorsqu'elles sont significatives (<0,05). ACC : Algues Corallines encroûtantes Calcaires

	Top Faces		Bottom Faces	
Taxon	P-value Sea	P-value Site	P-value Sea	P-value Site
Annelida	**0.029**	0.197	**<0.001**	0.997
Bryozoa	**0.031**	0.331	**0.032**	0.544
Mollusca	**<0.001**	0.615	**0.009**	0.588
Porifera	0.163	0.115	**<0.001**	0.993
ACC	**<0.001**	0.103	**0.008**	0.737
Other Rhodophyta	**0.009**	0.961	0.094	0.904
Chlorophyta	0.661	**<0.001**	**0.029**	0.677
Phaeophyta	**0.005**	0.683	**0.005**	0.947

Les PERMANOVA et ANOSIM réalisées sur l'ensemble des données ont révélé des effets hautement significatifs de la mer, du site (imbriqués dans la mer) et des faces des plaques sur la composition de la communauté (voir David et al., 2018 *submitted in Marine Pollution Bulletin*) pour les résultats détaillés de PERMANOVA. Il y avait aussi un effet hautement significatif de l'interaction entre la mer et la face, et entre le site (imbriqués dans la mer) et la face, mais celui-ci est moins important.

Toutes les comparaisons par paires des mers prises deux à deux (two-way crossed design of plate face and sea) étaient hautement significatives. PERMDISP n'a pas montré de différences significatives concernant la dispersion entre les mers, ni entre les sites. Toutes les comparaisons de faces appariées montrent des différences hautement significatives dans la composition de la communauté, sauf entre les faces P4T et P8T (non significative, P = 0.1483) et entre les données non transformées des deux faces compartimentées P4B-P8B (données transformées de la quatrième racine, P = 0.0118). La dispersion était fortement différenciée entre les faces des plaques (valeurs P de PERMDISP = 0,0001) et les tests PERMDISP par paires indiquent trois niveaux de dispersion : faces très dispersées (P1T, P8T), faces modérément dispersées (P4T, P1B) et faces moins dispersées (P4B, P8B) (valeurs détaillées non montrées).

La Non-Metric Multidimensional Scaling (N.M.D.S.) (Figure 31) illustre l'homogénéité des communautés des plaques P1T par rapport aux autres dans chaque mer et, bien que moins clairement et pas pour toutes les mers, les différences entre les faces supérieures et inférieures. Les groupes d'algues (*Chlorophyta*, *Phaeophyta*, autres Rhodophytes et autres algues, mais pas CCA) ont tendance à être plus abondants sur les faces exposées (P1T) dans les quatre régions, comme le reflète la position de ces variables sur le tracé nMDS (S1 Fig de la publication David *et al*, 2018, en annexe 4.4). En revanche, tous les points représentant les animaux sont éloignés des échantillons P1T dans le tracé nMDS.

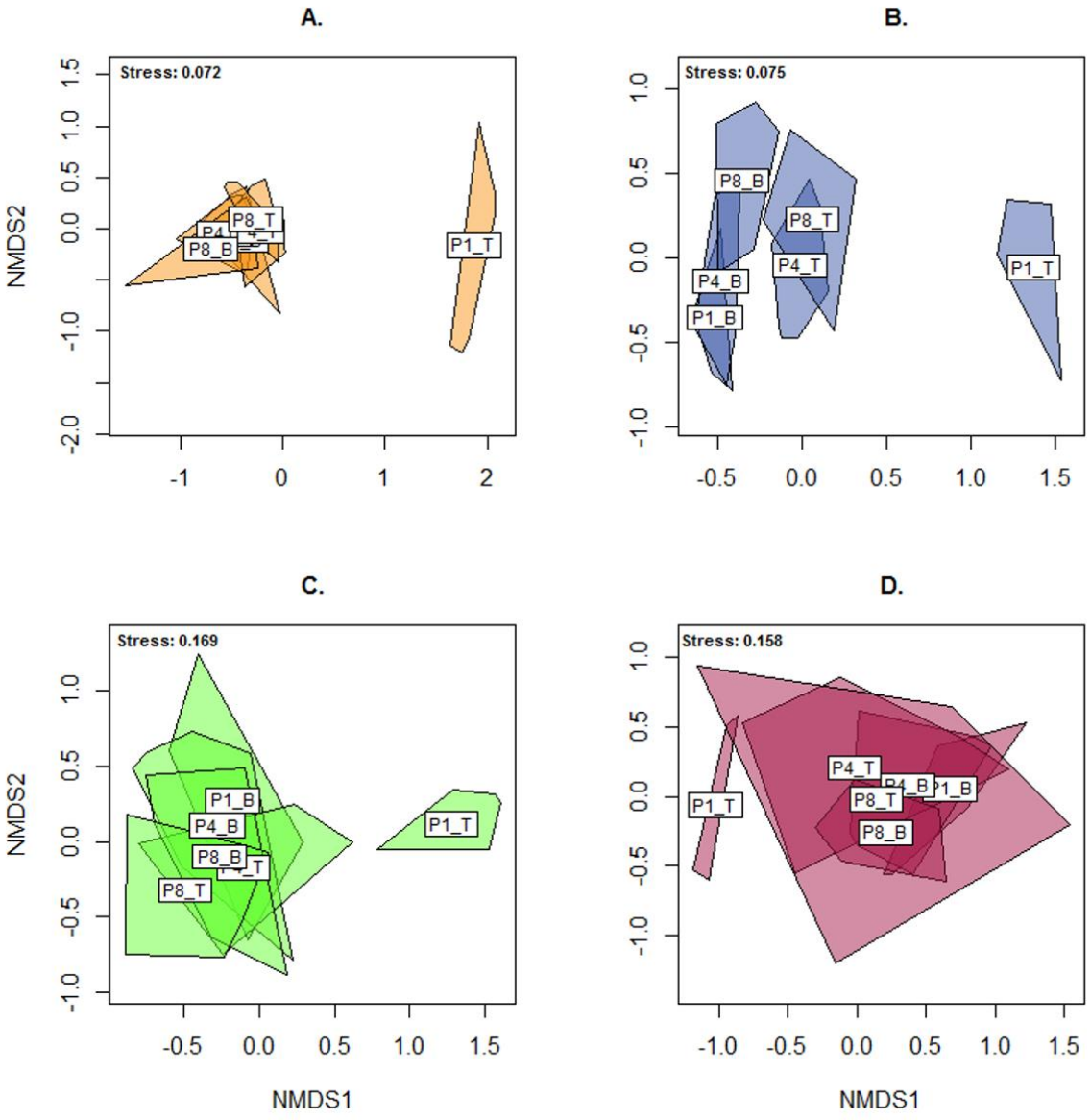


Figure 31 : N.M.D.S. (Non-metric Multi Dimensional Scaling representations) de la communauté pour chaque face dans chaque région. A : Golfe de Gascogne, B : Méditerranée du Nord-Ouest ; C : Mer Adriatique, D : Mer Rouge. Chaque point représente un ARMS pour une face donnée, la valeur de stress est indiquée.

Pour chacune des six faces analysées séparément, il y avait un effet très significatif de la mer et du site (Tableau 19) avec la seule exception du P8T pour lequel le site n'était pas significatif. Cela s'explique probablement par des données manquantes pour deux ARMS dans le golfe de Gascogne alors que des photographies étaient disponibles pour toutes les autres faces de ces ARMS. Lorsqu'elles sont analysées séparément dans chaque mer, les deux faces P1T et P4B (individuellement) permettent de détecter les effets significatifs des sites sauf dans la Mer Rouge (Tableau 5) mais le nombre de permutations distinctes est

réduit dans ces petits ensembles de données d'au plus 9 échantillons (trois ARMS dans chacun des trois sites).

Tableau 21 : Résumé des résultats PERMANOVA pour chaque plaque. P Value pour les régions et les sites dans les régions. Les particularités de la plaque sont entre parenthèses (le dessus de la plaque 1 est exposé à l'environnement extérieur, et les plaques 4 en bas et 8 en bas sont compartimentées par une croix centrale). Toutes les valeurs-P sont hautement significatives sauf pour P8T (il y a deux images manquantes, pour P8T, les deux dans BoB).

Plate face	Sea region (P-value)	Site (P-value)
P1T (exposed)	**<0.001**	**<0.001**
P4T	**<0.001**	**0.004**
P8T	**<0.001**	0.749
P1B	**<0.001**	**<0.001**
P4B (comp.)	**<0.001**	**<0.001**
P8B (comp.)	**<0.001**	**0.002**

Tableau 22 : Résultats de PERMANOVA unidirectionnel testant l'effet de sites pour deux faces de plaque (en colonnes): Les valeurs-P correspondant à l'effet de site sont reportées, écrites en gras lorsqu'elles sont significatives (P <0,05). La nomenclature des faces de plaques est expliquée dans la Figure 23.

Sea region	P1T (P-value)	P4B (P-value)
Bay of Biscay	**0.047**	**0.047**
Northwestern Mediterranean	**0.007**	**0.047**
Adriatic Sea	**0.035**	**0.024**
Red Sea	0.670	0.051

Parmi les 7 facteurs environnementaux booléens qui pouvaient être testés dans plusieurs mers, quatre étaient significatifs, généralement avec une valeur-P entre 2% et 5%, mais l'interaction du facteur environnemental avec la mer avait un effet beaucoup plus significatif (généralement valeur-P < 0,01 ou 0,001), suggérant que l'effet observé du facteur environnemental reflétait effectivement un effet du site (rappelons que l'effet du site dans la mer était hautement significatif avec la valeur-P la plus faible possible pour les 9999 permutations des données: 0,0001). Les facteurs significatifs (mais colinéaires avec les sites) étaient l'anthropisation globale, l'urbanisation, la pollution chimique et la présence d'un port. La boue à proximité, la production de sable et d'eaux usées à proximité n'étaient pas significatives. En effet, l'effet du facteur « présence portuaire », avec les données transformées en racine quatrième, était plus significative (valeur-P = 0,0285) que l'interaction mer x port (valeur-P = 0,0541 NS); ce résultat n'a néanmoins pas été obtenu avec des données d'abondances non transformées ou des racines carrées.(le facteur "présence d'un port" est devenu non significatif, alors que l'interaction des facteurs " présence d'un port x mer" est devenue significative).).

Pour les facteurs environnementaux qui varient seulement dans une mer, la seule façon d'étudier leur effet potentiel sur la composition de la communauté était de vérifier si le site qui a été distingué par rapport au facteur environnemental booléen était aussi le plus contrasté avec l'effet du facteur site. Globalement, nous avons obtenu 6 correspondances significatives sur 20, ce qui est très proche de la proportion attendue (20/3) sous l'hypothèse nulle qu'il n'y a pas d'effet des facteurs environnementaux sur la composition de la communauté (Tableau 20). Nous n'avons donc aucune preuve fiable de l'influence d'un

facteur environnemental sur la composition de la communauté, probablement due à notre design expérimental avec peu de sites par mer. Aucune variation dues aux facteurs environnementaux dans les sites n'a pu être mise en évidence alors que nous avons de forts effets des facteurs mer et site.
Dans la PERMANOVA imbriquée avec les facteurs mer et site (par mer), mais sans séparer les faces de plaques d'ARMS, la mer et les sites sont restés significatifs

4. Discussions et perspectives concernant l'efficacité des outils, méthodes et protocoles

L'observation *in situ* est contrainte par un certain nombre de facteurs. Les capacités d'observation d'une plongée scientifique dépendent en premier lieu de la profondeur à laquelle les plongeurs opèrent (le temps de plongée pour la première plongée sans avoir à mettre en œuvre de palier autre que celui de sécurité est de 35 minutes à 21 mètres, 15 minutes à 30 mètres, 7 minutes à 42 mètres dans les tables de plongée du Ministère du Travail M.T.92). La deuxième plongée, s'il y en a une, doit intégrer des pénalités qui dépendent des paramètres de la première plongée et de l'intervalle entre deux plongées.
En second lieu, le matériel utilisé (capacité des "blocs" de plongée, encombrement du plongeur avec les quadrats et/ou appareils photographiques, caméras, sacs ou portoirs à échantillons) influence la consommation en air, et un peu de courant ou une charge trop lourde peuvent nécessiter de raccourcir le temps de plongée. La prise en compte de ces paramètres est primordiale lors de la conception des protocoles. En complément, la formation, l'entraînement des plongeurs et la préparation de la plongée sont des facteurs très importants pour réaliser un échantillonnage ou une prise de données de qualité. Pour les améliorer, les débriefings post-plongée et la saisie des données aussitôt que possible après la sortie de l'eau ont été mis en œuvre pour tous les aspects de ce travail.

4.1 Discussion concernant l'intercalibration

L'intercalibration des observateurs en plongée

Concernant les observateurs en plongée, la formation taxonomique et la pratique de plongées sur cet habitat se sont révélées cruciales pour que les plongées soient réussies. Certains plongeurs scientifiques ont parfois aussi des problèmes de vision qui peuvent empêcher une bonne mise au point ou une erreur de reconnaissance lors d'un prélèvement. Mais au-delà de ces facteurs relevant de la compétence, d'autres aspects sont à prendre en considération.

Dans notre cas, au-delà de ces évidences, d'autres difficultés peuvent être amplifiées par l'utilisation à large échelle du protocole.

Lors des tests des différents protocoles, il s'est avéré que l'expérience des plongeurs d'une part, et leur entraînement d'autre part (et leur habitude à travailler ensemble) pouvaient influencer l'exploitabilité des quadrats. Quelques erreurs comme ne pas être bien perpendiculaire aux parois, ou prendre une image au moment où l'autre partenaire souffle son air peuvent rendre une grande partie de l'image non utilisable. D'autres sets de photos ont dû être écartés lorsque l'endroit choisi était trop dense en gorgones déployées. Lors de la manipulation d'un quadrat, certaines limites notamment la complexité du milieu engendrent certaines contraintes. La répartition qui est dite « aléatoire » des sets de quadrats doit se faire dans certaines zones de terrain permettant le placement correct de l'ensemble du set et ainsi permettre l'exploitation des données (autre exemple, si on place le quadrat central trop près d'un bord de paroi, il y aura impossibilité à placer les 9 quadrats). Ce discernement peut être mis à mal par la narcose, ou la fatigue liée à la présence d'un courant plus important, surtout lorsque l'on doit amener le quadrat sur le site en partant d'un peu loin (quadrat qui représente une résistance au courant à laquelle il faut s'habituer).

Dynamiques et efficacité de l'inter-calibration

L'évaluation et l'inter-calibration concernent l'analyse des assemblages benthiques coralligènes par observation directe, enquêtes photographiques / vidéo (directement influencées par les compétences et l'expérience des opérateurs). Ces approches, qui ont l'avantage d'être en principe non destructrices, permettent d'étudier les caractéristiques démographiques de base des principales populations d'espèces associées, en ayant pour objectif de détecter dans leur variation les perturbations et les menaces et de corréler ces variations avec d'autres facteurs biotiques (comme la présence d'espèces exotiques et/ou invasives) et abiotiques affectant les habitats coralligènes. Cela nécessite de comprendre l'effet de chaque facteur sur la variabilité des données.

Cette phase de caractérisation des sites par quadrats photo n'a été réalisée que dans la région de Marseille. Idéalement, elle devrait être reproduite dans d'autres localités et sur d'autres observateurs/opérateurs.

La qualité de l'ensemble de données « quadrats photo » souffre de la difficulté d'isoler correctement chaque variable étudiée. Ceci est principalement dû à la complexité des opérations sous-marines, mais cette inter-calibration peut être améliorée par le gain d'expérience.

L'étude des deux méthodes d'échantillonnage (transects linéaires ou aléatoires) montre que les observations effectuées au niveau d'identification d'un phylum sont comparables si elles sont effectuées selon l'une ou l'autre méthode, à part pour le groupe Porifera. Les espèces

de ce phylum sont difficiles à identifier dans le milieu, et globalement peu connues (avec une disponibilité faible des spécialistes). Néanmoins, ce sont des taxons abondants et d'une grande diversité ; ce groupe nécessiterait des tests d'inter-calibration plus poussés, car d'avis de spécialistes, même avec ce groupe difficile, des progrès (même in situ) sont tout à fait envisageables.

Dans l'état actuel de l'expérimentation, pour comparer les observations de Porifera d'un site à l'autre, il est recommandé de mettre en œuvre la même méthode dans les deux sites, car il est démontré qu'il pourrait y avoir une différence significative de comptes en fonction de la méthode appliquée. Comme la méthode du patch aléatoire est plus facile à exécuter, c'est celle qui a été recommandée dans le cadre du protocole CIGESMED.

En ce qui concerne l'effet de la qualité de l'appareil photo sur les observations faites, il a été montré que l'appareil de qualité moyenne est suffisant pour identifier autant de taxons différents que la caméra de haute qualité. La différence entre les deux est observée pour les espèces difficiles à différencier au niveau du genre. En effet, l'appareil photo de haute résolution permet d'identifier un niveau taxonomique plus précis pour certains individus qui étaient indéterminés ou moins précis avec la caméra de résolution moins bonne (en passant de la famille au genre, ou du genre à l'espèce par exemple, grâce aux critères d'identification qui peuvent ainsi devenir visibles). Étant donné que l'appareil photo de qualité moyenne est plus abordable pour les gestionnaires d'aire marine protégée (5 fois moins cher), cette solution peut être privilégiée par ceux qui ne disposent que de petits budgets.

L'inter-calibration formatrice

L'étude de l'influence du niveau de connaissances de l'opérateur montre que la discussion entre opérateurs permet aux novices d'améliorer rapidement leur capacité d'identification. La discussion est très efficace au début, puis les opérateurs atteignent une étape pendant laquelle aucun ou très peu de progrès sont observables et nécessitent une formation appropriée pour progresser, si une identification plus précise est nécessaire c'est-à-dire si elle est possible et utile dans le cadre des futures analyses.

Ces résultats prometteurs sont à renforcer en augmentant le nombre d'opérateurs participant. Le travail serait à prolonger avec des moyens d'envergure et sur le long terme, avec de nombreuses itérations et ajustements.

Inter-calibration, un défi dépendant de nombreux facteurs humains

Une difficulté généralement mal prise en compte est le facteur humain. L'objectif de cette étude de l'inter-calibration de méthodes opérationnelles en écologie est d'améliorer la fiabilité des connaissances sur la dynamique des habitats coralligènes et d'améliorer leur gestion en créant des protocoles et des outils reproductibles, à coût raisonnable.

Ce travail permet de mieux comprendre l'importance du niveau de compétence et de formation des opérateurs, de connaître l'impact du type de méthodologie d'échantillonnage employée et de comparer l'efficacité de la description de chaque méthode tout en améliorant l'accessibilité des informations et la facilité de mise en œuvre des protocoles.

Ces facteurs d'amélioration dépendent autant de la formation et de l'expérience de chacun que de sa culture et de ses motivations. Dans le cadre de CIGESMED, nous avons par exemple très rapidement observé des différences d'interprétations marquantes pour chaque protocole de la part des équipes turques, grecques et françaises. Parfois, celles-ci étaient liées à la langue et l'interprétation de l'anglais, parfois à la difficulté d'avoir une mesure de facteur effective sur un milieu trop différent (il n'y a pas de gorgones à 30 m en Grèce et en Turquie, les récifs à *Corallinaceae* y sont beaucoup moins denses et continus), et parfois encore aux habitudes déjà prises par les différentes équipes, qui résistent souvent à un changement dont elles ne comprennent pas la raison (par exemple, chacun souhaite garder son système de mesure expérimenté antérieurement pour avoir une comparabilité de ses propres données antérieures avec le programme innovant construit de manière commune).

4.2 Discussion concernant le protocole de cartographie

Qu'apporte la contextualisation basée sur la cartographie ?

L'hypothèse testée était que l'influence de certains paramètres relevés peut être difficile à mesurer tous contextes confondus, mais se montrer plus significative dans un contexte précis.

La cartographie nous a permis d'étudier séparément les données issues du traitement des photographies selon le type de profil. Cette approche laisse la possibilité de séparer les données pour les analyser indépendamment, et regrouper entre régions marines ce qui est réellement comparable.

Identifier des profils typiques d'un contexte et analyser leur concomitance constitue la phase de contextualisation, mais la difficulté est que chaque opérateur possède la même définition, une compréhension et une méthode de mesure identique de ces paramètres.

Il a été difficile de mettre "au diapason" les différentes équipes entre la Grèce, la Turquie et la France, donc nous avons basé ces premiers résultats sur les jeux de données relevés en France.

Les résultats de l'analyse du lien entre cartographie et type de peuplements illustrées avec la figure 28 montrent que nous pouvons interpréter le premier axe de variabilité comme fortement lié à l'intensité lumineuse, avec deux groupes de points bien distincts regroupant des mesures de fréquence de taxons vivant dans des conditions de luminosité contrastées. Le second axe s'explique par les facteurs en fonction de la variabilité de la lumière et des

courants (des études complémentaires sont à prévoir). Les préférences des assemblages d'espèces pour différentes associations de paramètres simples permettront de proposer une typologie coralligène et de comprendre quelles sont les différences de préférences à l'échelle méditerranéenne. Un "profil" est une combinaison de paramètres d'orientation, d'inclinaison et de rugosité. Une étude plus détaillée devra être menée sur les associations de caractéristiques de profil pour déterminer les profils préférentiels de différentes communautés coralligènes. Cette contextualisation est une étape nécessaire pour permettre d'identifier des paramètres pertinents, fiables et efficaces pour expliquer cette variabilité « naturelle » en tenant compte de la variabilité à l'échelle méditerranéenne et d'un large panel d'observateurs de niveaux de compétences différents.

Les paramètres issus des métadonnées ou de l'analyse des photos peuvent aussi constituer des éléments de contextes pour l'approche moléculaire et révéler une espèce cryptique *via* un préférendum pour un contexte différent. Lors des relevés, nous avons aussi fait des prélèvements de *Myriapora truncata* et *Lithophylum cabiochae*, dont les résultats seront présentés dans le travail de thèse de Aurélien De Jode (2018) et dans De Jode *et al.* (2018).

4.3 Discussion concernant l'analyse des photos CIGESMED

Dans le cadre de ce travail de thèse, nous nous sommes intéressés à un seul type de profil pour l'analyse statistique classique des quadrats photo (vertical à 30 mètres), car nous avons vu dans la partie contextualisation de ce chapitre que ces paramètres introduisent énormément de variabilité entre quadrats photo à l'intérieur d'un site. Par contre, les orientations n'avaient pas montré d'influences évidentes. Dans des profils verticaux, on pourrait s'attendre à ce que les ACC soient bien représentées car elles sont alors dans les conditions optimales pour se développer. Les rhodophytes représentent maximum 40% à MEJ, et moins sur les deux autres sites pourtant connus pour leurs concrétionnements.

Cnidaria, Porifera, Algae (très majoritairement des rhodophytes) représentent à eux trois près de 50% des recouvrements sur tous les sites : on peut se poser la question "est-ce un bon déterminant des sites avec ce faciès ?" (Vertical à gorgone autour de Marseille au moins).

Sur les sites, les profils avaient la même inclinaison, mais une grande variabilité de deux facteurs de profils (l'orientation et la rugosité) caractérise le site RMO. L'explication la plus plausible est que plus un site est hétérogène, moins il se différencie des autres, il se placera en "moyennant" entre les autres sites plus homogènes sur les analyses multidimensionnelles. C'est exactement ce que l'on constate dans la CAH et la n.M.D.S (Figure 29C) faite pour comparer les trois réplicas de neuf quadrats photo des sites MEJ, RMO et TFT. D'un point de vue statistique, en comparant trois sites, avec 900 points (9 photos * 100), on obtient une différentiation imparfaite des sites. Dans le cadre d'une même

région marine voire même une même localité, en augmentant le nombre de sites, il y a un grand risque que ce soit encore plus le cas.

Que ce soit en étudiant la structure en considérant le niveau des espèces, des genres ou d'un groupe plus élevé de taxons (jusqu'au phylum), on différencie bien les sites (Figure 29D, 29E et 29F). Donc on peut prendre le parti d'une détermination à un moindre niveau et ainsi espérer gagner du temps pour les protocoles à large échelle, surtout dans le cadre de suivis de gestion.

LPD est caractérisé par son faible nombre d'espèces et la surreprésentation de groupes comme les *polychaeta*, *echinodermata* et les *rhodophyta*.

Ce site a une orientation très majoritairement vers le Sud et une situation qui l'expose de manière singulière aux courants liguriens. Ce site est sur la côte et même s'il est éloigné des influences humaines (à part pour la plongée, mais de manière bien moindre que FTF et RMO), il comporte moins d'espèces dressées (notamment de cnidaires comme les gorgones). Une des difficultés est de comprendre quel est le facteur (ou les facteurs) déterminant(s) pour un site vraiment particulier. Nous essayerons d'apporter des éléments de solution dans le chapitre 5, en s'appuyant sur la visualisation des données sous forme de graphes.

Pour certains sites (certains segments de RMO mais surtout pour des segments de transects de FTF), la surreprésentation des grandes espèces dressées peut avoir un impact sur les résultats : sur la photo, c'est avant tout la strate élevée qui est visible, surtout lorsque les espèces comme les gorgones déploient leurs polypes (par exemple). Dans ce cas, une analyse faite en excluant de l'espèce peut permettre de montrer des informations supplémentaires.

Un des paramètres majeurs à surveiller pour bien utiliser les techniques de quadrats photo est l'apparition d'espèces saisonnières à fort recouvrement comme les algues filamenteuses ou des espèces comme *Asparagensis Armata*, qui, par leur recouvrement très majoritaire, peuvent rendre un jeu de données non exploitable dans le cadre d'un suivi à long terme.

Enfin, il faut noter que la détermination sur photo permet de comptabiliser sensiblement moins d'espèces qu'un plongeur. Un procédé ne pourra probablement pas remplacer l'autre sur tous les suivis possibles.

4.4 Discussion concernant l'analyse des photos des faces de plaques des ARMS

Malgré le fait que nous utilisions des rangs taxonomiques élevés, les photo-analyses des ARMS s'avèrent être un outil efficace pour comparer les communautés benthiques marines. Nous avons détecté des effets significatifs à tous les niveaux de notre plan expérimental : mer, site (dans une même mer) et face de la plaque (pour un même module). La capacité de l'analyse photo ARMS à discriminer entre les mers n'est pas surprenante car elles correspondent à des unités biogéographiques bien différenciées (Spalding *et al.*, 2007) avec des distinctions substantielles dans une variété de paramètres environnementaux (tels que la salinité, la lumière ou la disponibilité des nutriments). Les différences entre les régions ne peuvent pas être interprétées rigoureusement puisque les photographies provenant de mers distinctes ont été analysées par des observateurs distincts et les ARMS n'ont pas été installés pour des durées absolument identiques (et dans le cas de NWM, la profondeur d'installation est plus importante) dans les quatre mers. Des effets saisonniers évidents sont observables dans les mers tempérées, y compris les successions d'organismes et de communautés, par opposition à des latitudes plus équatoriales (où la variation de la durée du jour ou de la température est faible) (Mellin *et al.*, 2016, van Hoytema *et al.*, 2016). De plus, les écarts de date de déploiement et de retrait entre les mers contribuent probablement à la différence de composition de la communauté imputée au facteur « mer » dans nos analyses statistiques. Ceux-ci contribuent probablement à la différence de composition de la communauté imputée au facteur « mer » dans nos analyses statistiques.

Néanmoins, notre étude est à notre connaissance la première englobant une telle variété de régions non tropicales. Ce qui est particulièrement pertinent pour une étude pilote visant à évaluer l'utilité potentielle des dispositifs de surveillance ARMS pour la gestion, c'est leur capacité à distinguer les sites d'une mer ou d'une région, ou de définir un état initial propre à chaque mer pour pouvoir détecter des tendances et déterminer des seuils.
Les sites sont soumis à des conditions environnementales distinctes mais les groupes d'espèces ayant la possibilité de s'y installer sont régionalement les mêmes. Ce sont donc les conditions environnementales qui pourront faire varier la structure des communautés.

Le pouvoir discriminant des mers apparaît clairement, malgré le fait que les facteurs confondants décrits ci-dessus (écarts de date et de profondeur) aient pu empêcher cette différenciation claire. Ces facteurs confondants étaient plus susceptibles d'induire une différenciation entre les sites ou les conditions environnementales, mais ont des effets/impacts identiques pour tous les sites dans le cadre de chaque mer car ils sont

identiques entre sites dans chaque mer. Les sites ont été choisis pour leurs contextes très contrastés a priori pour correspondre à différents niveaux de pression dans chaque mer. Le plan expérimental de cette étude pilote n'était cependant pas principalement destiné à tester un effet des facteurs environnementaux booléens que nous avons examinés en second lieu mais plutôt à vérifier la faisabilité d'un suivi des structures de communautés dans des régions maritimes très distinctes. Il serait plus approprié d'utiliser beaucoup plus de sites dans une région pour étudier les effets environnementaux régionaux, ceci afin de séparer les effets spatiaux purs des effets environnementaux. Par exemple, avec au moins 6-12 sites linéairement placés parallèlement à la ligne de côte, et des variations de facteurs environnementaux intercalés dans chaque partie de ce transect, il devrait être possible d'isoler un effet purement spatial (en calculant la corrélation des distances dans la composition des communautés avec les distances spatiales entre les sites) des effets liés au contexte environnemental. En se concentrant sur une région donnée (par exemple le Golfe de Gascogne ou dans la Baie de Marseille au Nord-Ouest de la Méditerranée), il serait également possible d'utiliser des catégories taxonomiques plus fines, ce qui donnerait probablement plus de pouvoir pour détecter les effets environnementaux sur la composition des espèces. À titre d'exemple, dans le Golfe de Gascogne, le Nord-Ouest de la Méditerranée, la Mer Adriatique et la Mer Rouge, nous avions initialement recensé 31, 36, 33 et 34 taxons distincts, bien que notre analyse globale n'utilisât que 14 catégories.

En effet, nous n'avons trouvé aucune preuve d'un effet des facteurs environnementaux booléens étudiés (tels que définis à l'origine de ce travail). Nous ne pouvons pas conclure si les différences significatives trouvées entre les sites sont dues à des effets spatiaux typiques (le fait que des groupes différents d'espèces sont disponibles pour coloniser des ARMS dans des sites distincts) ou à certains effets environnementaux impliquant le filtre de sélection naturelle, ou n'ont pas été détectés avec les grandes catégories booléennes utilisées dans cette étude. Pour la région Nord-Ouest méditerranéenne, nous avons des arguments de génétique des populations indiquant que le flux et la dispersion des gènes ne sont pas illimités entre les sites, malgré leur proximité (Cahill et al 2017), un effet purement spatial expliquera au moins quelques différences de sites. Cette région maritime était également la seule où la plupart des facteurs environnementaux (4 sur 7) correspondaient au contraste le plus élevé parmi les sites (tableau 21).
On pourrait recommander que les effets de paramètres environnementaux, antagonistes ou potentialisateurs, soient étudiés plus précisément et séparément en commençant par ceux décrivant les conditions naturelles (habitats proches ou lumière, profondeur) puis en expérimentant sur ceux décrivant l'influence de facteurs anthropiques.

Nos résultats soutiennent la proposition que la photo-identification des plaques ARMS peut être utilisée comme un outil de diagnostic rapide pour détecter les changements dans la composition de la communauté à des échelles spatiales relativement petites (dizaines de km). Ce diagnostic doit pour l'instant être considéré avec prudence car nous avons utilisé des catégories taxonomiques très larges. En effet, les rangs taxonomiques identifiables définis par les scientifiques travaillant dans chaque mer ont parfois dû être considérés au rang taxonomique supérieur pour limiter le risque d'erreurs parmi les observateurs et permettre une comparaison entre mers. Comme cela a déjà été souligné, le pouvoir discriminant pourrait être accru dans les études futures portant sur une seule mer en utilisant des catégories taxonomiques plus fines et en incluant des sites plus nombreux. Certains taxons, considérés isolément, se sont montrés plus puissants que d'autres pour discriminer les mers. Cependant, l'abondance des taxons considérés un par un, contrairement à la composition de la communauté, ne permettait pas de discriminer les sites à l'intérieur des mers, à l'exception de *Chlorophyta* installés sur les faces supérieures. En revanche, les analyses au niveau de la communauté (PERMANOVA imbriquée des sites dans les mers) ont révélé des effets significatifs à la fois pour la mer et le site, même lorsque le facteur face n'a pas été pris en compte. De plus, pour cinq des six faces de plaques analysées séparément (l'exception, pour P8T, s'expliquant par deux images manquantes), les compositions communautaires diffèrent significativement entre les régions et entre les sites (au sein des régions). Ceci est une illustration de la limite inhérente aux approches basées sur des taxons identifiés peu précisément : l'utilisation des données correspondant aux abondances relatives est plus efficace que celle des données d'abondance des taxons considérés séparément pour détecter les changements et surveiller l'état écologique (Borja et al., 2015).

Le fait qu'il ne soit pas possible de différencier les sites dans les mers en examinant les plaques P8T dans le cas où deux sites avaient des données manquantes (et étaient donc représentés par des duplicatas plutôt que des tripliquas) suggère que tripler l'ARMS dans chaque site suivi est une nécessité. Lorsqu'une unité ARMS est accidentellement perdue, le fait d'avoir plusieurs faces dans les autres unités ARMS installées sur le même site peut partiellement compenser le manque d'informations, puisque les faces uniques dans une mer ont pu différencier les sites dans trois mers sur quatre (Tableau 5). En effet, une caractéristique frappante de notre étude est la forte différence dans la composition de la communauté entre les faces des plaques de l'ARMS : ceci établit que les faces représentent des micro-habitats distincts. Ils présentent des différences évidentes par exemple en termes d'exposition à la lumière, et/ou à la prédation et/ou au courant, et/ou à la sédimentation. Plusieurs taxons d'algues semblent contribuer fortement à la singularité de la plaque P1T (nMDS, Figure 31), ceci s'explique probablement par l'exposition à la lumière. Une étude

récente menée sur les ARMS colonisant les microalgues a fourni des conclusions similaires (Pennesi et Danovaro, 2017). Pour les faces compartimentées (P4B et P8B), chacune de ces faces représente 4 unités de colonisation indépendantes et / ou est moins susceptible d'être affectée par un événement aléatoire donné tel qu'une prédation par un brouteur. Cette interprétation est supportée par le fait que ces faces affichent les plus bas niveaux de dispersion des données, et significativement moins que les autres (analyses PERMDISP). La structure en sandwich des ARMS apparaît donc comme une caractéristique positive de ces systèmes, par rapport aux plaques de colonisation monocouche, car elle permet de suivre des micro-habitats distincts et peut garantir un design de test plus sensible.

Des recommandations peuvent être faites pour la mise en œuvre d'autres ARMS en dehors des zones de récifs coralliens pour en faire un outil de surveillance efficace des effets des conditions environnementales et à la pression anthropique.
Nous avons déjà invoqué la nécessité d'utiliser plus de sites dans les régions car un très fort effet de site sur la composition de la communauté a été détecté et peut être dû à des effets purement spatiaux (espèces disponibles dans une région marine) et non environnementaux. Considérer les faces des plaques séparément est susceptible d'améliorer la puissance statistique, mais l'analyse des 16 faces n'est probablement pas nécessaire pour séparer les types de faces. Nous recommandons cependant que les études futures comparent les 16 faces afin de fournir une image complète des différences par paires et de la dispersion d'un ARMS. Sur la base des résultats, un sous-ensemble de faces pourrait être sélectionné pour une surveillance temporelle répétée en fonction de leur sensibilité à tel ou tel paramètre environnemental.

Les ARMS doivent être immergés pendant au moins 1 an, ils ne peuvent pas être utilisés pour la surveillance sur des échelles de temps courtes (par exemple, des perturbations sur des échelles de temps de plusieurs jours, semaines ou mois). Les unités ARMS devraient être remplacées sur les mêmes sites lors de leur retrait afin de fournir une série chronologique, qui permettrait de détecter les changements dans la composition de la communauté aux perturbations chroniques (par exemple l'eutrophisation, le changement climatique). Enfin, l'utilisation de l'ARMS est complémentaire aux études portant sur la communauté benthique établie car les ARMS ne mettent en évidence que des pionniers ou des stades précoces de la communauté qui peuvent montrer des réponses différentielles aux impacts comparés à la communauté établie (Pearman *et al.*, 2016). En raison du statut de succession précoce de la communauté sur l'ARMS, ces unités pourraient éventuellement être utilisées pour surveiller l'arrivée d'espèces non indigènes (Hayes *et al.*, 2005, Marraffini *et al.*, 2017). Enfin, afin de rendre les données aussi reproductibles et vérifiables que

possible, elles doivent être placées dans un référentiel à accès libre avec les métadonnées appropriées. Cela permettrait alors d'améliorer les efforts de surveillance futurs, car les études futures pourront ainsi intégrer des jeux de données de référence[83] utilisables à des fins de comparaison temporelle.

Généralement, les programmes de surveillance ne sont pas entrepris à l'échelle pan-régionale et se concentrent plutôt sur des échelles régionales plus petites. Comme je l'ai déjà évoqué, afin de mieux comprendre comment les pressions environnementales distinctes affectent la communauté benthique marine, un plus grand nombre de sites devrait être évalué dans chaque région. Aussi, chaque fois que possible, la mesure régulière des données de température ainsi que l'analyse des concentrations de nutriments, de chlorophylle et de contaminants chimiques permettraient d'étudier les inférences entre les changements de la biodiversité et de la composition de la communauté avec les variables physico-chimiques. Les capteurs de température et de lumière ont maintenant un coût abordable et peuvent être reliés à au moins une unité par site.

Les substrats benthiques, observés avec des approches photographiques, ne révèlent qu'une partie superficielle de la biodiversité locale (Sini *et al.*, 2015). Les structures de communautés décrites sur photo peuvent ne pas représenter toute la diversité des habitats 3D très complexes, tels que les récifs coralligènes (notamment les espèces endogènes). Ceci est particulièrement vrai pour la composante sessile de la communauté. De plus, comme cette analyse a été entreprise par des chercheurs provenant de régions différentes, la résolution taxonomique a été perdue lorsque les données ont été combinées en un seul ensemble de données pour les rendre "compatibles". Bien que l'utilisation d'un seul expert pour analyser toutes les plaques pourrait supprimer ce problème, cela n'est pas toujours possible voire impossible pour la production des grands ensembles de données demandés par des institutions transfrontalières. Les techniques moléculaires (notamment le barcoding et le metabarcoding proposées par Leray et Knowlton (2015) dans leurs analyses des ARMS), avec leurs propres limites (Carugati *et al.*, 2015), pourraient offrir de nouvelles opportunités pour rendre l'analyse des ARMS plus efficace et standardisée (Ransome *et al.*, 2017). En effet, Pearman *et al.* (2016) ont montré que pour les ARMS dans la Mer Rouge, une plus grande diversité, englobant un plus large éventail de taxons, a été observée, en utilisant des techniques moléculaires, comparées à des approches morphologiques. Dans leurs analyses, le metabarcoding était capable de différencier les sites alors que les

[83] Les données de références sont des données utilisées donc partagées par l'ensemble des processus et partenaires d'une organisation. Elles sont à la base des prises de décision.

approches morphologiques ne montraient pas de différence significative dans la composition entre les sites (malgré la mise en commun de la fraction sessile de toutes les faces des plaques pour le metabarcoding). Un projet en cours présentera les résultats des analyses de metabarcoding effectuées sur les ARMS étudiés ici (et incluant d'autres régions européennes).

- Eléments clefs :
    - Les ARMS sont exploitables quelle que soit la mer (effet spatial entre sites) !
    - Les analyses en séparant les micro habitats donnent plus d'information.
    - 3 réplicas sont nécessaires pour avoir un suivi fiable.
    - 64 points sont suffisants pour détecter des différences significatives entre sites.
    - La fusion des catégories taxonomiques permet de montrer des différences entre sites.
- Limites de la méthode :
    - La méthode n'est pas encore pertinente pour détecter un effet anthropique
    - Donc plus de sites seraient nécessaires pour meilleure sensibilité,
    - Mais lorsque beaucoup de réplicas : la synchronisation est très délicate.
    - Quelques idées pour adapter ce procédé à la gestion :
        - Une simplification de la structure
        - Une approche depuis la côte (en plongeant du bord)

4.5 Inter-calibration, une nécessité face aux directives internationales

L'approche intégrée de la complexité des habitats coralligènes proposée dans le cadre de CIGESMED permet de mutualiser et de visualiser de vastes collections de données et de gérer les connaissances pour étudier les écosystèmes. Cette inter-calibration peut seule permettre de créer un socle commun, base d'une donnée plus standardisée et comparable d'une région à une autre. Elle permet aussi de travailler à grande échelle sur l'accessibilité des dispositifs d'observation[84] ce qui améliore leur efficacité.

Les indicateurs construits pour légiférer ou aider à la décision à large échelle ne peuvent être basés que sur des systèmes d'observation inter-calibrés et donc efficaces à large échelle (ou parfois sur des systèmes d'observation qui ne peuvent absolument pas s'appuyer sur les mêmes méthodes de mesure de facteurs ou d'indication, car ayant des objectifs et des

[84] L'accessibilité d'un système d'observation est définie par le niveau de facilité avec lequel un observateur va pouvoir mettre en œuvre un dispositif d'observation (coût du matériel + compétences sollicitées + temps à impartir en formation, observation, traitement). Elle a une influence sur la taille du panel d'observateur compétent à large échelle et sur les coûts.

contraintes trop différents). En considérant les facteurs utiles à mesurer autant pour les analyses de communautés que pour recherche sur une espèce en particulier, une typologie adaptable mais commune de mesure de paramètres contextuels doit être coconstruite et la variabilité de chaque paramètre doit être testée par tous les observateurs potentiels (les scientifiques, les parcs naturels marins et les gestionnaires de réserves, et les réseaux de « science citoyenne »). L'utilisation de nouvelles représentations de données et de qualifications contrôlées devient indispensable.

Les représentations graphiques utilisées aujourd'hui ne conviennent pas pour trier, organiser et analyser de très grands ensembles de données hétérogènes concernant le coralligène, et *a fortiori* concernant tous les aspects de la biodiversité. De nouvelles approches concernant la gestion et l'analyse de la donnée sont nécessaires, et devront s'appuyer sur une nouvelle organisation des acteurs (producteurs et utilisateurs de la donnée), et des processus d'inter-calibration entre tous les acteurs à chaque étape de création ou de transformation d'une action.

Pour construire un réseau plus efficace et plus interopérable (et donc avec une meilleure inter-calibration), la communauté CIGESMED a développé un catalogue de métadonnées et certaines typologies partagées ; à mon sens, la suite à donner à ce programme afin que celui-ci soit utilisé sur le long terme concerne: i) l'harmonisation des méthodes de collecte de données et la normalisation de l'accès aux données en intégrant les nouveaux prérequis générés par les consortiums internationaux (biodiversité, bases de données et fouille de données) ii) l'initiation / l'animation d'un réseau thématique sur les habitats coralligènes en Mer Méditerranée rassemblant tous les acteurs compétents. Ce réseau serait destiné à être pérenne, ouvert et entièrement décentralisé (pour permettre une mise à jour continue) aux échelles locale, régionale, nationale et internationale.

Cette organisation à mettre en œuvre est incontournable pour permettre une diffusion efficace de la donnée, sa transformation pour les différents usages qui peuvent en être fait. Pour cela, un système de gestion de la qualité de la donnée est nécessaire à chaque niveau géographique (local, régional et national, puis pan-méditerranéen) afin d'assurer une amélioration continue de la comparabilité des données à large échelle (et entre les échelles). La recherche sur l'analyse et la visualisation de données sur une large échelle est encore trop peu développée concernant les habitats marins. Le résultat à atteindre est pourtant une évaluation intégrée du G.E.S. dans le cadre de la D.C.S.M.M..

Chapitre 3 : Utilisabilité des données et systèmes de gestion et d'entreposage des données

Ce chapitre utilise des éléments des publications El guerrabi 2015, David *et al.*, 2016a, Féral *et al.*, 2016.

1. Généralités et questionnements concernant les systèmes de gestion et d'entreposage des données

1.1 Quelques observations sur les relations entre typologie et propriétés des données

La donnée d'observation automatisée est souvent mieux typée car produite par un automate qui en contrôle les formats de sortie. Cette donnée appartient à l'organisme ou la personne qui est propriétaire de ce capteur, sauf en cas de convention explicite (e.g. les données produites par un bureau d'étude sur une aire protégée, peuvent être par convention rétrocédées à la structure qui gère ce milieu, comme un parc naturel, un conservatoire d'espaces naturels, etc.). Néanmoins, pour un même type de donnée (par exemple, une température), les formats de lecture dépendent de la génération de ces automates, et il n'est pas rare que même en utilisant des formats très simples (un csv avec une date, une heure et une température pour reprendre le même exemple), l'agrégation de données d'un même type soit rendue laborieuse par la multiplicité des générations et des marques de capteurs déployés sur le terrain. Ceci est encore plus vrai à large échelle et sur une grande échelle de temps, et les données historiques produites par un opérateur privé, même si elles devraient être tombées dans le domaine public, continuent à rester accessible à un format propriétaire et nécessitent un abonnement pour le moins pour accéder à une donnée non floutée. Les erreurs appelées valeurs manquantes liées aux pannes des capteurs sont souvent non corrigées dans les bases de données (dérive des valeurs, ou non enregistrement pendant plusieurs jours, mois voire pour certaines balises, années).

Lorsque la donnée est produite par un observateur, pour un même type de données (par exemple la reconnaissance d'un taxon), ces problèmes de formats et de correction de la donnée deviennent prépondérants et chronophages. A cela s'ajoute le fait que cette donnée est une donnée issue d'une interprétation. A ce titre, elle n'a pas le même statut que la donnée produite par un capteur qui par essence est le bien du détenteur du capteur. L'observateur, parfois appelé "inventeur", interprète ce qu'il observe, soit en reconnaissant

un taxon, soit en estimant une abondance, et à ce titre, selon son statut, il peut détenir un droit d'auteur sur l'information récoltée. Un chercheur par exemple détient ce droit, au même titre qu'un bénévole d'une association, sauf conventionnement contraire. Un technicien, lui, rétrocède généralement contractuellement sauf convention contraire à l'organisme qui l'emploie (comme c'est le cas au C.N.R.S.). La donnée précédemment décrite est une donnée brute. Chaque traitement la transforme, et en fonction des traitements qui lui sont fait, elle peut changer de statut, passant d'un format libre ou au moins non propriétaire, à un format propriétaire ou au moins commercialisé, avec un accès contrôlé. Enfin, il faut faire le distinguo entre une information, et une donnée, plus formalisée et explicite : de nombreuses informations sont conservées sans traitement, et demandent une interprétation comme les textes, photos ou vidéos (N.B. à ce stade déjà, il peut y avoir un droit d'auteur selon les mécanismes énoncés ci-avant) : l'information est alors implicite et non directement exploitable. La plus-value et le statut de la donnée issue de l'information dépendront du type de traitement et du statut de l'opérateur qui l'auront rendue explicite.
Enfin, il est bon de préciser que quand l'auteur rend publique une donnée (par exemple sur internet), s'il ne l'a pas mis sous licence, il ne pourra pas poursuivre en justice un bureau d'étude pour le fait de l'avoir vendu, que ce soit avec ou sans amélioration(s). Cette crainte d'être dépossédé est souvent vaine car dangereuse pour le vendeur, et le réutilisateur préférera souvent aller la chercher sur le site du producteur pour s'assurer de sa qualité et de sa mise à jour (cela sous entend que le producteur ait bien référencé et rendu réutilisable sa donnée comme nous le verrons dans les parties suivantes de ce chapitre).

<u>Cadre légal du partage de la donnée en France et en Europe</u>

Convention d'Aahrus

La convention d'Aarhus sur "l'accès à l'information, la participation du public au processus décisionnel et l'accès à la justice en matière d'environnement" est un des premiers cadres internationaux incitatifs pour le partage de la donnée dans le domaine environnemental. Celle-ci est signée le 25 juin 1998 par trente-neuf États et établit un accord international visant la « démocratie environnementale » reposant sur 3 principes : améliorer l'information environnementale délivrée par les autorités publiques, favoriser la participation du public à la prise de décisions et étendre les conditions d'accès à la justice en matière de législation environnementale.

Elle reprend notamment le Principe 10 de l'article 2 de la convention sur la diversité biologique signée à Rio de Janeiro lors du Sommet de la Terre de 1992 stipulant que « La meilleure façon de traiter les questions d'environnement est d'assurer la participation de tous les citoyens concernés, en mettant les informations à la disposition de celui-ci ». Sa mise en

application progresse lentement et de manière différente dans chaque pays signataire grâce à un processus itératif reposant sur des rapports annuels regroupant états des lieux et possibilités de progresser. Par exemple en France, ses principes, d'abord adoptés par décret en 2002[85], ont été intégrés en 2005 dans le bloc de constitutionnalité du droit français avec la charte de l'environnement.

Directive INSPIRE

La directive européenne INSPIRE, plus récente (2007/2/CE du 14 mars 2007) « vise à établir une infrastructure d'information géographique dans la Communauté européenne pour favoriser la protection de l'environnement. La directive INSPIRE s'applique aux données géographiques détenues par les autorités publiques, dès lors qu'elles sont sous forme électronique et qu'elles concernent l'un des 34 thèmes figurant dans les annexes de la directive, donc sur un champ très large. Elle impose aux autorités publiques, d'une part de rendre ces données accessibles au public en les publiant sur Internet, d'autre part de les partager entre elles. La directive INSPIRE a été transposée dans le droit français par l'ordonnance 2010-1232 du 21 octobre 2010. »

Identification des freins à la mutualisation de la donnée

Afin d'identifier les processus nécessaires à la mise en place du partage des données, il est important d'identifier tous les freins qui aboutissent à un non partage, ou à un faible partage et de pouvoir les qualifier. Lorsqu'on enquête auprès des producteurs de données, de nombreux aspects sont évoqués comme étant mal maîtrisés ou pas assez financés comme les problématiques de volume, de complexité, de localisation, d'évolution, d'accessibilité, de traçabilité. L'absence de choix sur ce qui doit réellement être conservé et avec quel niveau de qualité est prégnant (donnée traitée ou brute, standards et normes à respecter, moyens à déployer). Pourtant, la velléité de développer des indicateurs fiables sur le long terme est partout, et pour que ces indicateurs puissent fonctionner dans la durée, il faut qu'ils soient basés sur une donnée bien conservée, et réutilisable sur le long terme. En fin de compte, des solutions existent et évoluent sur le plan international pour que la donnée soit partageable, et les infrastructures mis à disposition dans le domaine de la biodiversité sont sous exploitées et peu connues des écologues. Beaucoup de producteurs de données, notamment dans les équipes de recherche adoptent, quand ils s'en préoccupent, des solutions"partielles" et /ou "faites maisons" de partage de données. Ils peuvent pourtant rarement garantir ni la pérennité des infrastructures mises en place, ni la pérennité des compétences permettant leur maintenance et leur évolutivité au gré des besoins. Ce choix

[85] https://www.legifrance.gouv.fr/eli/decret/2002/9/12/MAEJ0230045D/jo/texte

est souvent fait par méconnaissance des architectures possibles et des solutions ayant montré leurs performances. Cette méconnaissance favorise les craintes de dépossession des données, car lorsqu'elles ont coûté cher, le producteur ne veut pas que cette donnée puisse être monnayée telle quelle ou avec des transformations, même majeures. La plupart des producteurs de données sont très loin de penser que c'est le partage de la donnée qui augmente sa valeur.

Focus sur les blocages au partage de la donnée

La volonté de partage et la reconnaissance du partage sont deux notions qui peuvent être fortement liées ; des exemples réussis de partage de données peuvent constituer un déblocage à d'autres producteurs. Pour aboutir à ce déblocage, il est important d'identifier et d'analyser la ou les raison(s) et raisonnement(s) qui aboutissent à ce non partage.

Quelles sont les différentes raisons invoquées pour ne pas partager les données :

- JE NE VEUX PAS – du pire au plus acceptable -> à discuter (la solution passe par une sensibilisation du producteur) (niveau -1)
    - J'ai trafiqué mes données, je ne veux pas qu'on puisse s'en rendre compte
    - Mes données ont de la valeur, je ne veux pas que quelqu'un les revende à ma place
    - Je n'y gagne rien
    - Je ne veux pas que quelqu'un d'autre écrive une publication en utilisant mes données
    - Je ne veux pas que ces données soient publiées vu leur qualité
    - Je souhaite publier avant
- JE NE PEUX PAS (quand c'est possible, il s'agit d'apporter des moyens et un accompagnement étape par étape) (niveau 0)
    - Contraintes légales
        - Données sensibles (exemple, une station d'une espèce protégée)
        - Données personnelles (liberté individuelle, droit à l'anonymat)
        - Données sécuritaires (sécurité de l'état ou autre motif régalien)
    - Temps et/ou moyens
        - Pas le temps de les améliorer pour qu'elles soient diffusables (Curation)
            - Données contenant des erreurs de syntaxe
                - Erreurs lexicales
                - Erreurs de formatage
                - Erreurs d'irrégularité
                - Données ambivalentes
            - Erreurs sémantiques
                - Violation des contraintes d'intégrité
                - Erreurs de contradiction
                - Erreurs de duplication
            - Erreurs de donnée invalide
                - Erreurs de couverture
                - Valeur impossible
                - Valeur manquante
                - Donnée manquante
                - Données dont la précision n'est pas connue
        - Pas le temps de les documenter pour qu'elles soient diffusables (qualification de la qualité)

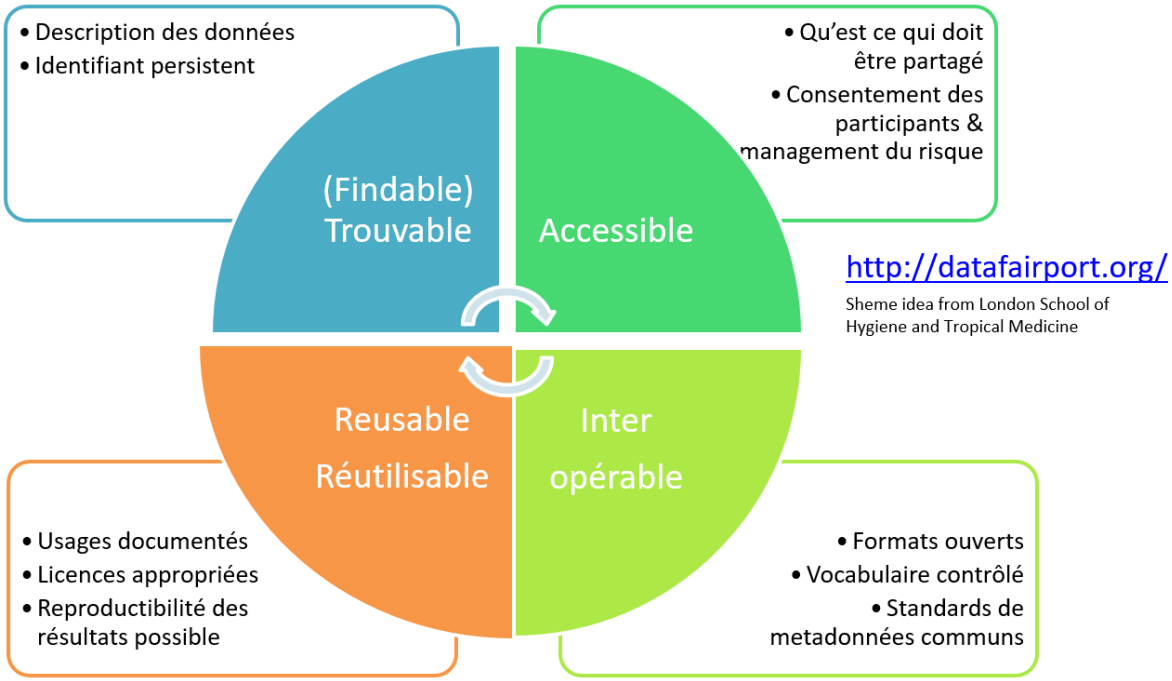
                - Pas de définition de champs
                - Pas de dictionnaire de données
                - Pas de respect des standards
                - Pas de data management plan
            - Pas le temps de me former sur les outils/méthodes indispensables
                - Standards de données
                - Sémantique des données de mon domaine
                - Curation de données
                - Préservation de données
                - Reproductibilité des données / Compatibilité des données
                - Interopérabilité des données
            - Pas le temps de les documenter pour qu'elles soient compréhensibles
                - Dictionnaire de données
                - Schéma relationnel et contraintes d'intégrité
                - Protocole de production
                - Transformations faites pour les obtenir
        - Pas le savoir-faire (pas le temps de me former)
            - Pour les mettre sur un entrepôt de données
            - Pour les transformer d'un format à un autre et respecter les standards
            - Pour les indexer afin qu'elles soient bien référencées par tous les catalogues et les utilisateurs potentiels
            - Pour créer des services autour de ma donnée
            - Pour la documenter convenablement
            - Pour préserver tous les éléments nécessaires à la réutilisation de la donnée (logiciel, environnement, algorithmes, lecteur…)
- JE PEUX MAIS JE MANQUE DE GUIDE (ici, les niveaux indique une proposition de qualification de la qualité du partage de données
    - Pas le savoir-faire (même si j'ai le temps pour cela)
        - Pour les mettre sur un entrepôt de données (niveau 1)
        - Pour les transformer d'un format à un autre et respecter les standards (niveau 2)
        - Pour la documenter convenablement (niveau 3)
        - Pour les indexer afin qu'elles soient bien référencées par tous les catalogues et les utilisateurs potentiels (niveau 4)
        - Pour créer des services autour de ma donnée (niveau 5)
        - Pour préserver tous les éléments nécessaires à la réutilisation de la donnée (logiciel, environnement, algorithmes, lecteur…) (niveau 6)

1.2 La mise à disposition de la donnée avec les principes FAIR

Les principes FAIR [Findable, Accessible, Interoperable, and Reusable (Figure 32)] sont issus d'une proposition d'un groupe de discussion appelé Force 11[86] et publiés en 2016 (Wilkinson *et al.*). Ils sont aujourd'hui adoptés par les tutelles européennes qui guident les bénéficiaires d'Horizon 2020 pour rendre leurs données de recherche trouvables, accessibles, interopérables et réutilisables, et cherche ainsi à s'assurer qu'elles sont bien gérées. L'argument avancé n'est pas anodin : "Une bonne gestion des données de la recherche n'est pas une fin en soi, mais plutôt le passage clé qui mène à la découverte des connaissances et à l'innovation, aux données ultérieures ainsi qu'à l'intégration et à la réutilisation des connaissances."[87]

Dans le domaine de l'environnement, une donnée est rarement déconnectée d'une unité ou d'un *objet géographique dont elle décrit un état ou une caractéristique à un instant T.

D'après " La directive Inspire pour les néophytes » par exemple, pour que les données puissent être publiées et réutilisées, il est nécessaire qu'elles « respectent des règles d'interopérabilité dans 2 domaines : i) Sémantique : il s'agit de définir, grâce à un modèle, le sens, le contenu et la structuration des données. Ii) Géographique : les coordonnées géographiques (longitude et latitude) des données dépendent du système géodésique utilisé et les coordonnées planes dépendent de la projection cartographique. »



[86] https://www.force11.org/group/fairgroup/fairprinciples

[87] Lignes directrices pour la gestion des données FAIR dans Horizon 2020 : http://www.donneesdelarecherche.fr/IMG/pdf/lignes-directrices_gestion-donnees-fair_horizon2020_version_3.0_tr-fr.pdf

Figure 32 : les 4 principes FAIR et les conditions sous-jacentes, pour chacun des principes, mis en œuvre à la manière d'un processus qualité, c'est à dire de manière itérative, à chaque fois qu'un nouvel usage apparaît, ou qu'une modification sur les données est faite (enrichissement, curation, analyse…).

Réalité de la mise en œuvre des principes FAIR et des textes associés

Les raisons de non partage sont, nous l'avons vu, souvent le fait de facteurs humains. La méconnaissance des besoins pour la gestion de l'accès à la donnée et des moyens à apporter pour réellement respecter les principes FAIR sont sans doute liés à l'inadéquation en terme de moyens techniques et humains, notamment pour les gros projets qui inscrivent principes et besoins FAIR dans leur programme, mais finissent par ne plus respecter ces critères d'éligibilité. Comme il n'existe pas encore de grille d'évaluation du respect de ces critères, ces actions ne sont pas souvent considérées comme prioritaires.

Par exemple, « L'étude du cas IPERION-CH semble souligner l'impossible gestion commune des données hétérogènes que le projet génère, et l'éclatement de la communauté de chercheurs qu'il est censé fédérer. » (Puren, 2016). Pourtant, nos tutelles nous demandent de plus en plus d'interdisciplinarité dans nos recherches. Dans notre cas, au-delà des travaux effectués sur les données pendant la réalisation des programmes CIGESMED et DEVOTES, même si les données sont documentées et accessibles en ligne, leur maintenance à long terme n'est pas garantie. Pour les données Turques dans le cadre de leur convention ANR adossée au « Seasera », nous avons même rencontré un blocage administratif et légal : il est interdit aux laboratoires turcs de mettre en accès libre leurs données de recherche. De gros travaux d'harmonisation des mentalités puis des contextes juridiques sont manifestement à entreprendre.

1.3 Cycle de vie de la donnée, à quelles conditions?

Le cycle de vie des données de recherche (Research data lifecycle) décrit le processus d'utilisation des données de leur création à la publication et à leur réutilisation ultérieure (Figure 33). (Définition de l'INIST : INstitut de l'Information Scientifique et Technique). Au-delà de l'idée de réutilisation, la notion de cycle est liée à la notion itérative de "démarche qualité".

Le cycle de vie des données comprend les étapes de description et de conservation des données. Un projet traditionnel comporte des étapes de planification, de collecte de données, de contrôle de la qualité des données puis d'analyse. Les projets s'appuyant sur des données existantes, pour tout ou partie de leurs analyses nécessitent aussi les étapes de la planification, de la collecte, de l'assurance qualité et du contrôle qualité (AQ / CQ), auxquelles

il faut ajouter la découverte de données supplémentaires, l'intégration de données et enfin l'analyse. Pour compléter ce cycle de vie des données, dans le cas de projets basés sur la réutilisation de données existantes, des étapes supplémentaires de documentation des données (métadonnées) et l'archivage des données dans un référentiel accessible au public sont absolument nécessaires. Dans le domaine de la recherche intégrée sur les macro systèmes en écologie, l'utilisation des informations nécessaires à la gestion de la donnée a montré ses capacités à apporter une plus-value considérable (J Rüegg *et al.*, 2014).

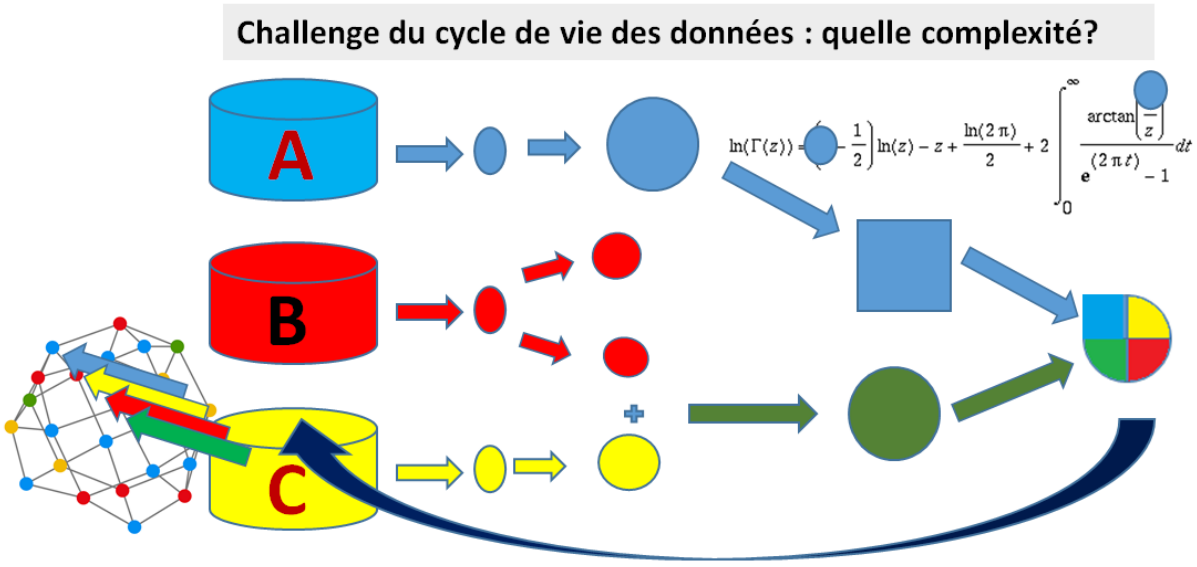


Figure 33 : Plus une donnée est réutilisée et plus le nombre d'utilisations différentes en est faite, plus elle est susceptible de subir des transformations (A), qui peuvent être simples (des additions, des moyennes par exemple) ou relativement complexes et en différentes étapes (symbolisées dans ce schéma par l'équation). Des transpositions peuvent être faites à de nouveaux objets (B) : par exemple une densité sur une région est reportée sur deux sous régions ou des agrégations pour en faire une donnée composite comme un nombre total d'individus sur deux surfaces (B+C). La dernière flèche indique qu'il n'est pas rare que cette donnée transformée devienne une nouvelle donnée brute, sans que soit identifié son origine et ou les transformations qu'elle a subies.

En Europe, UK Data Archive fournit un modèle de cycle de vie des données en 6 étapes sur lequel les chercheurs peuvent s'appuyer pour gérer leurs données :

- Création ou collecte des données (creating data) ;
- Traitement des données (processing data)
- Analyse des données (analysing data)
- Conservation des données (preserving data) ;
- Accès aux données (giving access to data / data discovery) ;
- Réutilisation des données (re-using data).

Au-delà de ce modèle conventionnel, la diversité des usages de données intégrées (et donc multi-source) dans le cadre de systèmes de suivi à large échelle temporelle et spatiale complexifie ce schéma en rendant toutes ces étapes plus ou moins simultanées.

Le cycle de vie d'une donnée est aujourd'hui multiplié en plusieurs cycles parallèles et n'évolue plus seulement à l'intérieur du système d'information du producteur initial de la donnée, mais chez tous les transformateurs de cette donnée dans d'autres systèmes extérieurs. L'organisation des communications entre ces réseaux est nécessaire, pour assurer la traçabilité de la donnée sur le long terme.
Pour continuer à construire chaque réseau de manière rationnelle, chaque communauté développe un catalogue de métadonnées et certaines typologies partagées; dans le cadre du réseau de CIGESMED (notamment dans le cadre du WP6) nous avons travaillé sur i) l'harmonisation des méthodes de collecte de données et la normalisation de l'accès aux données (en prenant en compte les normes notamment européennes lorsqu'elles existent) ii) l'initiation / l'animation d'un réseau thématique sur les habitats coralligènes en Mer Méditerranée rassemblant tous les acteurs compétents. Ce réseau est destiné à être pérenne, ouvert et entièrement décentralisé (pour permettre une mise à jour continue) aux échelles locale, régionale, nationale et internationale, et doit pour cela trouver un financement sur le long terme. En complément de ces moyens, seule une organisation des acteurs entre eux permettra la diffusion de données sur le long terme et une accessibilité garantie aux données simultanément sur les plans locaux, régionaux, nationaux et internationaux. Les systèmes de gestion de la qualité de la donnée doivent être ainsi garantis pour chaque niveau géographique (local, régional et national) afin d'assurer une évolution continue des usages de la donnée sans atteindre leur véracité.

2. Structure de données dans le cadre de CIGESMED et de DEVOTES

Dans le cadre du W.P.6. de CIGESMED, les données résultant des quadrats photo ont été structurées et intégrées dans un modèle de données après validation du protocole et un certain nombre d'essais de terrain. Il a pour cela été nécessaire, *via* une démarche classique d'analyse[88] de développer un Modèle Conceptuel de Données[89] (M.C.D.). Je ne présente pas ici toutes les étapes de cette modélisation d'une base de données au niveau conceptuel comme l'élaboration du dictionnaire des données, la description des dépendances fonctionnelles et du Modèle Conceptuel de Communication[90] (M.C.C.) qui n'ont pas d'intérêt particulier dans le cadre de cette thèse. Le M.C.D. est une image de la base de données décrivant les contenus des tables et les liens qui les relient entre elles. Il contient toutes les entités (aussi appelés objets) ainsi que les propriétés nécessaires pour le développement du système de gestion des données (Figure 34).

Une grande partie des données de CIGESMED utilisées dans le cadre de cette thèse concerne les quadrats photo prises par les observateurs en plongée. Une table intitulé "quadrat_photo_objet" stocke toutes les informations concernant ces quadrats photo, et constitue le cœur de la donnée analysée (c'est cette table qui contient le plus d'enregistrements). Voici donc le détail des typologies de données mises en place dans le cadre des protocoles développés dans CIGESMED et présentées précédemment dans matériel et méthode (Chapitre 2 Partie 2.4) :

[88] On parle d'analyse en informatique lorsque l'on développe un système répondant à des attentes utilisateurs

[89] Le M.C.D. est une représentation des données qui décrit de façon formelle les données utilisées par le système d'information sous forme d'entités.

[90] Le M.C.C. représente de manière normalisée les communications entre les différents acteurs dans le cadre d'un projet ou d'un métier.

Cette table contient des champs correspondant à la typologie de la nomenclature suivante : [Programme]_ [site]_[date]_D[profondeur]_T[transect]_Q[n° quadrat]_[auteur]

- [Programme] : le nom du programme pour lequel le prélèvement est réalisé (Ex : « CIGESMED »). Composé de 8 caractères en majuscules.
- [Site] : représente le code du site sur lequel ont été prises les quadrats photo. Cette partie est constituée de 3 lettres en majuscule. Exemple : « CAS » pour « Cassidaigne » « FTF » pour « Frioul -Tiboulen de Frioul »
- [Date] : la date de prélèvement de l'échantillon sous la forme suivante : YYYYMMDD
- D_ [profondeur] : la lettre « D » en majuscule (« D » pour « depth») suivie du niveau de la profondeur auquel le prélèvement a été (« 1 » pour 28 mètres environ ou « 2 » pour 43 mètres environ)
- T_ [n° transect] : la lettre « T » suivie du numéro à 2 chiffres du transect prélevé (ex : « T01 » pour le transect numéro 1)
- Q_ [n° quadrat] : la lettre « Q » en majuscule suivi du numéro à deux chiffres du quadrat pris en photo sur le transect (ex : « Q05 » pour le quadrat numéro 5).
- [Auteur] : initiales en majuscule (prénom suivi du nom) de la personne ayant prélevé l'échantillon suivi de son identifiant dans la base de données. Ex : « RD01 » pour Romain DAVID.

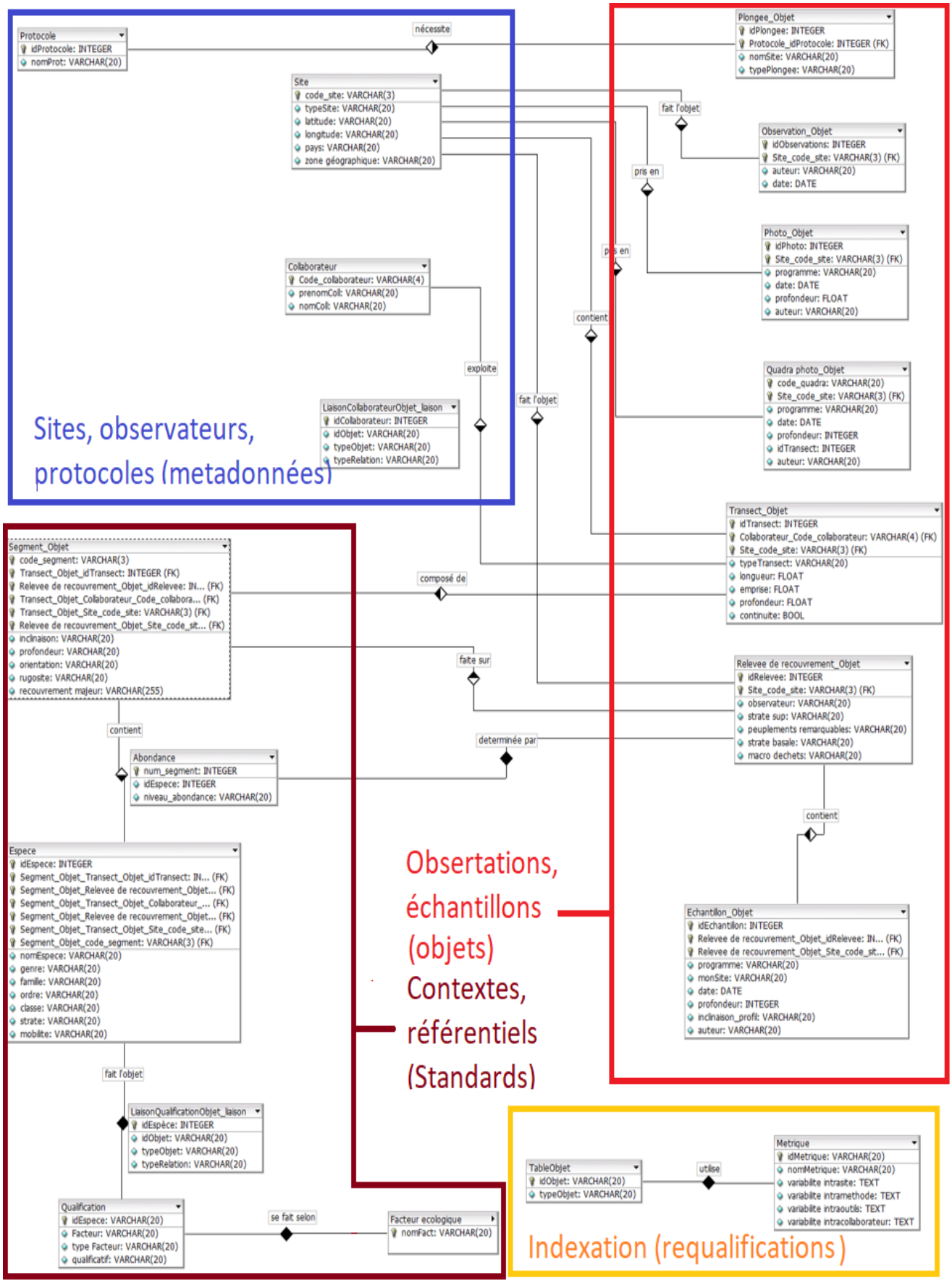

Protocole
idProtocole: INTEGER
nomProt: VARCHAR(20)
nécessite
Site
code_site: VARCHAR(3)
typeSite: VARCHAR(20)
latitude: VARCHAR(20)
longitude: VARCHAR(20)
pays: VARCHAR(20)
zone géographique: VARCHAR(20)
Collaborateur
Code_collaborateur: VARCHAR(4)
prenomColl: VARCHAR(20)
nomColl: VARCHAR(20)
exploite
LiaisonCollaborateurObjet_liaison
idCollaborateur: INTEGER
idObjet: VARCHAR(20)
typeObjet: VARCHAR(20)
typeRelation: VARCHAR(20)
Sites, observateurs, protocoles (metadonnées)
Plongee_Objet
idPlongee: INTEGER
Protocole_idProtocole: INTEGER (FK)
nomSite: VARCHAR(20)
typePlongee: VARCHAR(20)
fait l'objet
Observation_Objet
idObservations: INTEGER
Site_code_site: VARCHAR(3) (FK)
auteur: VARCHAR(20)
date: DATE
pris en
Photo_Objet
idPhoto: INTEGER
Site_code_site: VARCHAR(3) (FK)
programme: VARCHAR(20)
date: DATE
profondeur: FLOAT
auteur: VARCHAR(20)
contient
Quadra photo_Objet
code_quadra: VARCHAR(20)
Site_code_site: VARCHAR(3) (FK)
programme: VARCHAR(20)
date: DATE
profondeur: INTEGER
idTransect: INTEGER
auteur: VARCHAR(20)
Transect_Objet
idTransect: INTEGER
Collaborateur_Code_collaborateur: VARCHAR(4) (FK)
Site_code_site: VARCHAR(3) (FK)
typeTransect: VARCHAR(20)
longueur: FLOAT
emprise: FLOAT
profondeur: FLOAT
continuite: BOOL
composé de
Segment_Objet
code_segment: VARCHAR(3)
Transect_Objet_idTransect: INTEGER (FK)
Relevee de recouvrement_Objet_idRelevee: IN... (FK)
Transect_Objet_Collaborateur_Code_collabora... (FK)
Transect_Objet_Site_code_site: VARCHAR(3) (FK)
Relevee de recouvrement_Objet_Site_code_sit... (FK)
inclinaison: VARCHAR(20)
profondeur: VARCHAR(20)
orientation: VARCHAR(20)
rugosite: VARCHAR(20)
recouvrement majeur: VARCHAR(255)
faite sur
Relevee de recouvrement_Objet
idRelevee: INTEGER
Site_code_site: VARCHAR(3) (FK)
observateur: VARCHAR(20)
strate sup: VARCHAR(20)
peuplements remarquables: VARCHAR(20)
strate basale: VARCHAR(20)
macro dechets: VARCHAR(20)
contient
Abondance
num_segment: INTEGER
idEspece: INTEGER
niveau_abondance: VARCHAR(20)
determinée par
contient
Obsertations, échantillons (objets)
Contextes, référentiels (Standards)
Espece
idEspece: INTEGER
Segment_Objet_Transect_Objet_idTransect: IN... (FK)
Segment_Objet_Relevee de recouvrement_Objet... (FK)
Segment_Objet_Transect_Objet_Collaborateur_... (FK)
Segment_Objet_Relevee de recouvrement_Objet... (FK)
Segment_Objet_Transect_Objet_Site_code_site... (FK)
Segment_Objet_code_segment: VARCHAR(3) (FK)
nomEspece: VARCHAR(20)
genre: VARCHAR(20)
famille: VARCHAR(20)
ordre: VARCHAR(20)
classe: VARCHAR(20)
strate: VARCHAR(20)
mobilite: VARCHAR(20)
Echantillon_Objet
idEchantillon: INTEGER
Relevee de recouvrement_Objet_idRelevee: IN... (FK)
Relevee de recouvrement_Objet_Site_code_sit... (FK)
programme: VARCHAR(20)
monSite: VARCHAR(20)
date: DATE
profondeur: INTEGER
inclinaison_profil: VARCHAR(20)
auteur: VARCHAR(20)
fait l'objet
LiaisonQualificationObjet_liaison
idEspèce: INTEGER
idObjet: VARCHAR(20)
typeObjet: VARCHAR(20)
typeRelation: VARCHAR(20)
Qualification
idEspece: VARCHAR(20)
Facteur: VARCHAR(20)
type Facteur: VARCHAR(20)
qualificatif: VARCHAR(20)
se fait selon
Facteur ecologique
nomFact: VARCHAR(20)
TableObjet
idObjet: VARCHAR(20)
typeObjet: VARCHAR(20)
utilise
Metrique
idMetrique: VARCHAR(20)
nomMetrique: VARCHAR(20)
variabilite intrasite: TEXT
variabilite intramethode: TEXT
variabilite intraoutils: TEXT
variabilite intracollaborateur: TEXT
Indexation (requalifications)

Figure 34 : La structure des données est constituée de quatre domaines différents :
i) les objets observés dont les descripteurs sont décrits dans le protocole (en rouge),
ii) les descripteurs de contexte (en marron), respectant lorsqu'ils existent des standards et/ou reposant sur des référentiels (p.e. ici le(s) référentiel(s) espèce(s),
iii) les données assimilables à de la métadonnée (en bleu) respectant les standards de type sampling-event,
iv) une table d'indexation de chaque objet (en orange) permettant de qualifier chaque objet de la base dans une table commune (structuration de la base de donnée de type graphe).

Le M.C.D. (ici simplifié) de la partie du système d'information de Cigesmed (El Guerrabi W., 2014) concernant les quadrats photo présente deux entités[91] centrales : "Sites" et "Collaborateurs", et des entités objets qui appartiennent à la fois à un collaborateur et à un site : Transect_objet, Segment_objet, Observation_objet, Quadraphoto_objet, Echantillion_objet et Espèce_objet. Une table "qualification" indépendante sert à qualifier les attributs des entités objets. La qualification de certains objets se fait grâce à des référentiels (exemple : utilisation du référentiel taxonomique pour les noms de taxons). L'entité "Site" est reliée aux entités objets, par exemple si on prend Transect, un site contient plusieurs transects différents. Un transect est caractérisé par une longueur, un type (transect aléatoire[92] ou permanent), une emprise, une profondeur et une continuité (transect continu ou discontinu). "Collaborateur" est aussi relié à l'ensemble des tables objets, ce qui signifie qu'un collaborateur joue un rôle très important dans l'acquisition, la création et le traitement de nouvelles données, un collaborateur est soit un observateur (un plongeur) chargé de l'acquisition et la création de données soit un opérateur qui manipule ces données pour en extraire des valeurs de facteurs décrivant l'état écologique des sites (fréquence de taxons, profondeur…). Une relation de liaison entre l'entité Segment_objet et l'entité Transect_objet décrit une dépendance type un - l'infini, c'est a dire qu'un segment ne peut appartenir qu'à un transect, alors que le transect peut être constitué d'autant de segments qu'il est nécessaire. Cette relation permet de modéliser plus précisément le contenu d'un transect composé de plusieurs segments, chaque segment étant décrit par un profil prélevé du site d'étude par quatre attributs : orientation, inclinaison, rugosité et recouvrement majeur.

[91] Une entité est la représentation d'un élément matériel ou immatériel dans un système d'information
[92] dans le cadre de CIGESMED, se dit d'une trajectoire où l'on prend des photos de quadrats non bout à bout. Ces transects peuvent donc se faire à bathymétrie constante.

En seconde phase, une table « indexation » (en orange dans la figure 34) a été ajouté à la base de données. Cette table a pour objectif de représenter la base de données sous forme des graphes construits à l'aide des qualifications communes à tous les objets de la base. Elle indexe les liaisons objets-attributs entre les différentes données de la base à l'aide des identifiants uniques. Pour cela, un champ « UID » structuré comme les U.U.I.D. est ajouté à toutes les tables objets de la base. Cette table a une structure correspondant à une base de données graphes et est utilisée pour générer les flux X.M.L. et J.SON utilisés plus tard par le prototype présenté au Chapitre 4 partie 2.1 de ce manuscrit.

Focus sur l'indexation des données CIGESMED

Le principe de l'indexation repose sur l'utilisation d'un répertoire d'identifiants pour accélérer la recherche d'informations. Il permet d'éviter de dupliquer des informations et faciliter la mise à jour à chaque modification. Des identifiants uniques ont été ajoutés aux tables "objets" de la base de données « CIGESMED » afin de pouvoir traiter les relations entre objets sans le contenu de la base de données.

Un identifiant qualifié d'"unique" est utilisé en informatique pour désigner un enregistrement dans une base de données sans ambiguïté (il doit donc ne pas avoir la même valeur dans une autre table). Dans le cas des données CIGESMED, c'est un identifiant du type U.U.I.D. qui est généré aléatoirement *via* un algorithme (Figure 35) basé sur l'adresse MAC de la machine physique et sur l'heure où est généré l'identifiant :

```php
function getGUID(){

    if (function_exists('com_create_guid')){
        return com_create_guid();
    }else{
        mt_srand((double)microtime()*10000);//optional for php 4.2.0 and up.
        $charid = strtoupper(md5(uniqid(rand(), true)));
        $hyphen = chr(45);// "-"
        $uuid= substr($charid, 0, 8).$hyphen
            .substr($charid, 8, 4).$hyphen
            .substr($charid,12, 4).$hyphen
            .substr($charid,16, 4).$hyphen
            .substr($charid,20,12);

        return $uuid;
    }
}
```

Figure 35 : L'algorithme de génération des identifiants uniques, en open acces sur le web, est montré ci-dessus. Ces identifiants uniques sont produits en utilisant des composantes pseudo-aléatoires codées sur 128 bits ainsi que les caractéristiques d'un ordinateur (numéro de disque dur, adresse MAC, etc.). Un UUID se présente habituellement sous la forme de groupes de caractères hexadécimaux en minuscules séparés par des tirets. La structure de l'identifiant généré est de cette forme :

D5657E09-8875-9D78-2602-246D434F7A31 et génère un des 10^{38} possibilités d'identifiants laissées par un code alphanumérique de cette longueur. Pour donner une idée de ce que représente ce nombre, d'après la NASA, il contient 10 milliards de fois le nombre estimé d'étoiles dans l'univers. Un UUID est donc initialement conçu de manière à être unique dans le monde ; cependant, le risque que deux ordinateurs produisent un même identifiant est non nul. Un événement peu

probable de ce type n'aurait pas de conséquence sur les résultats issus d'une analyse de ces données, et la comparaison avec d'autres erreurs plus impactantes (souvent d'origines humaines) justifie de négliger l'impact d'une génération de deux U.U.I.D. identiques (El Guerrabi, 2014).

3. Regard critique concernant les systèmes de gestion et d'entreposage des données dans le cas d'une application aux données issues des programmes CIGESMED et DEVOTES

La métadonnée : un outil nécessaire mais pas suffisant

Une métadonnée donne des informations sur des enregistrements, des jeux de données ou des services rendus par des données.

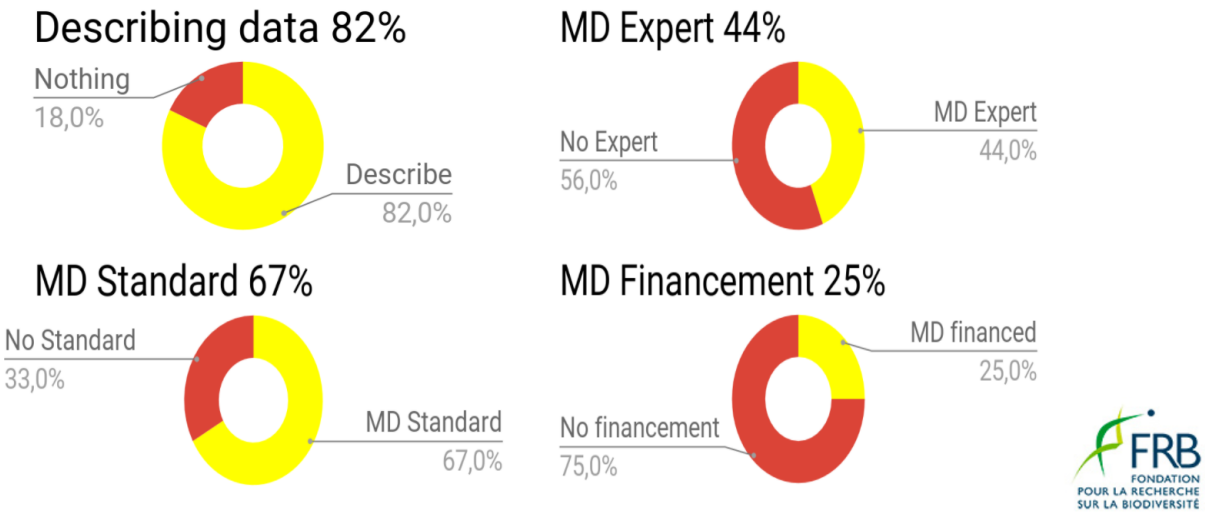


Figure 36 : Résultat d'une enquête réalisée par la FRB en 2015 à laquelle 40 établissements producteurs de données dans le domaine de la biodiversité ont répondu (qui sont probablement les plus sensibilisés). Les données n'appartiennent pas à tous les répondants (pour près de 50%, les données appartiennent aux établissements). 82% de ces 40 établissements décrivent leurs données, et 67% utilisent des normes de métadonnées. Les leviers pour produire des métadonnées s'appuient dans 44% sur le soutien d'un expert externe, et 25% allouent un financement spécifique à cette tâche, et donc un support avec ressources matérielles ;

Ces métadonnées peuvent être très réduites (libellé, date de création, point de contact, projection cartographique utilisée dans le cas de données géographiques par exemple) ou très détaillées (mesures de qualité des données pour chaque jeu de données voire même chaque enregistrement, mode de création et opérateur pour chaque donnée, version(s) et

niveau de maintenance des équipements utilisés pour effectuer la mesure, contraintes d'utilisation...).

La structuration et le renseignement des métadonnées sont souvent négligés lors de l'utilisation d'un système d'observation (Figure 36) et lors du renseignement du système d'information associé, alors qu'elles s'avèrent indispensables à la réutilisation par des tiers de la donnée d'origine ou de son analyse et de sa compréhension en tenant compte de son contexte.

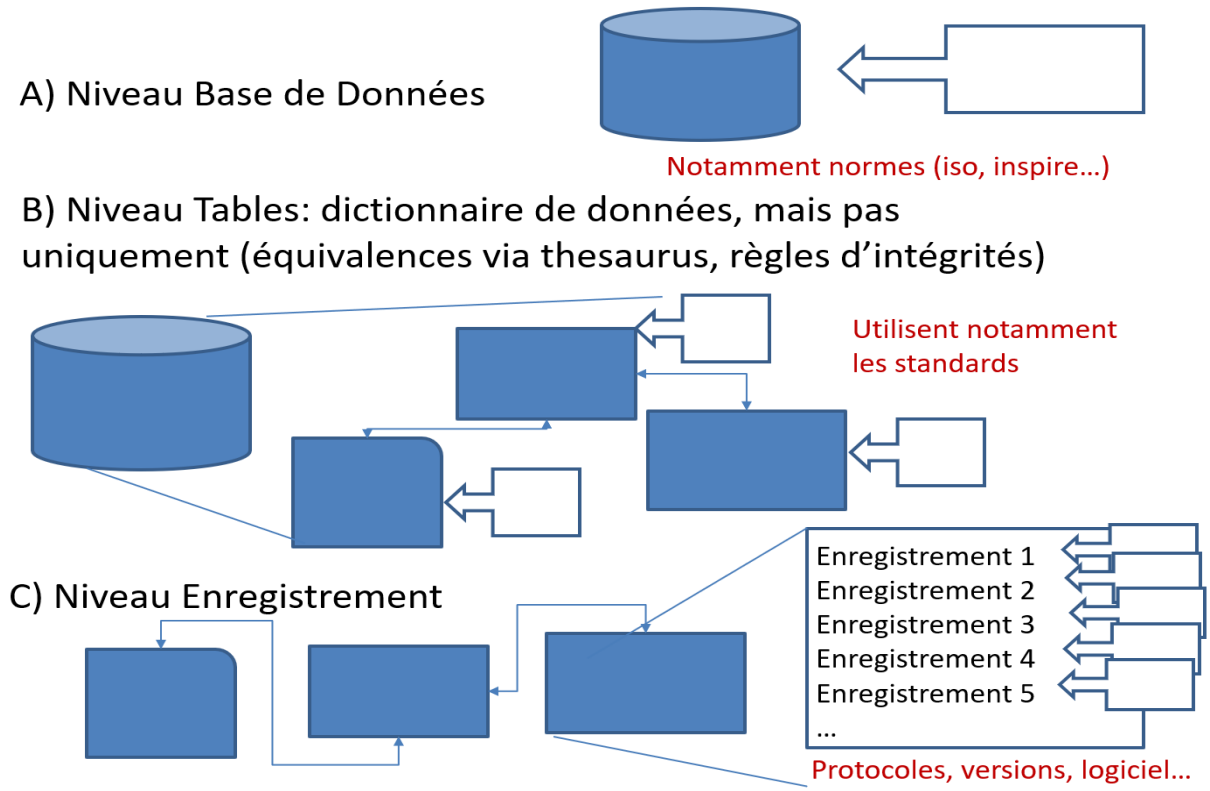


Figure 37 : Une métadonnée est une donnée qui définit ou décrit d'autres données. Ces métadonnées sont souvent normées, pour décrire des bases de données (A), des producteurs de données, ou la structure des données en elle-même (B). Le concept de métadonnées est remplacé dans de nombreuses recherches par la dénomination descripteurs de données qui sont des annotations décrivant la qualité des données de la base de données dans son ensemble jusqu'au niveau des enregistrements, voire même l'état d'un enregistrement à un temps T (comme par exemple un numéro de version où la version du protocole ou du référentiel utilisé pour la produire (C)).

Même dans le cadre des programmes CIGESMED et DEVOTES, et à cause d'une mise en place, de tests puis de validations du protocole en plusieurs itérations, les données résultant des quadrats photo ont été complétées par des métadonnées après une grande partie des récoltes de données. De ce fait, certains champs qui auraient été utiles resteront non

renseignés, notamment au niveau des enregistrements. Pourtant, ce niveau renseignement de métadonnées s'avère le plus utile pour mieux analyser les données et/ou formuler de nouvelles hypothèses à partir des données (Figure 37).

Objectifs de la métadonnée à large échelle, en général, puis pour les programmes CIGESMED et DEVOTES

Les métadonnées permettent de savoir à quel niveau deux jeux de données d'origines différentes peuvent être analysés ensemble de manière cohérente (nombre de champs communs, formats communs de ces champs, échelles de temps ou d'espace décrits, etc.). Cette analyse des possibilités d'"alignement" peut concerner soit le même système observé mais avec des systèmes d'observation différents, soit deux systèmes d'observations identiques mais appliqués à deux systèmes observés différents, ou dans des cas plus particuliers mais plus fréquents sur deux systèmes observés différents en utilisant des systèmes d'observation différents. Ce troisième cas est pourtant contraire aux désirs les plus fréquents du scientifique qui souhaite proposer un comparatif fixant le système observé et le système d'observation pour en montrer la différence à deux états T1 et T2 sans biais d'observation. Nous montrerons dans le reste de cette étude que dans le domaine de l'écologie, *a fortiori* dans le domaine de l'écologie marine, ni le système observé ni le système d'observation n'ont de paramètre stable et donc qu'il est impératif de connaître et de conserver un maximum d'informations sur ces facteurs de variabilité. Ces facteurs de variabilités sont appelés dans leur ensemble un « contexte ».

Dans la plupart des cas, les contextes ne sont pas ou peu décrits, ce qui handicape une réutilisation efficace des données. Dans des situations plus vertueuses, ces contextes peuvent être renseignés, parfois très précisément, mais leur utilisation est empêchée par les formats de publications et de mise à disposition. Dans le pire des cas, sous la forme d'un PDF ou d'une image qu'un utilisateur potentiel, pour des raisons de rentabilité (et parce qu'il n'en a souvent pas les moyens humains) n'essaiera jamais d'exploiter. Dans des cas intermédiaires, les fichiers de données sont fournis en un seul format dont i) la bonne exploitabilité et le contrôle qualité n'ont pas été pensés, ii) l'évolutivité des technologies pour lire ou exploiter la donnée a été négligée.

Pour illustration et sans être exhaustif, on peut citer quelques exemples de mauvaises pratiques :

- Un format unique comme un tableau en H.T.M.L. impossible à intégrer dans un tableur.

- Un fichier C.S.V. avec un séparateur de champs utilisé aussi dans les champs, ce qui rend les procédures de “parsing[93]” peu efficace,
- Une métadonnée globalisée et donc potentiellement difficile à redistribuer sur les enregistrements),
- Un téléchargement des données uniquement sous la forme d’un “Dump[94]” qui nécessite non seulement de connaître la technologie utilisée mais aussi de connaître la structure des données pour pouvoir faire des requêtes

Dans tous ces cas, le “ticket d’entrée” pour espérer réutiliser la donnée nécessite de faire appel à un curateur et/ou un technicien, ce dont ne bénéficient que peu de réutilisateurs potentiels.

Afin de pallier le manque de personnel qualifié, et afin de garantir un accès facile aux données, il est important de se mettre au niveau de chaque utilisateur potentiel. Une documentation trop volumineuse est un des exemples courants de frein à la réutilisabilité des données. De la même manière, le data paper, qui doit avoir un format scientifique, n’est du coup pas forcément l’outil unique pour diffuser de la donnée en dehors de la sphère scientifique. Enfin, dans la description des données, il ne faut pas oublier de décrire les champs servant de descripteurs de données ainsi que les variables aux niveaux les plus fins (c’est à dire de l’enregistrement, avec par exemple la version de la donnée, les validations, la version des protocoles et/ou des référentiels).

Méthode de cartographie des métadonnées dans le cadre de CIGESMED

Afin d'établir une carte dynamique des métadonnées dans le cadre de CIGESMED, il a été nécessaire d'identifier les types possibles de métadonnées, puis définir les champs obligatoires, d'une part ceux imposés par les normes européennes, d'autre part ceux rendus nécessaires à une description sans équivoque des données de terrain issues du protocole CIGESMED.

Une description dite robuste de chacun de ces champs a été réalisée (cf glossaire) afin de permettre de comprendre et de partager la signification exacte d’une variable et de ses valeurs et ainsi améliorer la reproductibilité des données et leur interprétation. Cette description a été réalisée dans le même temps que le développement de protocoles de terrain (WP2 de CIGESMED) et de mesure de variables de contextes et améliorée de manière itérative pour pouvoir être réutilisée dans le cadre des approches génétiques (WP4 de

[93] Le parsing est une analyse syntaxique d’un flux de caractères (X.M.L., J.SON ou autre) qui permet soit de le segmenter en éléments plus petits, soit d’utiliser un motif, constitué d’un ou plusieurs modèles pour extraire du flux les données qui correspondent au motif, en vue de les manipuler.

[94] Un dump est une sauvegarde des bases de données, dans laquelle on peut choisir de conserver la description de la structure des données.

CIGESMED) et les constructions d'indices (WP3 de CIGESMED) [voir annexe 3]. La validation des protocoles a demandé de tester chaque facteur mesuré et son efficacité descriptive, en prenant notamment en compte le niveau de compréhension par les observateurs mais aussi par les potentiels utilisateurs n'ayant aucune connaissance du domaine. Afin d'avoir un bon niveau de renseignement, les métadonnées ne doivent pas être trop complexes (pour être remplies par le plus grand nombre et avoir une réelle cohérence à grande échelle). Elles doivent néanmoins être suffisamment robustes pour pouvoir être réutilisées dans des contextes différents par des non-spécialistes. Malgré cette volonté simplificatrice en vue d'améliorer l'accessibilité de l'information, il faut noter que dans le cadre de l'expérience en vraie grandeur que représente CIGESMED, la maîtrise de la sémantique de tous les termes par les observateurs[95] a été jugée difficile et fastidieuse à obtenir, et que l'amélioration des outils de formation semble encore nécessaire.

Contenu des métadonnées dans le cadre de CIGESMED

Les métadonnées recueillies concernent :

- La cartographie des profils de contexte et des assemblages d'espèces
- La bibliothèque de taxons communs aux différentes régions ou particuliers à une région marine pour décrire le contenu des quadrats photo
- Les quadrats photo par site et la description des sites
- L'inter-étalonnage des types et modes d'acquisition des données (matériels et conditions)
- L'échantillonnage permettant l'approche phylogéographique (nommage des spécimens en particulier)
- L'échantillonnage de substrats (sous forme de grattages de surface)

Le modèle de données utilisé pour développer la base de données y insère les métadonnées et les complète en y ajoutant des règles d'intégrités. Les métadonnées et les interdépendances entre données sont décrites dans le dictionnaire de données et permettent d'éviter les problèmes de polysémie.

[95] La définition des formats de métadonnées a été réalisée lors des différents ateliers organisés lors des séminaires CIGESMED et en ligne via des formulaires (exemple : http://www.cigesmed.eu/Structural-descriptors-on-species).
Ces formats sont basés sur des recommandations "nationales" (normes de métadonnées SINP, agrégateurs nationaux comme le MNHN) et les recommandations internationales fournies par la directive INSPIRE, ou le TDWG). Ils utilisent les standards et langages proposés à l'échelle internationale (EML, XML, JSON). L'utilisation de normes linguistiques s'est avérée difficile à appliquer à tous les facteurs surveillés et doit être ajustée lors de la modification des data papers pour chaque ensemble de données.

4. Discussions et perspectives d'amélioration concernant les systèmes de gestion et d'entreposage des données appliqués aux données d'écologie marine.

Cycle de vie des données et des métadonnées dans le cadre de CIGESMED et de DEVOTES

Lors de la mise en œuvre des protocoles, des modifications ont été rendues nécessaires sur les données car des améliorations ont été apportées sur les modes d'acquisition. D'autres modifications se sont faites « à l'usage », en considérant les ambiguïtés relevées par de nouveaux observateurs ou les problèmes de compréhensions d'opérateurs distants. Afin de conserver leur utilité, les métadonnées doivent évoluer en même temps que les données, sur le même rythme itératif.

Ces itérations ont permis d'améliorer la méthode de description des métadonnées (en allant vers les méthodes les plus simples possibles) et ont été insérées dans les protocoles disponibles en ligne sur le site web de CIGESMED.

Cette construction d'un système d'information évolutif concernant tous les pays est nécessairement itérative et ne peut être considérée comme finalisée tant que les protocoles ne sont pas mis en œuvre dans les différents contextes. La première description de métadonnées décrivait de manière théorique un système de métadonnées nécessaires, qui évoluait ensuite selon les besoins et la mise en œuvre des protocoles (années 1 et 2). D'autre part, un suivi de l'évolution de ces normes a été mis en œuvre et sera nécessaire pour maintenir le niveau d'interopérabilité du système d'information. Ce maintien à long terme de l'interopérabilité, pierre angulaire de l'accessibilité et de la réutilisation des travaux réalisés se confronte à la limite dans le temps des moyens alloués (crédits à dépenser sur la période de financement du programme de recherche), et au peu de stabilité des équipes, laboratoires et instituts de recherche, et ceci particulièrement en France où l'on crée facilement des structures dites "en mille-feuille".

L'organisation des acteurs de l'observation à large échelle, si on souhaite que ses données soient réutilisables sur le long terme ne peut pas être déconnectée de la recherche mais pour autant doit être indépendante. Les moyens nécessaires pour mettre en oeuvre cette organisation ne peuvent pas dépendre de programmes de recherche soumis aux effets de mode et limités dans le temps. La donnée doit être conservée sur le long terme et maintenue à jour.

De prime abord, la description et la cartographie des métadonnées testée dans le cadre des programmes CIGESMED et DEVOTES devra évoluer non seulement en fonction de

l'évolution des normes ou la nouvelle élaboration de normes communes, mais aussi et surtout à cause des besoins générés par les nouvelles utilisations des données. À l'avenir, de nouvelles « couches » d'information seront nécessaires (types de données intermédiaires) qui vont accroître le potentiel d'utilisation de ces données et leur intégration dans les processus d'exploration de données transdisciplinaires. Ces couches supplémentaires, qui enrichissent la description des données, sont basées sur un thésaurus en cours de construction (- Descripteurs structurels sur espèces et taxons - Descripteurs anthropocentriques sur espèces et taxons - Descripteurs sur espèces et taxons - Descripteurs contextuels de l'étude des sites). Pour développer ce thésaurus, la communauté CIGESMED s'appuie sur un outil développé par le CESAB (Centre de synthèse et d'analyse de la biodiversité), le Thesauform[96], qui permet pour chaque terme / qualificatif de données, de proposer une définition puis, par votes successifs, d'obtenir un consensus sur la définition des descripteurs. Ce travail devrait être poursuivi après la fin du projet CIGESMED, avec un consortium IndexMEED, si celui-ci parvient à débloquer les moyens nécessaires.

Afin de cartographier les données à grande échelle dans un système générique, l'architecture nécessaire a été définie lors de l'atelier E.G.I. (European Grid Infrastructure) à Amsterdam et décrite dans David et al. (2016b) et fait l'objet du chapitre suivant.

[96] http://thesaurus.cesab.org/ThesauformCesab/home

Chapitre 4 : Une architecture pour la ré-exploitation et la fouille des données hétérogènes en écologie

Ce chapitre utilise des éléments des publications David *et al.*, 2015, El guerrabi 2015, David *et al.*, 2016a, David *et al.*, 2016b, David et al., 2017.

1. Généralités, questionnements et hypothèses concernant la ré-exploitation et la fouille des données

1.1 Objectifs à court et long terme concernant la donnée

La majorité des données scientifiques est souvent produite pour une “consommation immédiate”. C’est aussi le cas dans le cadre de CIGESMED et DEVOTES ou le premier enjeu est de faire une démonstration scientifique étayée par l'affirmation ou l’infirmation d’une hypothèse. Elles sont issues d’un dispositif expérimental, de simulation ou d’observation à plus ou moins longue durée de vie, dont on connaît souvent mal la répétabilité / reproductibilité. Les programmes européens imposent aujourd’hui comme une condition *sine qua non* la présence d’un Data Management Plan, surtout lorsque les données produites sont relatives à un suivi environnemental. Pour autant, malgré un objectif affiché de tests et recherches de système de suivi et d’évaluation du bon état environnemental, chaque programme a des financements finis fixés sur une période donnée, ce qui rend plus difficile la possibilité de rendre la donnée réutilisable a posteriori. Aucun dispositif de reconnaissance d’un chercheur n’est actuellement efficace (même si la généralisation des data papers pourraient changer la donne). Des efforts de synthèse et de bancarisation sous forme de “données élémentaires d’échanges” sont proposés par le S.I.N.P.. Des référentiels et “couches de références” sont créés et tenus à jour par le M.N.H.N. et l’I.N.P.N. A plus large échelle, l’évolution des systèmes internationaux en terme de nombre de données agrégées (par exemple, le GBIF qui proposait presque 400 millions d’occurrences accessibles en Mars 2013 va dépasser le milliard d’enregistrements en 2018) marque un changement profond de ces systèmes de production et de bancarisation des données. Pour autant, les “producteurs vertueux sont encore très peu nombreux et leur nombre évolue pour l’instant plus lentement (le GBIF rassemble au niveau mondial seulement 1200 contributeurs en 2018 !).

C’est l’automatisation de systèmes de reconnaissance [machine learning *p.e.* à partir de photos d’ailerons de requins (Hughes et Burghardt, 2016) et de plancton *via* des systèmes

de cytométrie en flux (Gorsky *et al.*, 2016)], deep learning (balbutiant dans le domaine de la biodiversité), screening autre fréquence du moléculaire à la donnée satellite (par exemple (Pesant *et al.*, 2015) qui dope l'évolution de données partagées. Par ailleurs, ces données sont dans la plupart des cas des occurrences simples (on note la présence, parfois l'absence) ou observations ponctuelles et opportunistes (et donc sans vue d'ensemble), des suivis de population (et donc, par exemple, 1000 occurrences peuvent uniquement concerner le même individu) et, dans le meilleur des cas, des fréquences (souvent des fréquences relatives par unité de surface dans le programme CIGESMED, d'unité d'observation pour les ARMS et ASUs du programme DEVOTES). Ces données sont des données dites "d'interprétation". Que ce soit avec les gestionnaires d'espaces naturels, ou même les spécialistes, seul un pan taxonomique ou les espèces et / ou taxons évidents à reconnaître sont prises en compte. L'observation photo, notamment du benthos, montre aussi ses limites à large échelle lorsqu'elle s'appuie sur des opérateurs en grand nombre, pour lesquels l'inter-calibration devient alors cruciale, à laquelle s'ajoute les effets de l'épuisement à la tâche qui peut nécessiter un *turnover* important des opérateurs. Ceux-ci sont d'ailleurs le plus souvent des stagiaires ou des contractuels précaires, dont l'embauche est liée à l'obtention d'un programme de recherche (ce qui démultiplie les coûts de formation sur le long terme). Néanmoins, les bases photo ainsi analysées et annotées doivent être conservées précieusement car elles permettront de développer des techniques d'apprentissage automatisé, d'abord supervisées par ces analyses humaines. Puis si les contextes correspondent, elles permettront d'utiliser des systèmes de classification non supervisés soit pour faciliter ensuite le rôle de l'expert, soit pour mener des analyses en beaucoup plus grand nombre. Notons aussi qu'aujourd'hui, ces observations d'occurrences ou de fréquences relatives d'espèces bancarisées sous forme d'image, de sons ou de vidéos, ne reflètent qu'un prisme d'observation de la biodiversité biologique (comme le montre le développement des E.B.V.). Pourtant, seule une compréhension des interactions entre ces espèces et communautés dans le temps et, selon les contextes, permettraient de mieux comprendre les aspects fonctionnels des écosystèmes et les enjeux liés à la modification de ces contextes.

1.2 Enjeux de la préservation de la donnée :

La nécessité de sauvegarder les données scientifiques est devenue une évidence. Mais lesquelles ? pour quel coût ? Dans un contexte de "big data", sauver veut nécessairement dire aussi "effacer, trier, nettoyer…". Durant cette démarche, l'essentiel est souvent oublié : les exemples de données sauvegardées mais rendues inutilisables par l'inaccessibilité du code pour les lire, les extraire ou les analyser pullulent…. De nombreux problèmes concernant la capacité des données à être réutilisés doivent être pris en compte et résolus (maintenabilité des workflows, processus, interfaces, fréquence d'accès, criticité dans les

algorithmes, réplicabilité, redondance, niveau d'abstraction…) pour éviter le constat finalement fréquent "*si j'avais su*…". La préservation de la donnée est une problématique complexe qui demande des ressources et des compétences sur le long terme (David *et al.* 2018, *in press in PREDON*).

Les méthodes de fouille de données ont vu le jour à la fin des années 90 (Fayyad *et al.*, 1996) et se définissent comme une discipline visant à extraire des connaissances nouvelles et compréhensibles pertinentes de l'analyse des ensembles de données préexistants. Cette discipline a évolué vers une approche de plus en plus complexe incluant l'écologie, entre autres disciplines. Bien que la plupart des producteurs d'information et des utilisateurs des disciplines scientifiques et de l'industrie considèrent celle-ci comme la piste la plus prometteuse pour faire des progrès et de nouvelles découvertes, l'écologie est encore en retard par rapport à d'autres disciplines (Peters *et al.* 2014). En particulier dans le domaine de la biodiversité marine et les systèmes socio-écologiques côtiers (SES) qui y sont liés, la production de données reste très coûteuse et avec un faible niveau d'automatisation. Les études sur les séries de données à long terme et / ou sur les grandes zones spatiales sont difficiles à réaliser et, lorsque plusieurs observateurs doivent être impliqués, malgré les nécessaires tâches d'inter-calibration décrites précédemment, la robustesse et la reproductibilité de l'observation sont souvent plus difficiles à obtenir.

Dans ce contexte, il est nécessaire d'analyser les relations entre les données hétérogènes pour une meilleure compréhension de la réalité se basant sur l'existant en terme d'observations (et de systèmes d'observation). La fouille de données sur des données hétérogènes est complexe. Elle est compliquée par un travail nécessairement plus long pour améliorer la qualité des données. Néanmoins, celle-ci permet l'extraction de connaissances ou de données pertinentes pour l'aide à la décision à partir de grandes quantités de données, en utilisant des algorithmes supervisés ou entièrement automatiques.

Acquérir une meilleure compréhension globale des équilibres des Systèmes Socio Écologiques (S.E.S.) et de leurs impacts sur la biodiversité est un défi scientifique majeur. Avancer dans ce processus de compréhension ne sera possible qu'en construisant et en testant de nouvelles méthodes pour l'interprétation commune de ces données hétérogènes.

Certains systèmes de recherche récents tentent de montrer des interdépendances logiques dans les S.E.S. en vue de faciliter la compréhension des liens et services fournis par la biodiversité et les écosystèmes aux sociétés humaines (Laporte M.A. *et al.*, 2014).

D'autres études portent sur l'intégration des connaissances déclaratives avec des données numériques et qualitatives (Gibert *et al.*, 2014) ou sur le post processus des résultats nécessaires pour fournir des connaissances compréhensibles à l'utilisateur final (Cortez *et al.*, 2012, Gibert *et al.*, 2012). Les méthodes d'exploration de données doivent être en mesure d'apporter de nouvelles perspectives à la recherche disciplinaire sur ces systèmes

complexes, en étudiant des objets interconnectés dans la réalité, mais encore trop étudiés séparément (chimie environnementale, génomique, transcriptomique, protéomique, métabolomique, écologie des peuplements, systèmes socio-écologiques ou écologie paysagère). A cette quête du chaînon manquant s'ajoute le challenge d'une compréhension de ces mécanismes en prenant en compte la spatio-temporalité intrinsèque du phénomène écologique.

1.3 Qualifications de la donnée

La qualification, outil pour l'interopérabilité

La qualification de données (qui correspond à une annotation dans le domaine de la sémantique[97]), vise à augmenter l'efficacité d'un système d'information par l'enrichissement d'une bases de données avec de nouveaux termes et/ou champs. C'est un des aspects de la curation de la donnée, particulièrement utile lorsqu'on cherche par exemple à agréger des données sur un même sujet qui n'ont pas le même type. Les données peuvent être des séquences de mesures de type numérique, et permettant une ordination (par exemple une température). On parle alors de données quantitatives ordonnables. Si ces données fonctionnent en un binôme de mesure (par exemple, la température et la luminosité à un instant T), le binôme est une donnée quantitative non-ordonnable. Dans le cas d'associations de différentes données, on parle de données multimodales. Cette multimodalité peut exister pour les données catégorielles, et celles-ci peuvent être ordonnées (par exemple petit, moyen, grand). La répartition entre catégorie des données dépend alors de la valeur que l'on a donnée à chaque limite. Enfin, beaucoup de données sont catégorielles et non ordonnées (comme par exemple rouge à pois noirs en comparaison avec bleu à rayure verte). Pour un facteur mesurable, lorsque différents systèmes d'observation produisent des typologies de données différentes, il faut les rendre comparables. Pour cela, il faut choisir le plus grand dénominateur commun et tenter de le faire progresser ensuite en améliorant les observations les moins précises. On requalifie les données avec un descripteur ayant des modalités communes.

La qualification comme outil d'enrichissement de la donnée

Cette qualification peut aussi être effectuée avec des informations contenues dans d'autres champs de la base de données, grâce à un système déductif. Elle permet la complétion de champs non renseignés dans une partie de la base de données ou dans une des bases de données agrégée avec les autres. Cette qualification peut aussi se faire en utilisant des

[97] Une annotation sémantique consiste à relier le contenu d'un texte à des entités dans une ontologie

référentiels extérieurs (par exemple renseigner la profondeur ou l'altitude à partir des coordonnées G.P.S. d'une observation) et/ou des analyses de la donnée (par exemple, qualifier sa qualité, sa précision en fonction de standards, lui donner une "note", etc.)
La qualification est un processus crucial pour la production d'analyses à large échelle. Pour des données multi sources, elle nécessite des données accessibles, un système d'indexation et une identification unique de la donnée, afin que ces annotations puissent être effectuées en dehors des systèmes d'information "métiers" (et pourquoi pas ensuite être réintégrées dans ces systèmes d'information "métiers").

La qualification de données est souvent catégorielle

Le cri que Sir Tim Berners-Lee[98] a lancé lors d'une conférence TED[99] en 2009 : « We want raw data, now !»[100], est symptomatique de l'état actuel de l'inaccessibilité des données sources, alors que la qualification ou la requalification de la donnée est plus efficace lorsque l'on peut partir des données initiales. Pour autant, la donnée source n'est la plupart du temps pas une donnée interprétée, et les systèmes de traitement qui fournissent une donnée explicite à partir de celle-ci ne donnent pas le détail des transformations faites. Il en découle des agrégations de données à large échelle plus difficiles et des données finalement moins précises. Dans l'état actuel et pour ces raisons, dans le domaine de la biodiversité, un grand nombre de descripteurs sont et resteront catégoriels voir même catégoriels et non ordonnés à large échelle. L'enjeu est aujourd'hui de pouvoir s'en servir pour construire des scénarios utilisables dans des contextes d'aide à la décision.

1.4 La fouille de données basée sur les graphes

Quelques notions sur les représentations basées sur les graphes

Un graphe est un ensemble de points que l'on appelle des nœuds (sommets en mathématique ou entités en informatique) reliés par des traits (segments en mathématique ou relations en informatique) ou flèches nommées liens (ou arêtes ou arcs). L'ensemble des liens entre les nœuds forme une figure similaire à un réseau (Aggarwal, C. & Wang, H., 2010). La représentation de données sous forme de graphes permet de relier des objets (champs/entité de la base de donnée ou valeurs de ces champs/attributs/relation) ayant des natures différentes (valeurs quantitatives, qualitatives ordonnées ou non ordonnées) ; les

[98] Timothy John Berners-Lee est connu comme un des principaux inventeur du World Wide Web en 1984. Il a l'idée de réaliser un partage de document en associant le principe de l'hypertexte à l'utilisation d'Internet. https://home.cern/fr/topics/birth-web
[99] https://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide?language=fr
[100] « Nous voulons des données brutes, maintenant ! »

attributs contenus dans un second champ décrivant une qualité de l'objet permettent de créer les liens entre ces objets et/ou de les pondérer. Les liens sont matérialisés par des descripteurs, c'est-à-dire des variables ayant plus d'une valeur possible. Les objets ayant le plus de liens en commun sont les plus proches, ceux ayant les liens les plus ténus (c'est à dire le moins de chemins possibles pour les relier entre eux et beaucoup de nœuds intermédiaires) sont les plus éloignés dans la représentation. On peut traiter les champs un à un ou bien en groupe de valeurs pour former - selon la combinaison de leurs valeurs respectives - un motif appelé patron (pattern en anglais). Ces patrons peuvent décrire des objets et/ou des liens et/ou des contextes. Les champs de « contextes », sont ensuite utilisés pour différencier les nœuds entre eux (couleur, forme, grosseur des nœuds). Ils ne participent pas à la topologie du graphe (c'est-à-dire à sa forme et ses propriétés[101]). Les motifs ainsi projetés dans le graphe peuvent être (i) dispersés, auquel cas les liens qui organisent le graphe ne sont pas liés aux éléments de contexte ; ou bien (ii) regroupés dans une ou plusieurs parties du graphe auquel cas il existe un lien entre la façon dont les nœuds sont organisés et un ou plusieurs contextes. Dans certains cas, il est possible de pondérer les liens, donnant plus d'importance à certains qu'à d'autres, ce qui agit sur la topologie du graphe. On parle alors de graphe valué[102].

Le clustering : Analyser les regroupements de nœuds pour aller un peu plus loin

Le clustering (classification non supervisée en français, mais c'est le terme anglais qui est le plus usité à la place de "classification", "groupe", ou "regroupement") consiste à regrouper des éléments. Cette agrégation est un élément-clé pour l'analyse de grands graphes. Une fois les groupes de nœuds obtenus, on peut réappliquer l'opération pour obtenir un clustering hiérarchique (basé sur une autre variable par exemple). Cette décomposition hiérarchique (ou multi-échelle) permet de modifier la complexité des algorithmes de fouille, de faciliter l'exploration des données, et de proposer une visualisation paramétrable : on parle aussi de navigation multi-échelle (Auber *et al.*, 2014 ; Lambert *et al.*, 2013). Les descripteurs quantitatifs sont en général transformés en classes de valeurs. L'analyse des fréquences relatives des "motifs" et des redondances entre "plus proches voisins" par rapport à leur fréquence dans tout ou partie du graphe montre l'importance des corrélations entre certains contextes et des clusters du graphe. La significativité de ces motifs peut ensuite être testée par des méthodes statistiques spécifiques analysant les qualités des clusters de graphes. Dans des graphes plus complexes où le nombre de combinaisons et de liens peut croître

[101] Ce terme est utilisé pour décrire la forme d'un graphe, donnée par les propriétés de ses composants (type de nœuds, nombre de nœuds, type de liens, propriété des liens, etc.).
[102] Graphe dont les arêtes sont pondérées (par exemple par une fréquence d'espèce), et dont la pondération agit sur les distances relatives entre les sommets.

exponentiellement, l'étude de la corrélation entre fréquences de contextes et "clusters" de nœuds peut demander de paralléliser les calculs[103] nécessaires à une investigation des parcours possibles. Selon la question scientifique sous-jacente aux objets représentés par un graphe, certains éléments dans les liens ou les nœuds peuvent être ignorés ou simplifiés.

Utilisation en écologie/environnement :

Les graphes permettent de représenter tout ou partie d'un système observé. Un tel système serait par exemple un ensemble de sites plus ou moins ressemblants en terme de composition d'espèces où les nœuds sont des sites, les liens sont les observations d'un taxon commune à différents sites et les clusters de nœuds correspondent aux sites les plus identiques et donc ayant le plus de liens entre eux. De la même manière, on peut représenter des individus reliés par la fréquence de leurs contacts, des groupes de taxons reliés par des traits, des groupes de personnes reliés par des réponses d'enquêtes sociologiques…) en intégrant des données de contextes de format différents (température, altitude, âges des individus, ensoleillement représentés par la grosseur ou la forme du noeud…). Ils permettent de mélanger des objets (graphes bipartites ou tripartites) ou de représenter de nouveaux objets complexes en combinant les valeurs de différents champs. Ils peuvent aussi être utilisés pour étudier le système d'observation ainsi que les efforts de prospection, ou pour avoir une approche visuelle de la répartition des compétences utilisées dans un projet ou évaluer un système d'information.

L'analyse des contextes liés aux clusters de graphes

L'analyse des fréquences relatives de ces "motifs" et des redondances entre "plus proches voisins" par rapport à leur fréquence dans tout ou partie du graphe montre l'importance des corrélations entre certains contextes et des clusters du graphe. La significativité de ces motifs peut ensuite être testée par des méthodes analysant les qualités des clusters de graphes. Le clustering est un élément-clé pour l'analyse de grands graphes (Figure 38).

[103] La parallélisation des calculs permet de raccourcir le temps de calcul en découpant ce calcul et en le répartissant sur différentes machines. La grille de calcul européenne permet de solliciter plusieurs dizaine de milliers de machines en même temps, ce qui permet par exemple de raccourcir un calcul qui durerait une année entière à quelques heures.

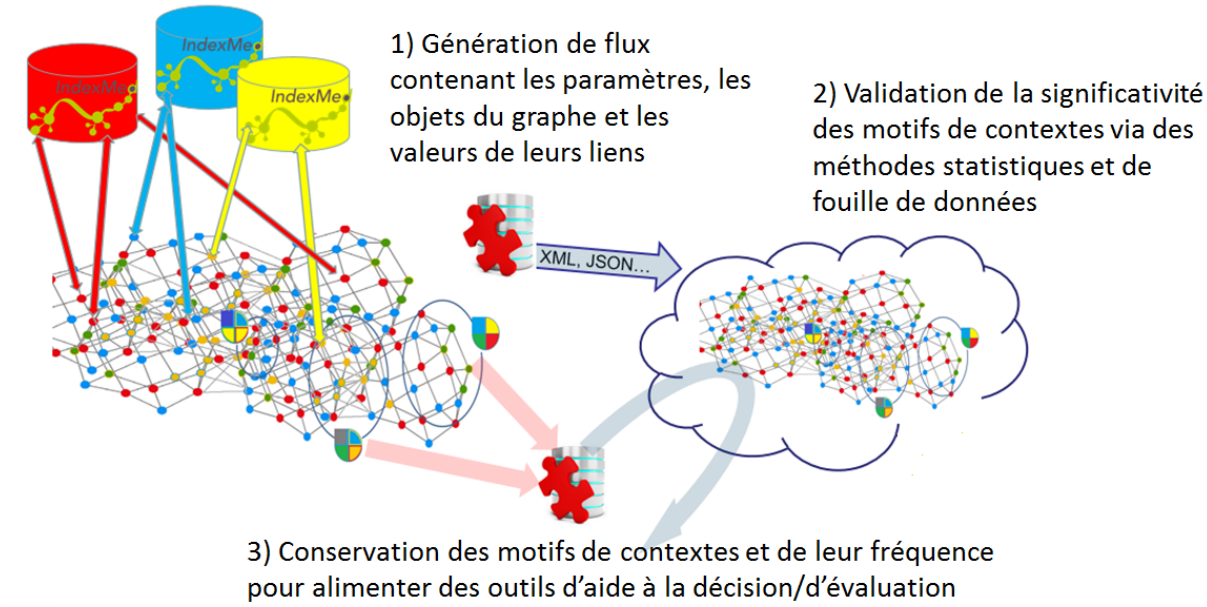


Figure 38 : Les trois grandes étapes de la fouille de graphe et de la comparaison de la significativité des contextes dans chaque cluster

Dans des graphes plus complexes où le nombre de combinaisons et de liens peut croître exponentiellement, l'étude de la corrélation entre fréquence de contextes et "clusters" de nœuds peut demander de paralléliser les calculs nécessaires à une investigation des parcours possibles (Figure 39). Selon la question scientifique sous-jacente aux objets représentés par un graphe, certains éléments dans les liens ou les nœuds peuvent être ignorés ou simplifiés. Au sein d'IndexMEED, cet aspect prospectif dans les graphes est en cours d'élaboration avec la communauté S.T.I.C..

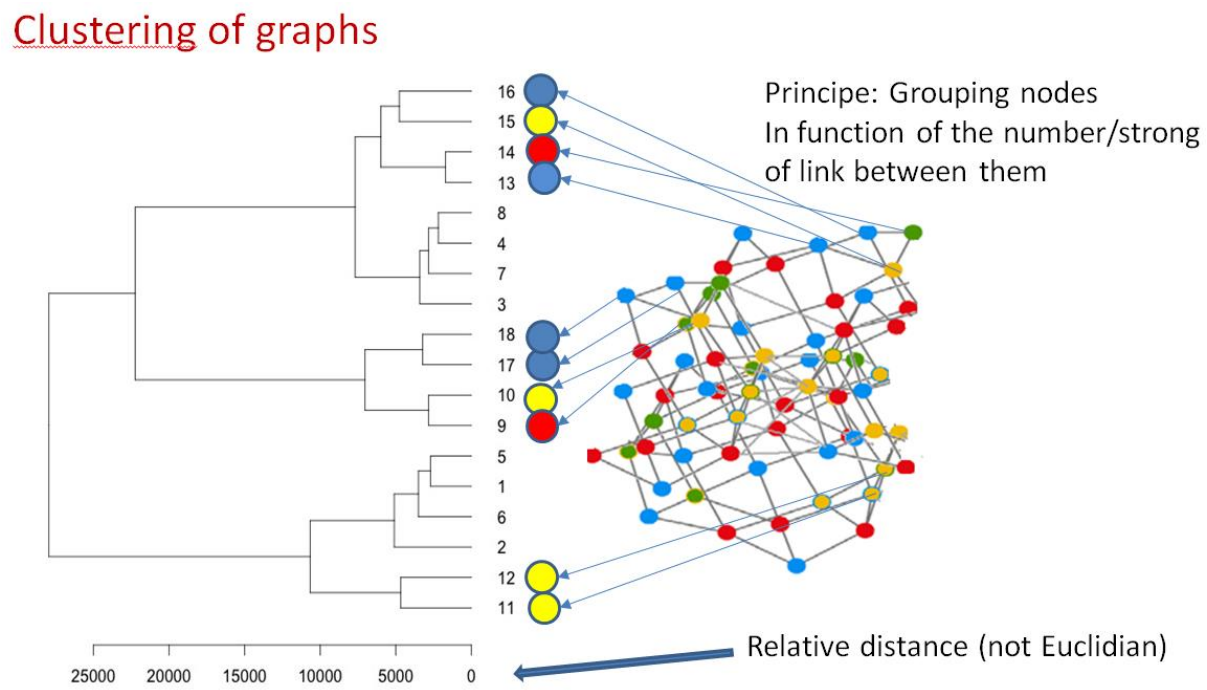


Figure 39 : Principe du clustering de graphes : plus il y a de liens, plus les nœuds sont proches. On peut pondérer les liens selon leurs types si le graphe en contient plusieurs différents. Dans les clusters, on teste ensuite la co-occurrence de toutes les combinaisons de facteurs possibles.

Les graphes permettent d'appréhender tout ou partie du système observé (un ensemble de sites ayant une caractéristique en commun par exemple) en intégrant des données de contextes de formats différents (valeurs numériques ou catégorielles, ordonnées ou non, simples ou multimodales). Ils permettent aussi d'étudier le système d'observation ainsi que les efforts de prospection, ou d'avoir une approche visuelle de la répartition des compétences utilisées.

2. Méthodes de conception des processus de test

2.1. Conception de l'architecture et des services

La méthode proposée pour utiliser des graphes pour fouiller des données hétérogènes en environnement et écologie a été élaborée dans le cadre d'IndexMEED. Elle s'appuie sur une architecture d'indexation elle-même définie lors du séminaire «Design your infrastructure» organisée par European Grid Infrastructure http://www.egi.eu/ (avril 2016) (David *et al*., 2016b). Celle-ci a l'avantage de pouvoir, soit laisser les jeux de données chez le repository officiel (ce sont alors des flux paramétrables qui sont interrogés), soit de se baser sur des

imports (CSV, XML, JSON…). Il a été décidé de construire les graphes *via* un prototype de visualisation de graphes (sur un serveur WEB) à partir d'informations agrégées grâce à des points nodaux d'indexation[104] et de qualification des données de contexte sur l'environnement (qui peuvent être hébergées sur des serveurs de partenaires externes). Le système d'information et *a fortiori* les données auxquels il organise l'accès, doivent être évolutifs. L'architecture du système a été conçue pour organiser ces itérations qualitatives (Figure 40) selon un modèle qui se veut générique et transposable à tout type de données scientifiques. Les services d'indexation ont été voulus réplicables à la manière d'un plugin pour permettre la création d'un nouveau point nodal pour chaque nouvelle thématique disciplinaire ou interdisciplinaire, ou le développement d'une thématique à un niveau géographique différent. A l'image des registrars répliquant les bases d'IP, ces index se recopient d'un système à un autre avec toutes les qualifications de données en rapport avec son périmètre thématique ou géographique de recherche. Une généralisation de ce système devra s'appuyer sur une gouvernance des autorités gérant un point nodal, et une gestion des autorités produisant ou transformant la donnée et administrant un service d'accès aux données.

[104] On appelle point nodal d'indexation un service web qui va moissonner les flux de données pour indexer les enregistrements en utilisant des descripteurs communs à ces flux de données. Un point nodal peut être spécifique d'un domaine et/ou d'une zone géographique, et peut sélectionner une partie d'un flux servi par un autre point nodal, ou regrouper des flux de plusieurs points nodaux.

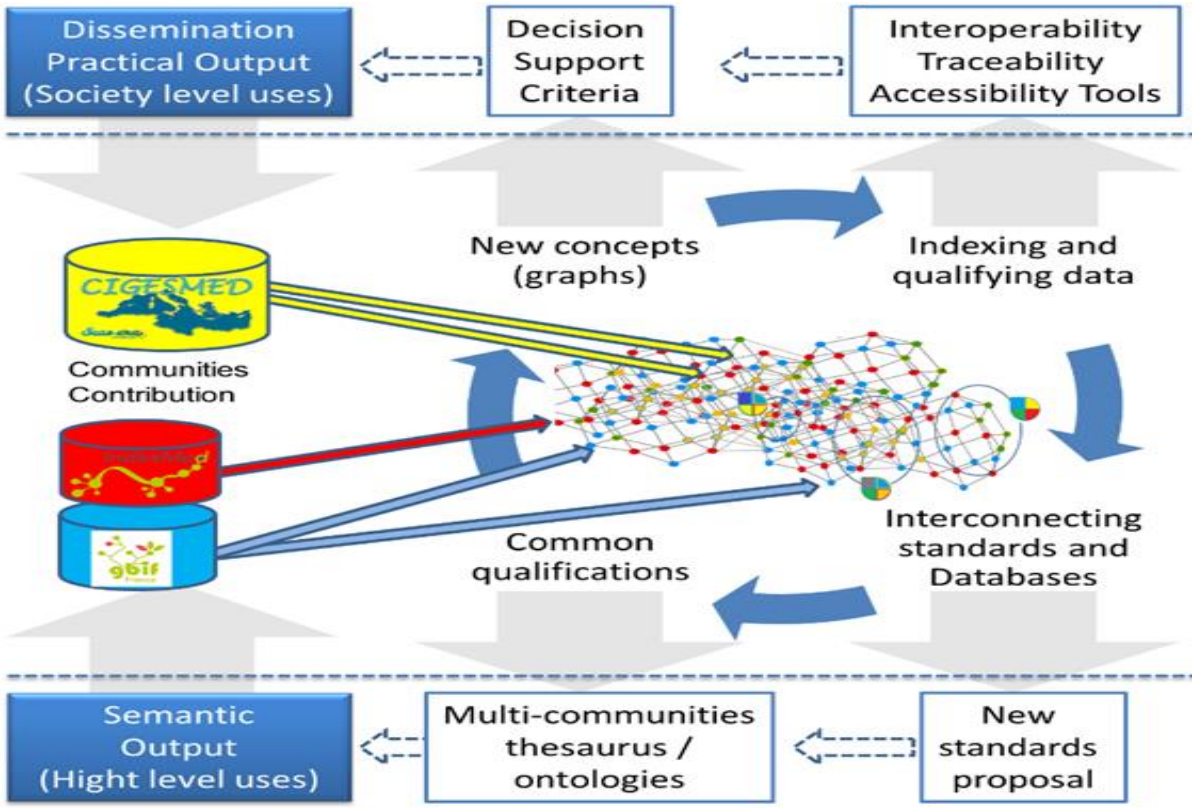


Figure 40 : Cette figure montre l'approche itérative utilisant les jeux de données CIGESMED et les workflow IndexMed. Les données provenant de différents fournisseurs/producteurs sont tout d'abord indexées par le prototype, puis qualifiées avec des descripteurs communs. Ces descripteurs sont choisis en fonction des standards disponibles si il y en a , sinon, de nouvelles propositions sont faites en intégrant au mieux les travaux antérieurs. On aboutit à une qualification commune des données (avec le processus de curation si nécessaire, il est alors possible de construire des graphes en effectuant des hypothèses à partir des données et les valeurs de descripteurs communs aux enregistrements étant les liens entre ces enregistrements (des objets de la base de données). À large échelle, ces descripteurs peuvent porter de nouvelles informations (enrichir la base de données) et aboutir à de nouveaux concepts, ce qui permet d'alimenter des thésaurus voire des ontologies. L'analyse des clusters de graphes permet ensuite de faire émerger des motifs de contextes qui sont significativement fréquents dans certains clusters, et sont conservés dans une base d'aide à la décision ou en tant qu'indicateur pour un gestionnaire. Les "output" sont génériques, et peuvent s'appliquer dans différents champs disciplinaires, et les bases de données, si elles décrivent les mêmes objets peuvent être de disciplines différentes.

2.2. Développement d'un prototype

Figure 41 : L'interface permet de paramétrer des requêtes en langage "Cypher"[105] ou de les générer via des formulaires (à gauche de l'image) et de visualiser les données sous forme de graphe ou de flux aux formats JSON ou XML (ici, on peut voir le flux JSON comportant les informations rattachées à un noeud du graphe). Ce fichier est généré à chaque requête, il peut être sauvegardé pour être utilisé à distance par une autre plateforme (bouton {URL} en haut à droite) ou être téléchargé directement (bouton avec la flèche dans le nuage en haut à droite). Ces fonctionnalités permettront d'interroger l'interface depuis les centres de calcul pour réaliser la fouille des graphes préparés et peuvent être conservés sur le serveur ou dans un centre de donnée distant.

Dans un premier temps, un prototype de visualisation est développé et testé pour visualiser des analyses de quadrats photo. Le modèle de données est simplifié pour rendre toutes les données adaptables. Ce modèle prend la forme "objet / attribut / valeur d'attribut", un modèle qui peut être formalisé en différents langages (OWL, RDF) et permettant de connecter des systèmes distants et multiformats (RSS, WMS, WFS, XML, JSON voir figure 41). Pour réaliser cette visualisation, les requêtes sont configurables et décrites dans David *et al.* 2016a.

Pour pouvoir visualiser et explorer simultanément des ensembles de données différents et distribués, un "service de résolution d'objet" (c'est-à-dire un service Web qui trouve des liens et des dépendances entre objets indexés, basés sur l'identifiant d'objets uniques) est partagé par différents partenaires sur un point nodal expérimental. Il est destiné à être géré comme un logiciel libre, installable sur un site web sous la forme d'un plugin et s'appuiera pour les besoins de fouille de la donnée sur un service sur la grille européenne (notamment via EGI). Les objectifs de ce prototype sont de :

i) Lister les données et séries de données disponibles sous forme de flux,
ii) Analyser le contenu des flux de données et le niveau de correspondance avec des standards existants,
iii) Qualifier les flux, les séries de données avec des identificateurs uniques s'il n'y a pas d'identifiants,
iv) Suggérer des correspondances entre les champs aux utilisateurs et des correspondances entre lignes de données équivalentes.

Un des rôles de ce service de résolution d'objet est d'établir des liens entre des lignes de données avec des "identificateurs uniques" différents (par exemple, des versions différentes

[105] Cypher est un langage informatique de requête orienté graphe utilisé par Neo4j.

de données brutes, des interdépendances entre des données brutes et des données transformées, etc.).

2.3. Méthode d'animation d'ateliers concernant la curation et la visualisation sous forme de graphes

Des ateliers ont été mis en place dans le cadre d'IndexMEED (ateliers soutenus par le G.D.R. MaDICS et la F.R.B., organisés à la demande des participants au séminaire IndexMed de 2016). Ce séminaire a été suivi de deux journées d'échange et de réflexion sur les questions et hypothèses envisageables en considérant *a posteriori* chaque jeu de données[106], et en proposant des améliorations qualitatives de ces jeux de données. Un besoin impérieux de curation des jeux de données est ressorti des discussions après les premiers essais d'intégration des données dans le logiciel de visualisation Gephi[107] : la plupart des participants n'avaient pas structuré leurs données sous forme "nœuds et liens", et tous avaient de gros problèmes de consistance des informations contenues dans chaque champs. A l'évidence, pour effectuer les premières visualisations, un processus de curation plus élaboré et plusieurs étapes de qualification de données étaient nécessaires (présentées dans l'annexe 9.1).

Le chapitre 4 partie 1.3. présentait deux aspects de la qualification très importants pour construire des graphes (l'interopération d'une part, et l'enrichissement d'autre part). Le travail sur la qualité des données a permis de rendre les jeux de données plus consistants (c'est à dire, ayant une information systématique dans un champ de la base de données), et que cette information soit d'un type et d'un format comparable aux champs portant la même information dans toutes les bases considérées en vue de les agréger. Ce travail d'homogénéisation, première étape de la curation, devait être priorisé pour chacun des participants. Durant l'atelier curation, une formation théorique sur les différents aspects de la curation et les différents outils utilisables a été proposée, suivie d'exemples pratiques sur un jeu de donnée préparé puis un travail de "débroussaillage" sur chaque cas d'étude a clôturé la journée. Lors des ateliers visualisations, après de nouvelles présentations théoriques cette fois ci sur les aspects visualisations de données hétérogènes sous forme de graphes, l'essentiel des deux journées a été consacré à l'expérimentation de premières visualisations par chaque participant (résultats les plus aboutis présentés en annexe 9). Ces premières visualisations ont mis en évidence de nouveaux potentiels pour les données qui eux même

[106] https://indexmed2016.sciencesconf.org/program
[107] https://gephi.org/

nécessitent une curation plus ambitieuse, et une meilleure intégration des standards existants.

La conclusion de ces ateliers est qu'il est très difficile pour un thématicien d'expérimenter seul, même avec la meilleure des documentations, la visualisation de données sous forme de graphes, même s'il en est à l'origine. Cette conclusion est d'autant plus vraie à large échelle lorsque le contenu des bases de données n'est pas décrit précisément : le travail en équipe et interdisciplinaire est le seul à même de permettre la construction de systèmes pérennes de qualification de données, systèmes eux même condition *sine qua non* de la visualisation de données environnementales hétérogènes sous la forme de graphes.

3. Résultats de la phase de conception des processus et de la phase d'expérimentation du prototype

3.1. Schéma décrivant l'architecture du système d'information

A l'issue des ateliers de E.G.I. "design your e-infrastructure"[108], nous avons pu proposer un schéma simplifié présentant les grands types de "e-services" et leurs relations (figure 42)

[108] https://indico.egi.eu/indico/event/2895/

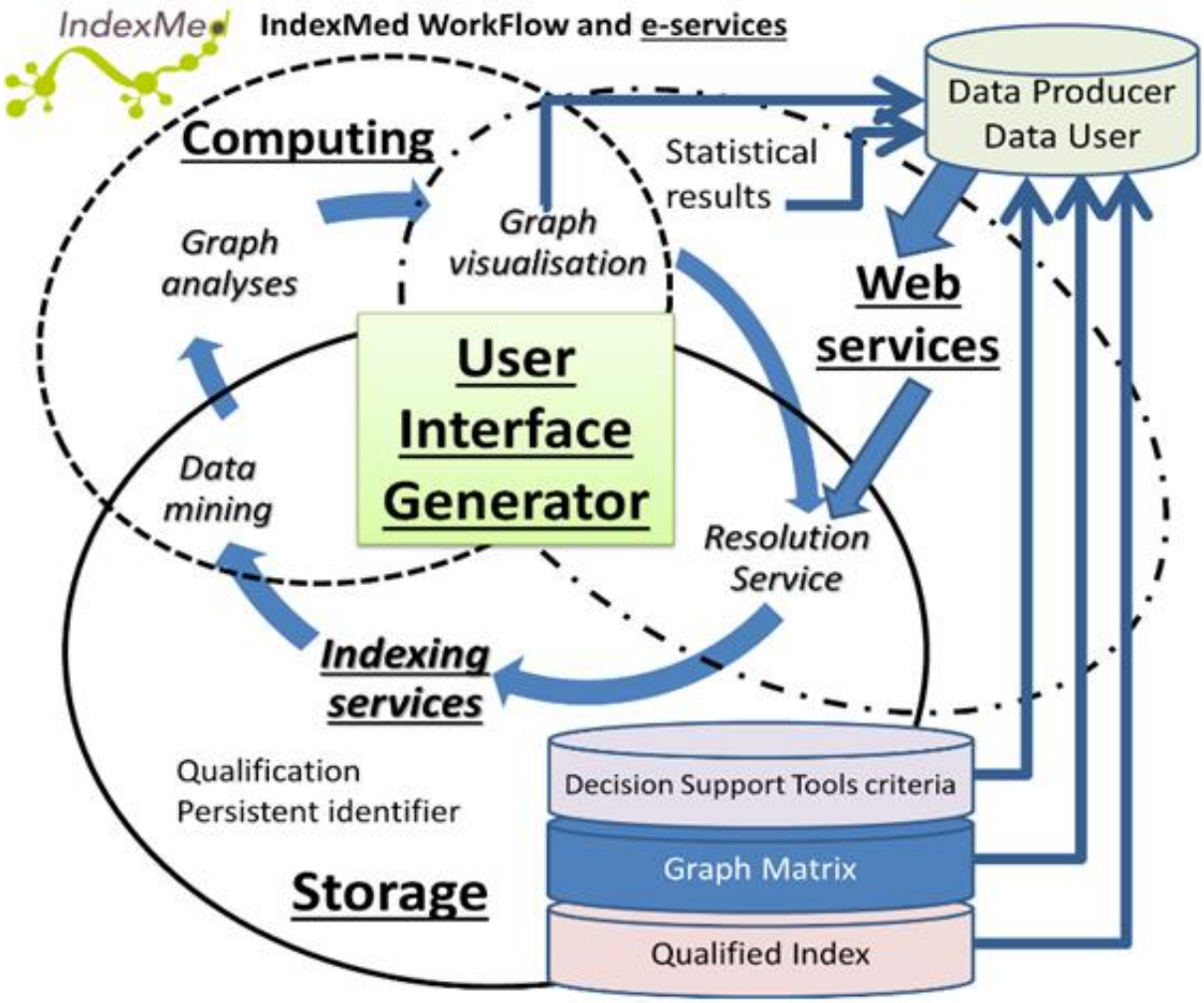


Figure 42 - Flux de travail IndexMed et e-services : Le service de résolution est capable de comparer l'index avec les données de stockage dans les e-infrastructures et d'autres XML distants, ou autres flux type JSON à partir de différentes bases de données. Lorsque cela est nécessaire / possible, il crée un identificateur persistant ou lie des ensembles de données ou des enregistrements de données avec des identifiants existants. Une interface scientifique, adaptée au niveau et aux besoins de chaque utilisateur permet un processus de qualification. Le service d'indexation accepte / gère les données pour des services de calcul comme l'exploration de données et l'analyse de graphiques. Les résultats statistiques et les modèles de graphes sont stockés et proposés comme un nouveau flux persistant (David et al., 2016b).

3.2. Le prototype et ses spécificités fonctionnelles

Présentation du prototype

L'agrégation de données multi-sources en écologie marine (Auber *et al.*, 2014) ou en archéologie a permis de tester ces approches grâce au prototype "open source" développé par le consortium IndexMEED (David *et al.*, 2015) dans le cadre du projet « Vigi-Geek[109] ».

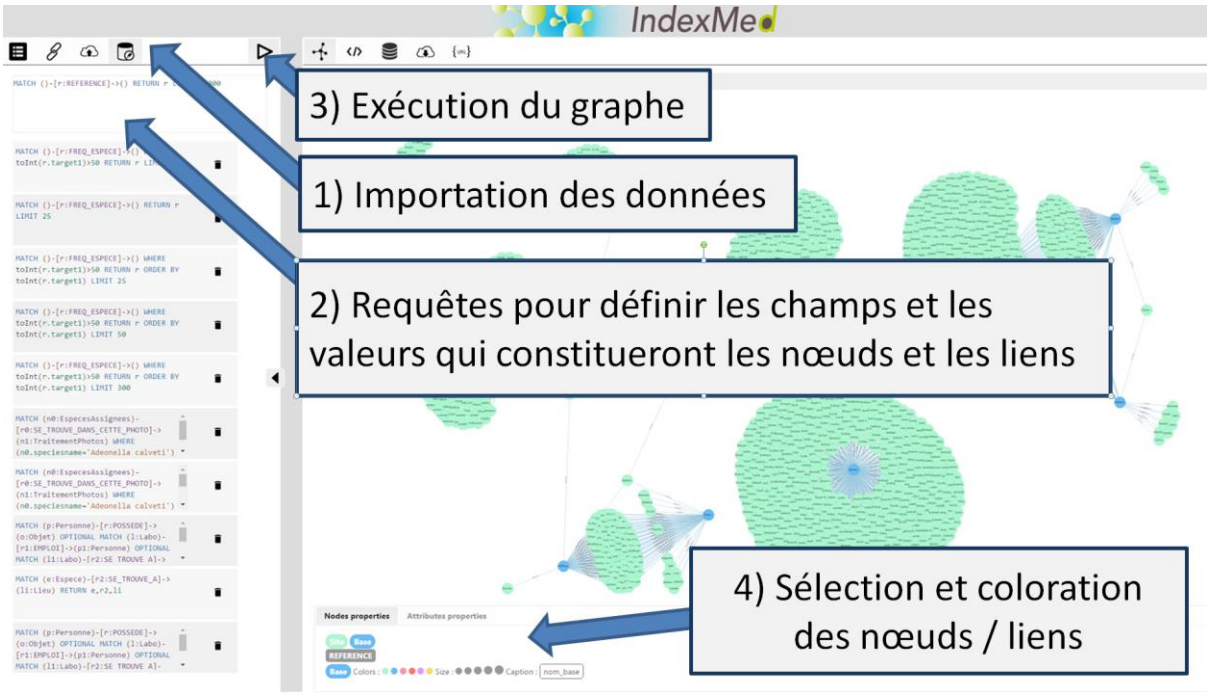


Figure 43 : Présentation générale du prototype d'IndexMed[110] de visualisation des données représentant dans cet exemple 1492 sites d'archéologie en vert et 12 bases de données sous la forme d'un graphe bipartite.
Les informations proviennent de l'import d'ArkeoGIS[111] mais pourraient directement être interrogées à distance (1) par le prototype sur les systèmes d'information des partenaires, (format JSON ou XML). La colonne de gauche permet d'importer, d'effectuer les requêtes (2) avec un formulaire ou le langage cypher et de les enregistrer. Un bouton (3) permet de lancer l'exécution de la nouvelle requête ou d'une requête pré-enregistrée. Le bandeau du bas (4) permet de configurer les couleurs des noeuds et des liens en fonction des valeurs de descripteurs (david et al, 2017).

[109] VIGI-GEEK : VIsualisation of Graph In transdisciplinary Global Ecology, Economy and Sociology data-Kernel

[110] Le prototype d'IndexMEED est accessible à http://data.imbe.fr/neo4j/

[111] "ArkeoGIS » est un projet porté par des archéologues, permettant d'agréger des données issues de bases aussi bien archéologiques que paléo-environnementales. Les données unifiées ont permis suite aux premières journées IndexMEED de produire le graphe de la figure 1.

Ce prototype (Figure 43) permet la mise en place de liens entre objets de bases de données différentes et distantes.

L'interface du prototype utilise Neo4j <neo4j.com/>, une base de données graphique mise en œuvre en java et publiée en 2010. L'édition communautaire de la base de données est sous licence GNU GPL v3. La base de données Neo4j et ses modules supplémentaires (sauvegarde en ligne ou haute disponibilité) sont disponibles sous licence commerciale. Le prototype d'IndexMed permet à un opérateur néophyte d'importer des données (en C.S.V., X.M.L. ou J.SON). Il permet d'interroger Neo4j pour produire le graphe et d'interagir avec lui à l'aide du navigateur Web. Le personnel technique d'IndexMed a développé un frontend Web spécifique à l'aide du langage Ajax / Jquery. Il peut être possible de demander une base de données demandant des objets spécifiques et des relations spécifiques entre eux, sans utiliser un langage de requête technique tel que S.Q.L. ou Cypher.

Le prototype est développé pour pouvoir être générique et permet d'intégrer n'importe quel type de données sous la forme "objet, attribut de l'objet et valeur d'attribut". Il suffit ensuite à l'opérateur de sélectionner la base à utiliser, les champs qui servent de nœuds, les champs qui servent de liens, et ceux qui servent à mettre en évidence des éléments de contextes (Figure 43). Il est aussi possible de faire ces opérations en sélectionnant certaines valeurs de champs. Ce prototype sera disponible avec ses codes et sources pour développer, à moyen terme, l'utilisation de ces graphes pour l'aide à la décision en matière de gestion environnementale et dans le cadre d'un projet de recherche à soumettre aux appels à projet européens (BiodivERsA, ERDF, Seasera, H2020 ...).

Spécifications fonctionnelles du prototype :

Dans le cadre de travaux sur l'interface utilisateurs du prototype IndexMed, un certain nombre de spécifications fonctionnelles ont été définies, certaines développées, d'autres présentées comme souhaitables. Ces spécifications fonctionnelles constituent les rouages permettant la mise en œuvre de nouvelles méthodes et de nouveaux outils prévus lors de la conception de l'architecture du système d'information.

Ces spécifications fonctionnelles prennent en compte l'usager et ses différents niveaux de compétences, et proposent une interface adaptable permettant une découverte de l'approche visuelle par les graphes des jeux de données d'une part, et des fonctions plus avancées d'autre part permettant à un usager expérimenté de faire des requêtes très précises et des constructions complexes. Ce prototype n'est pas abouti, car cela aurait demandé trop de temps et de moyens humains, mais il montre le potentiel de ce type d'outils qui peut être amélioré par le développement des fonctionnalités listées ci-dessous comme souhaitables.

Ce descriptif fonctionnel correspond donc à un état des lieux sur le prototype, et contient des fonctionnalités à améliorer/faire évoluer. Certaines d'entre elles, se complexifiant au cours du développement sont notées « (en cours) », d'autres, n'ayant pas encore débutées, sont notées « (à prévoir) ». Le test itératif de chaque fonctionnalité auprès des différents usagers (avec des niveaux de compétences et de pratiques différents) sera nécessaire pour aboutir à une interface intuitive, pratique et pédagogique.
Ces spécifications fonctionnelles sont exportables et conjugables avec n'importe quel autre projet d'interface sur la fouille de données et peuvent être décrites en faisant abstraction du langage ou de l'environnement de développement.

Ce prototype a une fonctionnalité d'import de données environnementales permettant :

- D'explorer des flux de données XML, JSON et CSV normés et référencés par le prototype. Deux préalables sont nécessaires : des services WEB produisant des flux doivent être mis en place sur les bases de données distantes des partenaires, et ceux-ci doivent être volontairement référencés grâce à leur dictionnaire de données sur la plateforme d'indexation. Sont à développer : un parseur au format le plus générique possible, et des tests de lectures ayant été réalisés sur des modèles de flux simples pour le moment,
- D'indexer toutes les données disponibles dans ces flux et de leur attribuer les éléments nécessaires à leur traçabilité (en cours),
- De produire un identifiant unique de donnée pour chaque enregistrement, « opaque » et basé sur les DOI et un système « d'autorités déclarées » chez un « enregistreur d'autorités» (en cours),
- D'associer à cet identifiant unique d'enregistrement toutes les métadonnées disponibles et si disponible, une url pérenne (en cours),
- De rechercher les identifiants uniques de donnée de versions précédentes ou formats différents pour chaque enregistrement, ou des données sources « parentes » et de conserver ces relations à des fins de résolution (trouver les données/jeux de données les plus récents sur un « objet environnemental », de construction de graphiques du cycle de vie de cette donnée et de recherche de tous les auteurs impliqués dans sa production/transformation (à prévoir dans une version déployée à large échelle).

Ce prototype a une fonctionnalité de sélection des données environnementales à partir des flux distants mis en place par les producteurs de données permettant :

- De sélectionner / désélectionner parmi ces objets, types d'objets, descripteurs / attributs, types de descripteurs/attributs et valeurs de ces descripteurs/attributs ceux qui constitueront les nœuds du graphe. Les nœuds peuvent être constitués d'un ou

plusieurs de ces éléments combinés, et leur importance relative peut être pondérée. (Développement réalisé),

- De sélectionner des objets et leurs descripteurs au sein de ces flux, en fonction des types d'objets, des types de descripteurs (aussi appelés attributs), des normes et standards qu'ils respectent et des valeurs que peuvent prendre chacun de ces descripteurs, et ceci de manière générique d'une « discipline » ou « thématique » d'une base de données à une autre (réalisé, à prévoir une augmentation de l'ergonomie, répondant aux qualités pédagogiques nécessaires pour l'interface),
- De préciser le nombre maximal d'enregistrements à prendre en considération dans chacun de ces flux afin de manipuler une représentation graphique exploitable de graphe. Des alertes sont disponibles lorsque la sélection est inadéquate ou trop importante (à prévoir).
- De créer manuellement des correspondances entre les types ou les valeurs prises par ces attributs/descripteurs, afin de travailler sur plusieurs flux (à prévoir).

Ce prototype a une fonctionnalité de visualisation/manipulation des graphes produits à partir de bases distribuées sur le client permettant :

- d'afficher les valeurs / types / noms / origines des nœuds et / ou liens et de gérer la mise en forme de ces valeurs (taille et couleur) (réalisé),
- de colorer / changer l'aspect (forme/épaisseur) d'une ou plusieurs sélections de nœuds/liens parmi ces objets, types d'objets, descripteurs / attributs, types de descripteurs/attributs et valeurs de ces descripteurs/attributs, ceux qui constituent les nœuds et liens du graphe. Les aspects différents peuvent concerner un ou plusieurs de ces éléments combinés, et leur importance relative peut être pondérée (atténuation, motifs etc.). (Réalisé, à prévoir une augmentation de l'ergonomie, répondant aux qualités pédagogiques nécessaires pour l'interface),
- de supprimer manuellement l'affichage de certains nœuds par un "clic" dans le graphique, ces suppressions ne changent pas la forme du graphe généré. (En cours),
- d'afficher le nombre d'enregistrements disponibles pour chaque combinaison d'objets, types d'objets, descripteurs / attributs, types de descripteurs / attributs et valeurs de ces descripteurs/attributs utilisés en tant qu'objet ou lien pour construire le graphe. (En cours),
- de mémoriser les requêtes et les mises en correspondances issues des manipulations précédentes (sous la forme d'un journal et sous la forme de mise en favoris) et d'y associer des notes, concernant la/les question(s) scientifique(s) prospectée(s), les perspectives que cela donne et les verrous, et enfin d'ordonner les questions / perspectives/verrous en fonction d'un système de notation sous forme d'étoiles. (En cours),

- de générer des graphes en sélectionnant les types de graphe et les propriétés afférentes (fonctionnalité à développer dans le cadre de futurs appels à projet),
- d'afficher des suggestions de combinaison en fonction des données disponibles et non utilisées. (Fonctionnalité à développer dans le cadre de futurs appels à projet),
- de paramétrer ces graphes *via* une représentation graphique pour fouiller et visualiser ces données pluridisciplinaires en mettant sur le même plan des données de type socio-écologique, économiques, écologique, moléculaire et fonctionnelle (relations trophiques, traits fonctionnels…). (fonctionnalité à développer dans le cadre de futurs appels à projets).

Ce prototype a une fonctionnalité de visualisation/manipulation des graphes produits à partir de bases distribuées sur le serveur permettant :

- La génération d'une visualisation statique des graphes paramétrés avec un grand nombre de données (limitation par le serveur) (en cours)
- Des alertes concernant les incohérences sur les données, les manipulations interdites ou les erreurs générées (en cours et à prévoir pour les parties « statistiques » et « qualité des données »)

Ce prototype a une fonctionnalité d'aide en ligne accessible :

- sous forme de tutoriels pas à pas accessibles sur un onglet (En cours)
- à côté de l'ensemble des boutons fonctionnels en lien direct vers le paragraphe les concernant dans le tutoriel. (En cours)

En perspective :

D'autres fonctionnalités sont en cours de définition mais leur maturation est nécessaire, grâce aux tests utilisateurs prévus avec les participants au projet et lors des futurs ateliers. La phase de test n'a pas encore commencé.

Ce prototype aura une fonctionnalité d'export permettant :

- de sauvegarder les données du graphe généré, soit sous la forme d'un visuel, soit sous la forme flux XML ou J.SON
- de créer un service WEB sous forme de flux XML des graphes
- de permettre la bancarisation des qualificatifs et des équivalences entre jeux de données

Un certain nombre d'outils sont envisagés notamment

- Des bibliothèques de normes
- Des bibliothèques de représentations graphiques

- Des bibliothèques de test statistiques sur la sélection de données affichées par le graphe
- L'utilisation d'ontologies téléchargées

Afin de permettre un passage à l'échelle, un travail est nécessaire sur le format et la qualité des données. Afin de mettre en lien les objets de ces bases de données de natures différentes, un travail sur la sémantique de ces données doit impérativement être développé et une recherche de moyen est en cours pour lui donner toute l'envergure nécessaire.

3.3. Les graphes générés par le prototype d'IndexMed

L'utilisation de modèle de visualisation de données basé sur des graphes valués a été mis en place sur les métadonnées et les ensembles de données de CIGESMED (David *et al.*, 2016a). Il permet de les considérer et de les visualiser, malgré leurs différences, à un niveau similaire et améliore la perception et la compréhension des contextes nécessaires à l'aide à la décision en utilisant les nouvelles méthodes d'exploration de données (clustering collaboratif, Représentation des connaissances, etc.).

Nous l'avons vu précédemment, les possibilités de sélection des données et de structuration des graphes sont immenses. On peut choisir en nœuds n'importe quel des champs de la base de données, le nombre d'enregistrement sera le nombre de nœuds, sauf si l'on choisit de faire une sélection (ce que l'on fait pour les très grandes bases pour ne pas trop ralentir la visualisation). L'affichage des contextes comporte lui aussi de nombreuses possibilités : il est possible d'afficher plusieurs valeurs de qualifications de données en leur donnant une couleur et/ou une taille différente ou d'afficher leur valeur sur les nœuds et/ou les liens[112]. A titre d'exemple, dans la figure 44, nous avons affiché des quadrats photo en tant que nœuds, initialement tous verts, reliés par les taxons qu'ils contiennent en commun (les liens). Les quadrats photo ont été sélectionnés pour avoir des orientations tranchées (seules les quatre modalités Nord, Sud, Est et Ouest sont représentées). Les initiales en trois lettres sont les noms des sites. Puis nous avons sélectionné une orientation, (les nœuds correspondant aux quadrats photo pris avec une exposition à l'Est se colorent en violet.). Ce premier niveau de visualisation montre que des quadrats photo de CCA (Cassidaigne - Ile de Cassidaigne) forment un cluster assez homogène, et sont pratiquement tous pris à l'Est. On remarque à gauche de la figure 44 deux photos-quadrats FTF (Frioul - Tiboulen du Frioul) reliés et isolé

[112] Lors de ces sélections successives, si l'opérateur choisit de colorer un noeud avec une valeur d'un attribut, puis un autre noeud avec une valeur d'un autre attribut, certains noeuds vont changer deux fois de couleur. Afin d'éviter cela, lorsqu'il ne s'agit pas de valeurs du même attribut, il faut choisir un autre mode de représentation du nouveau paramètre de contexte (par exemple, "couleur", puis "taille", puis "étiquettes des noeuds").

du reste du graphe, et un groupe de 8 photos-quadrats dont 7 de RMO (Rioux - Moyade) et un de FTF.

Dans la figure 45, nous avons modifié le graphe de la figure 44 en ajoutant aussi les expositions Nord en rouge, et diminué la taille des noeuds des quadrats photo situés à la profondeur D1 (dans la zone des 30 mètres de profondeur). Dans cette seconde représentation, on peut voir que des quadrats photo de FTF sont majoritairement à la profondeur D1, et que beaucoup des photos-quadrats sont orientés au nord. Une grande proportion de ces quadrats photo qualifiés avec les valeurs de descripteurs “D1” pour la profondeur et “N” pour l’orientation se regroupent dans des clusters.

Enfin, dans la figure 46, nous avons modifié le graphe de la figure 45 en ajoutant aussi les expositions Sud (S) colorées en orange. On observe en supplément que beaucoup de RMO sont qualifiés par l’orientation S et la profondeur D1. Les quadrats photo pris dans ce contexte à RMO sont répartis par petits groupes dans les différents clusters du graphe, alors que les quadrats photo pris à RIT pour le même profil (l’orientation S et la profondeur D1) sont majoritairement regroupés. Pour La figure 47, nous avons modifié le graphe de la figure 46 en ajoutant aussi les expositions Ouest (O) colorées en rose. On constate que tous les nœuds en vert se sont bien colorés en rose (pour la dernière modalité restante). *A contrario*, dans la figure 48, on visualise une sélection des photos-quadrats par exclusion des autres catégories “Plat” (Flat) et à l’ombre (Ceiling). Il ne reste donc plus normalement que les deux autres, à savoir “Incliné” (Inclined) et “à l’ombre” (Ceiling). En premier lieu, on remarque un fort regroupement des nœuds de chaque catégorie qui montre une forte structuration des données issues des fréquences relatives d’espèces sur les photo-quadrats en fonction de ce paramètre. En deuxième lieu, on peut observer qu’un nœud vert se situe en plein milieu du graphe : il met en évidence une erreur dans les données de contexte (une erreur de syntaxe sur une modalité de l’inclinaison qui crée artificiellement une cinquième modalité).

La figure 49 (A, B et C) montre qu’avec quelques paramètres, on peut faire des représentations plus complexes qui révèlent la structure de l’échantillonnage et la répartition des données en fonction de contextes et des méthodes employées. Dans cet exemple (Figure 49A), on constate que les photo quadrats prises en linéaire ne semblent pas se regrouper dans un cluster en particulier. En revanche, ils sont absents des clusters ou les nœuds représentent des photo quadrats ou l’inclinaison est “inclinée” (en orange sur la figure 49B et c’est normal vu que les quadrats photo ont été effectués sur des parois verticales). Ils sont concentrés sur les profils verticaux réalisés à la profondeur que D1 (partie basse de la figure), et clairement regroupés dans les clusters du bas de la figure. Afin de les mettre en exergue dans la figure 49C, tous les nœuds correspondant à des photo quadrats faits sur le site FTF ont été grossis. (On n’affiche donc plus l’information liée à la profondeur pour les photo quadrats faits sur les sites FTF, sauf pour les photo quadrats faits sur un transect

linéaire et avec flash car ils étaient uniquement faits à des profondeurs D1). Les résultats des analyses de quadrats photo ne sont pas sensibles à l'utilisation du flash par rapport au phare à la profondeur D1 sur le site FTF. En haut de la figure 49C, on constate que certains quadrats photo faits sur les sites FTF sur des profils inclinés avec des transects en patchs sont situés dans des clusters presque exclusivement constitués de nœuds provenant d'un quadrat photo fait sur un site en particulier (sur paroi inclinée à CCA ou sur paroi verticale à RMO : les nœuds étiquetés FTF sont en orange pour CCA et en vert pour RMO). Un gros nœud rouge (donc correspondant à un photo quadrat fait sur une paroi verticale à FTF) se trouve même au milieu du plus important cluster de nœuds correspondant à des photo quadrats faits à CCA.

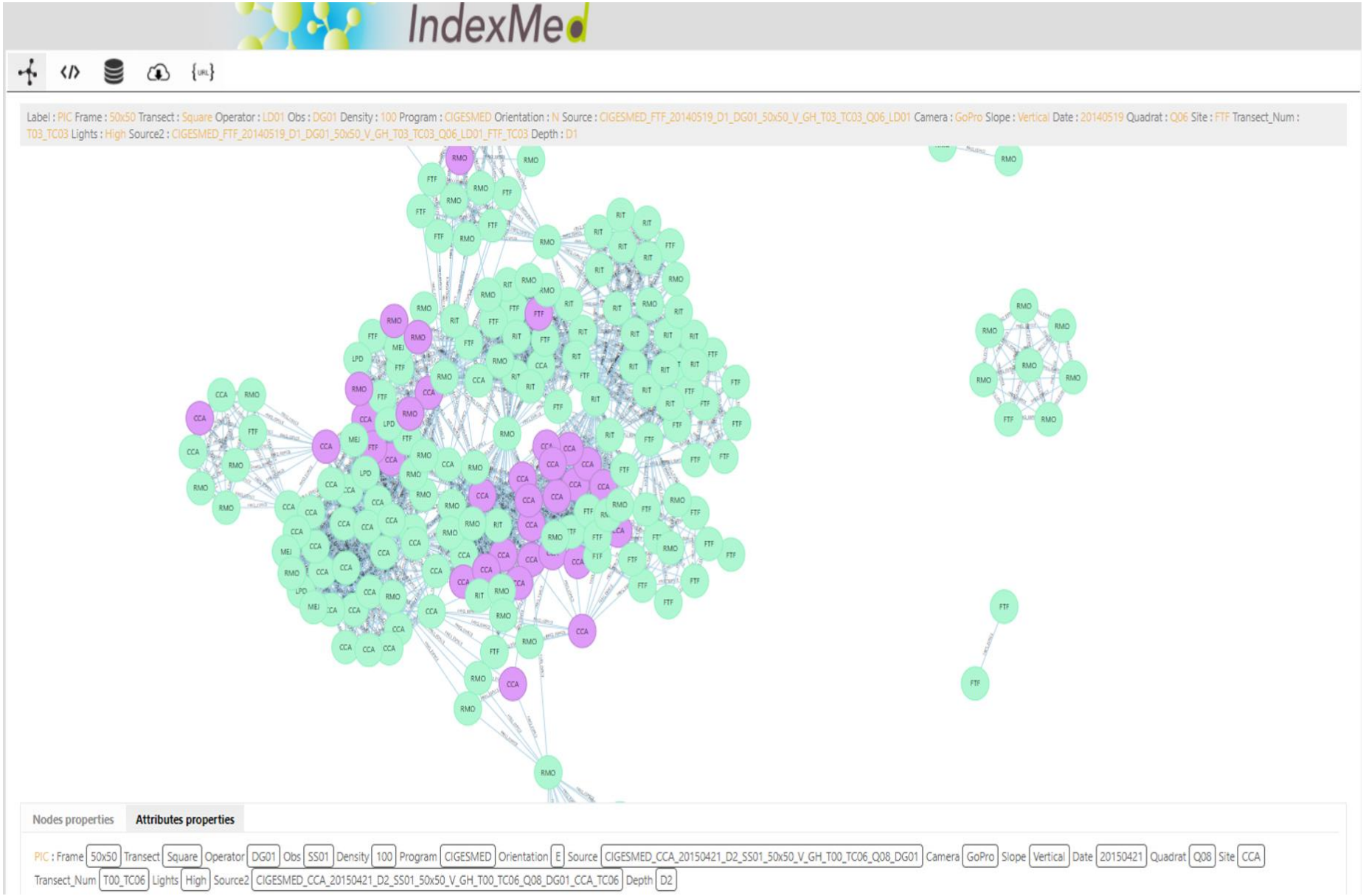

IndexMed
Label : PIC Frame : 50x50 Transect : Square Operator : LD01 Obs : DG01 Density : 100 Program : CIGESMED Orientation : N Source : CIGESMED_FTF_20140519_D1_DG01_50x50_V_GH_T03_TC03_Q06_LD01 Camera : GoPro Slope : Vertical Date : 20140519 Quadrat : Q06 Site : FTF Transect_Num : T03_TC03 Lights : High Source2 : CIGESMED_FTF_20140519_D1_DG01_50x50_V_GH_T03_TC03_Q06_LD01_FTF_TC03 Depth : D1
Nodes properties
Attributes properties
PIC : Frame 50x50 Transect Square Operator DG01 Obs SS01 Density 100 Program CIGESMED Orientation E Source CIGESMED_CCA_20150421_D2_SS01_50x50_V_GH_T00_TC06_Q08_DG01 Camera GoPro Slope Vertical Date 20150421 Quadrat Q08 Site CCA
Transect_Num T00_TC06 Lights High Source2 CIGESMED_CCA_20150421_D2_SS01_50x50_V_GH_T00_TC06_Q08_DG01_CCA_TC06 Depth D2

Figure 44 : Représentation des photos-quadrats (les nœuds du graphe) prises sur différents sites qui sont reliés par des liens représentant des occurrences de certains taxons (ce qui équivaut à un graphe valué). Les sites sont désignés par des abréviations de trois lettres, dans les nœuds, les taxons sont affichés sur les liens (on peut aussi choisir de ne pas les afficher), les photos-quadrats les plus proches dans le graphe sont ceux pour qui les distributions de fréquences d'espèces sont les plus proches. En haut de la figure, on voit les différentes valeurs des attributs d'un nœud sur lequel on a cliqué. En bas, il est possible de changer les étiquettes des nœuds et des liens par une valeur d'un attribut, ou de choisir les valeurs d'attributs pour lesquels on souhaite une coloration ou une taille différente. Ici, les nœuds colorés en violet sont à l'Est.

Figure 45 : Représentation des photos-quadrats présentés en figure 44, sur lesquels on a ajouté deux visualisations de contextes : les nœuds sont colorés en rouge lorsqu'ils sont au Nord, et les nœuds sont plus petits lorsqu'ils sont à la profondeur D1.

Figure 46 : Représentation des photos-quadrats présentés en figure 45, sur lesquels on a ajouté une visualisation de contexte : les nœuds sont colorés en orange lorsqu'ils sont au Sud.

Figure 47 : Représentation des photos-quadrats présentés en figure 46, sur lesquels on a ajouté une visualisation de contexte : les nœuds sont colorés en rose lorsqu'ils sont à l'ouest.

(Ces trois figures 45, 46,47 sont en pleines pages et se trouvent sur les trois pages suivantes).

IndexMed
Label : PIC Frame : 50x50 Transect : Square Operator : LD01 Obs : DG01 Density : 100 Program : CIGESMED Orientation : N Source : CIGESMED_FTF_20140519_D1_DG01_50x50_V_GH_T03_TC02_Q06_LD01 Camera : GoPro Slope : Vertical Date : 20140519 Quadrat : Q06 Site : FTF Transect_Num : T03_TC02 Lights : High Source2 : CIGESMED_FTF_20140519_D1_DG01_50x50_V_GH_T03_TC02_Q06_LD01_FTF_TC02 Depth : D1
Nodes properties
Attributes properties
PIC : Frame 50x50 Transect Square Operator LD01 Obs DG01 Density 100 Program CIGESMED Orientation N Source CIGESMED_FTF_20140519_D1_DG01_50x50_V_GH_T03_TC02_Q06_LD01 Camera GoPro Slope Vertical Date 20140519 Quadrat Q06 Site FTF
Transect_Num T03_TC02 Lights High Source2 CIGESMED_FTF_20140519_D1_DG01_50x50_V_GH_T03_TC02_Q06_LD01_FTF_TC02 Depth D1

Figure 45

IndexMed
Label : PIC Frame : 50x50 Transect : Square Operator : DG01 Obs : SS01 Density : 100 Program : CIGESMED Orientation : E Source : CIGESMED_CCA_20150421_D2_SS01_50x50_V_GH_T00_TC05_Q06_DG01 Camera : GoPro Slope : Vertical Date : 20150421 Quadrat : Q06 Site : CCA Transect_Num : T00_TC05 Lights : High Source2 : CIGESMED_CCA_20150421_D2_SS01_50x50_V_GH_T00_TC05_Q06_DG01_CCA_TC05 Depth : D2
Nodes properties
Attributes properties
PIC : Frame 50x50 Transect Square Operator DG01 Obs RD01 Density 100 Program CIGESMED Orientation S Source CIGESMED_RIT_20140416_D0_RD01_50x50_V_GH_T00_TC01_Q03_DG01 Camera GoPro Slope Vertical Date 20140416 Quadrat Q03 Site RIT
Transect_Num T00_TC01 Lights High Source2 CIGESMED_RIT_20140416_D0_RD01_50x50_V_GH_T00_TC01_Q03_DG01_RIT_TC01 Depth D0
S Colors : Size :

Figure 46

Label : PIC Frame : 50x50 Transect : Square Operator : DG01 Obs : SS01 Density : 100 Program : CIGESMED Orientation : W Source : CIGESMED_FTF_20150416_D2_SS01_50x50_I_GH_T00_TC09_Q06_DG01 Camera : GoPro Slope : Sloping Date : 20150416 Quadrat : Q06 Site : FTF Transect_I
T00_TC09 Lights : High Source2 : CIGESMED_FTF_20150416_D2_SS01_50x50_I_GH_T00_TC09_Q06_DG01_FTF_TC09 Depth : D2
Nodes properties
Attributes properties
PIC : Frame 50x50 Transect Square Operator DG01 Obs SS01 Density 100 Program CIGESMED Orientation W Source CIGESMED_FTF_20150416_D2_SS01_50x50_I_GH_T00_TC09_Q06_DG01 Camera GoPro Slope Sloping Date 20150416 Quadrat Q06 Site FTF
Transect_Num T00_TC09 Lights High Source2 CIGESMED_FTF_20150416_D2_SS01_50x50_I_GH_T00_TC09_Q06_DG01_FTF_TC09 Depth D2
W Colors :
Size :

Figure 47

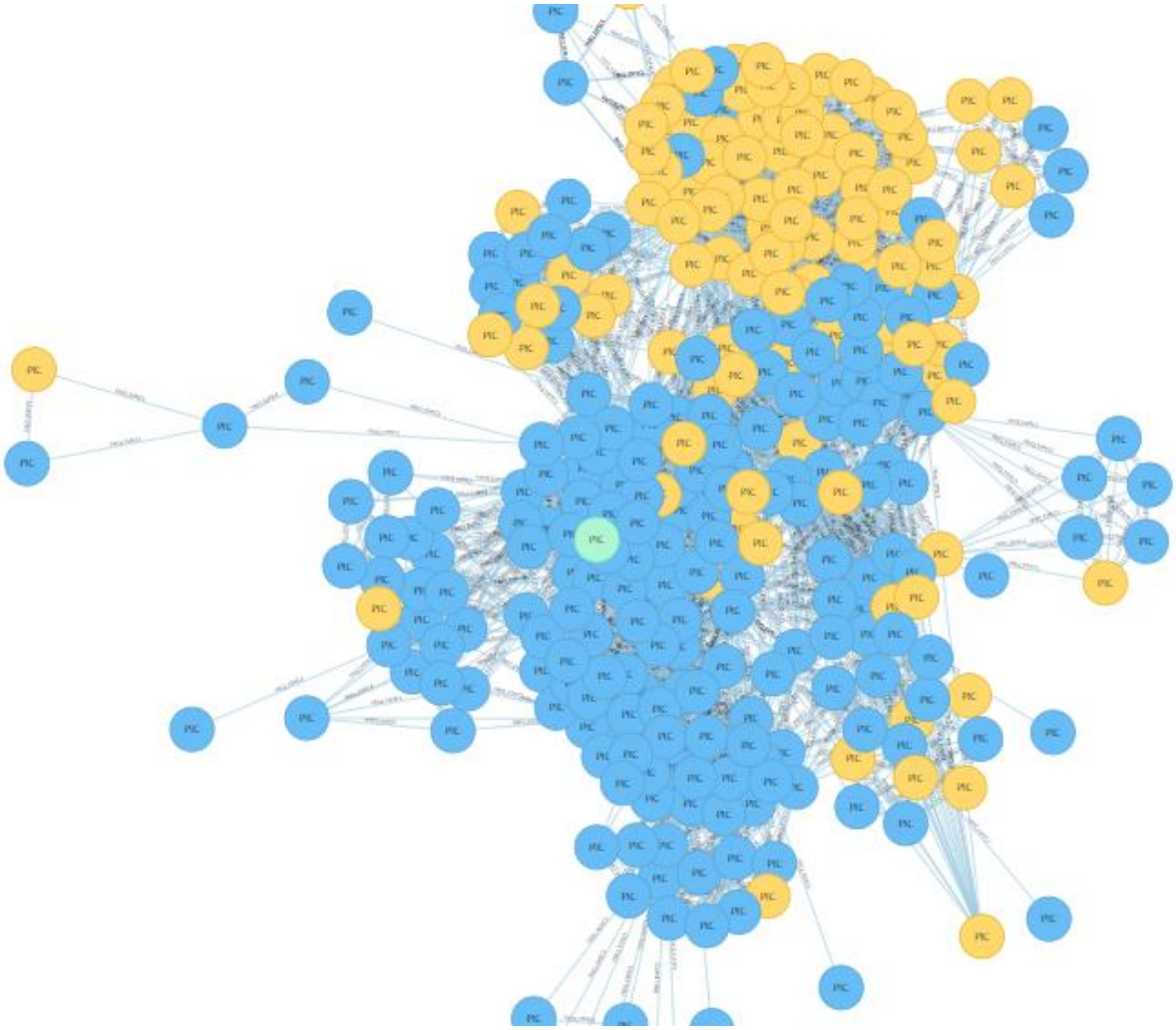


Figure 48 : Représentation d'une sélection des photos-quadrats par exclusion des autres catégories (il reste donc normalement les nœuds de quadrats photo pris sur une inclinaison intermédiaire et vertical) sur lesquels on a ajouté une visualisation de contexte : les nœuds sont colorés en bleu lorsque les photos-quadrats sont pris sur une paroi verticale et en orange lorsque les photos quadrats sont faits sur une inclinaison de paroi intermédiaire. Il subsiste alors au milieu un nœud vert qui est une erreur dans les données (une erreur de syntaxe sur une modalité de l'inclinaison qui crée artificiellement une cinquième modalité)

Dans la figure 49, en procédant de la même manière, nous avons affiché des paramètres de contexte (caractérisant les systèmes observés) et des descripteurs du système d'observation (en 49A, on affiche les photos quadrats faits sur un transect linéaire en rouge, et en vert les photos quadrats réalisés avec la méthode des patchs). En 49B, on a sélectionné les pentes intermédiaires pour les colorer en orange et sélectionné l'utilisation du flash sur les pentes verticales pour les colorer en bleu (les transects linéaires ont été uniquement faits sur des

profils à inclinaison verticale et les tests de comparaison entre phare et flash ont été uniquement faits sur des profils à inclinaison verticale sur les transects linéaires. Les nœuds représentant des quadrats photo faits à la profondeurs D1 sont les plus petits. En 49C, Les NOEUDS ayant la valeur FTF sont grossis pour mettre en évidence le site sur lequel ont eu lieu ces inter-calibration (tous les nœuds représentant des quadrats photo faits sur les sites FTF à la profondeur D1 sont donc repassés en très gros).

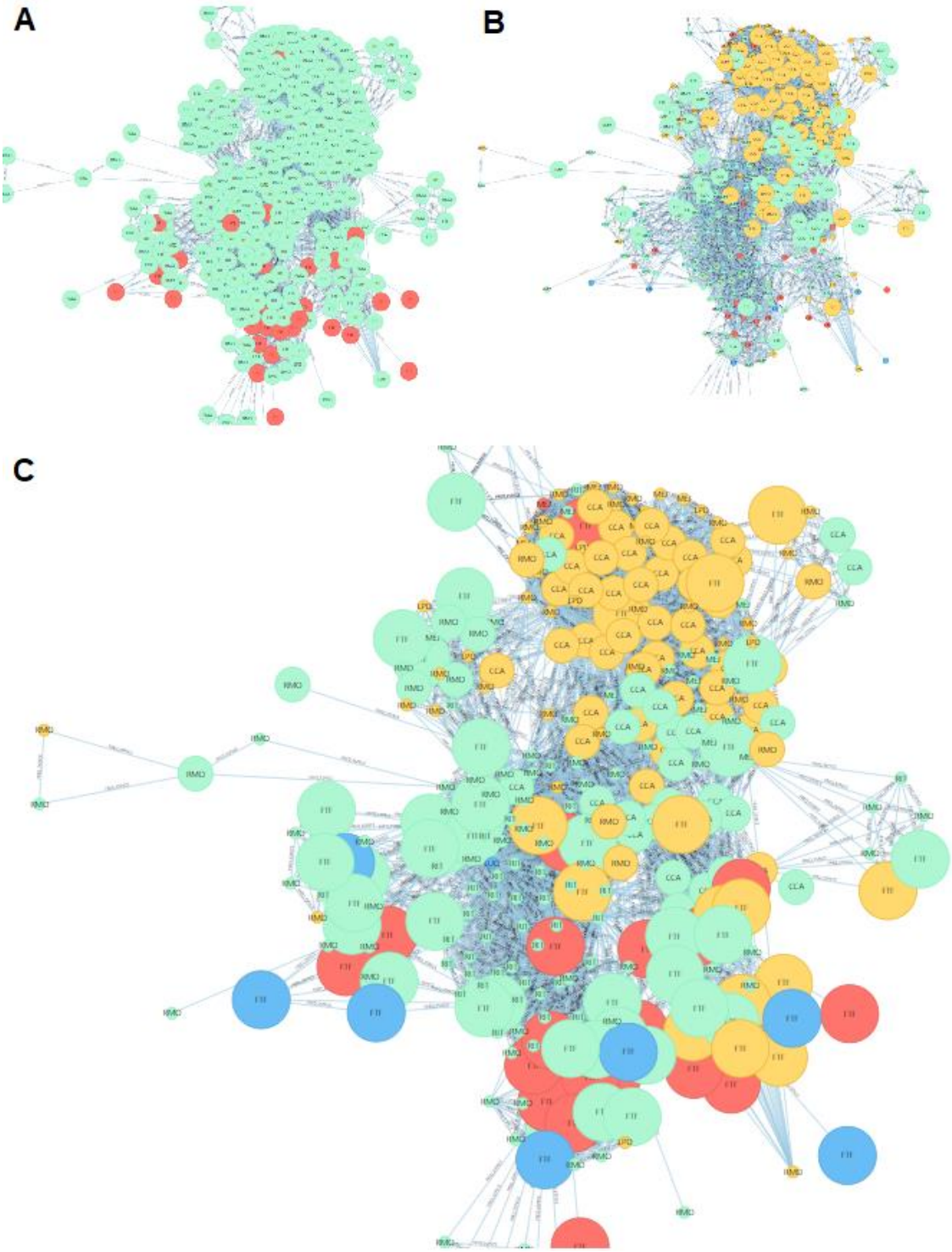


Figure 49 (A, B et C) : Représentation d'une sélection des photos-quadrats avec sélections successives de paramètres correspondant aux méthodes de mesure et au contexte des mesures. Dans la figure A, les photos quadrats faits sur un transect linéaire sont en rouge, et les photos quadrats réalisés avec la méthode des patchs sont en vert. Puis on a sélectionné les pentes intermédiaires pour les colorer en orange et sélectionné l'utilisation du flash sur les pentes verticales pour les colorer en bleu. (B)
Il est possible de faire des animations sous forme de GIF animés pour détailler chaque étape de sélection de paramètres.

3.4. Atelier sur la qualification et la curation des données : intérêts, méthodes, difficultés.

A cause de leur hétérogénéité, les données sont souvent difficiles à agréger pour avoir une vue d'ensemble (on parle parfois d'empilement de bases de données). Améliorer l'interopérabilité nécessite de s'appuyer sur la curation de ces données avec une sémantique commune. Cette sémantique commune demande d'organiser les échanges entre les différents spécialistes, tout en tenant compte des préconisations des spécialistes de la donnée et de la visualisation sous forme de graphes. Il faut de plus documenter les méthodes déterminées et les choix faits lors des discussions.

Pour curer un jeu de données, il faut aller plus loin que le "nettoyage" de ses imperfections. La curation de données, du latin *curare* qui signifie "prendre soin", est essentielle avant tout processus d'analyse ; elle consiste à améliorer la capacité des données à décrire un système de manière univoque et explicite. Elle est essentielle pour préparer un jeu de données pour chaque nouveau type d'analyse ou agréger différents jeux de données d'origines, de structures et de formats différents. La curation permet notamment de construire des graphes, qui sont pertinents pour modéliser plus efficacement les facteurs "composantes" des interactions biologiques malgré leur hétérogénéité.
Dans l'optique de créer des graphes, il faut expliquer comment choisir dans la base de données les "objets" (c'est-à-dire le ou les champs de la base de données) qui vont servir de "nœuds"/"sommets" aussi appelé "entités", et les champs descripteurs (que l'on appelle attributs) des objets choisis pour en faire des liens/arêtes aussi appelés "relations". Toute qualification de la donnée ou requalification de la donnée peut être utilisée comme nouvel objet ou comme attribut d'un objet existant, à condition de suivre un processus d'adaptation.
Un processus de nettoyage des erreurs a été mis en place. A ce titre, une première version

d’un document a été construit pour organiser sous forme d’étapes une base de consignes (voir “*how to do”,* partie 3.6 de ce chapitre). Ensuite, un processus de requalification des données a été mis en place. Les principales transformations effectuées sur les données sont de deux ordres : celles qui permettent de découvrir la donnée (et ses qualités et potentiels) et celles qui permettent de les améliorer. Chaque participant a dû, à partir de l’étape de découverte du potentiel de ses données, faire des choix concernant les entités et les types de relations et de graphes qu’ils voulaient construire, en fonction de la qualité initiale de leurs jeux de données et du temps imparti. Les premiers essais de curation de chaque participant ont montré que cet atelier n’était qu’une première étape dans le processus d’amélioration et de restructuration de leurs données. L’utilisation de démonstrations s’appuyant sur le prototype d’IndexMed a néanmoins permis de les convaincre de l’intérêt de persévérer dans cette démarche. (Exemple de démonstration avec la figure 50 (issue de David *et al.*, 2017) avec la visualisation de l’empilement de bases de données d’ArkeoGIS et des descripteurs communs aux différentes bases de données).

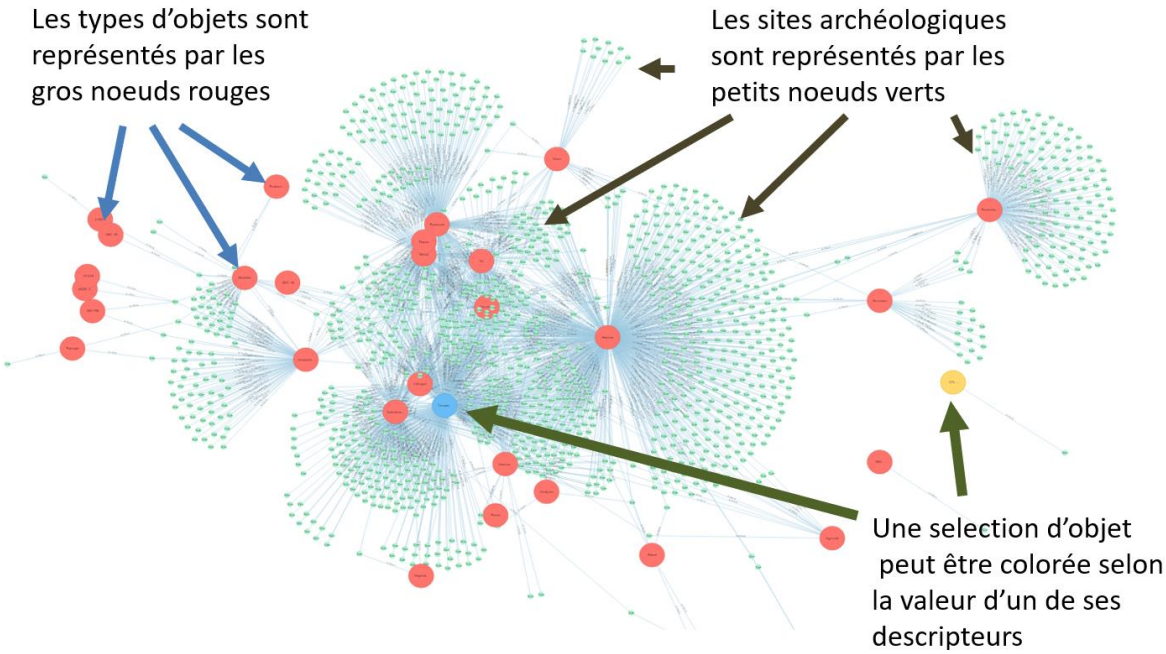


Figure 50 : Exemple de graphe démonstratif de type bipartite (c'est à dire mélangeant deux types de nœuds) issu de l'interface IndexMEED, utilisant les données exportées d'une sélection géographique sur ArkeoGIS , représentant 1492 sites archéologiques et 4 950 liens matérialisant les types d'objets trouvés sur ces sites. Les sites sont représentés par les petits nœuds verts, les clusters rapprochant les sites ayant les mêmes topologies / patrons de descripteurs (ici sont utilisés uniquement les patrons d'objets de niveau 1) matérialisés par des nœuds rouges. Une sélection a été faite sur deux valeurs de descripteurs : "Bleu" pour les types d'objet de niveau 1 qualifiés de "céramique" (le site contient donc au moins un objet de type céramique s'il est lié à ce nœud), "Jaune" : le nœud correspond à un objet daté de la période exactement égale à "-900 -à -726". Cet exemple simple a permis de montrer aux participants de l'atelier comment utiliser un graphe pour visualiser la répartition des valeurs de descripteurs dans un jeu de bases de données empilées.

3.5. Ateliers visualisation et résultats préliminaires concernant la visualisation des clusters de photos de plaques d'ARMS :

Premiers résultats de visualisation des clusters de photos de plaques d'ARMS

Le résultat attendu lors de l'atelier concernant le clustering consistait à visualiser avec le logiciel Tulip les plaques et les regroupements de plaques d'ARMS reliées à partir des fréquences relatives de taxons recueillis sur des faces des plaques réparties à différents

endroits du récif (et donc exposés différemment aux facteurs de contextes comme l'exposition à la lumière, au courant, ou aux différentes pressions décrites soit par des capteurs, soit en s'appuyant sur des avis d'expert. Figure 51)

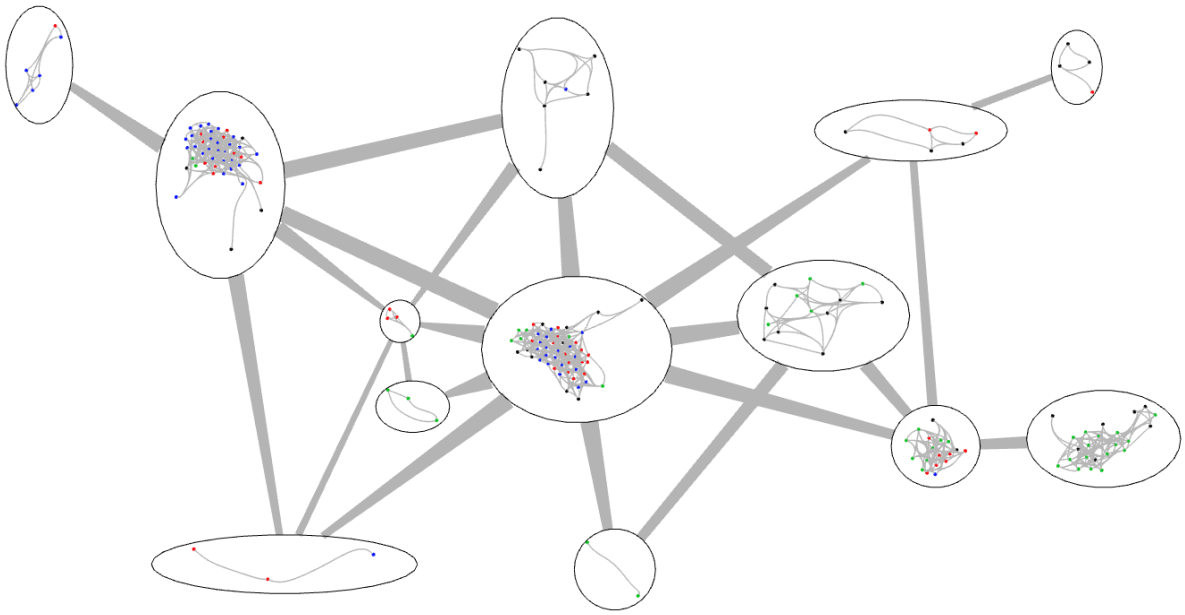


Figure 51 : Dans cette visualisation, qui est une démonstration préliminaire réalisée par Romain Bourqui avec le logiciel Tulip, les fréquences relatives ont été déterminées à partir de photos. 6 facettes de plaquettes ont été analysées dans 3 récifs, pour 3 sites pour chacune des 4 mers régionales (Golfe de Gascogne, méditerranée nord occidentale, Adriatique et Mer Rouge). Les nœuds de ce graphe représentent des facettes de plaques en PVC, et les liens relient les facettes pour chaque pourcentage d'un taxon en commun avec une autre facette. Chaque couleur représente une mer régionale différente. L'intérêt de cette visualisation est de mettre en exergue les clusters les plus importants, où les plaques avec les compositions en taxons les plus similaires sont les plus proches, et les sites des différentes mers sont colorés de manière différente avec des regroupements plus ou moins hétérogènes. La figure a été construite en sélectionnant uniquement des fréquences relatives en commun supérieures à 10%. Lorsque les formes, tailles et/ou couleurs des nœuds correspondant à l'affichage d'une valeur de paramètre de contextes se regroupent dans un cluster, il faut tester la significativité de ces corrélations grâce à des méthodes statistiques propres aux graphes.

Prochaines étapes concernant le clustering et la visualisation des plaques ARMS

Les combinaisons de facteurs de contextes peuvent être associées à chaque site (soit des combinaisons de descriptions de pression sous forme de booléen comme présence absence, soit plus finement avec des catégories ordonnées ou non ordonnées). Lorsque des combinaisons de facteurs de contextes se regroupent dans un cluster (une même gamme de température avec une même gamme de courant, lumière, et/ou facteurs de pression), on peut considérer que la structure de communauté de ce cluster correspond à la combinaison en question. Les premiers tests montrent qu'un travail sur la sémantique concernant les descriptions de contexte est incontournable, surtout à large échelle, pour améliorer le potentiel de l'analyse et les qualités du suivi comme sa comparabilité dans le temps. Les prochaines étapes auraient pour objectif de les normaliser, puis de "*temporaliser*" les suivis pour analyser les effets saisonniers et les mettre en évidence sous forme de « graphes temporels » (et utiliser les successions biologiques comme des changements dans des fréquences comme indicateurs de changement d'état (avec une cause ou une conjonction de causes naturelles, et/ou anthropiques). Lors du recueil des données, des analyses de communautés faites via des analyses métagénomiques ont été effectuées. Ces représentations pourraient inclure les analyses métagénomiques avec le même type d'approche (des clusters de gènes en tant que liens entre plaques). Il est aussi possible de "*temporaliser*" le plan d'échantillonnage pour la génomique surtout lorsque celui-ci devient un dispositif permanent avec des objectifs de gestion (On observe par sa/ses transformation(s) que l'arbre perd alors des éléments au fur et à mesure de la survenue des anomalies, tout en en affichant la cause, si celle-ci est connue).

3.6. "How to do"

L'énonciation de règles et de principes à respecter est un préalable au développement de comportements vertueux, notamment concernant les méthodes et bonnes pratiques de valorisation des données, de leur partage et de leur visualisation. La publication de ces règles et la mise en place de sanctions pour ceux qui les ignorent et de gratifications pour ceux qui les appliquent le mieux ne sont néanmoins pas une condition suffisante pour leur application par un plus grand nombre d'acteurs. Suivre ces règles devient vite complexe et un acteur volontaire pour les appliquer est vite confronté à des questions de mise en oeuvre (quels sont les aspects prioritaires de la curation, comment dimensionner les moyens nécessaires, quels objectifs sont réalistes et raisonnables en fonction de mes moyens, etc.). Afin de faciliter la mise en oeuvre de ces règles, nous avons proposé aux participants des ateliers sur la curation de données de construire un document "*how to do*" sur la préparation des

données en vue de construire des graphes regroupant et expliquant toutes les étapes nécessaires à cet objectif et tous les verrous à lever (document en version 1.0 en Annexe 9.2). Ce document a pour but de mieux quantifier les moyens nécessaires en se fixant des objectifs plus détaillés, et en préparant toutes les actions nécessaires avant d'en mettre une en oeuvre dans un plan de curation. Il a été enrichi par chaque participant par toutes les remarques jugées utiles pour une compréhension par le plus grand nombre (c'est à dire des non spécialistes des données). Il liste notamment pour les phases de préparation, d'exécution puis de contrôle tous les problèmes qui ont été rencontrés et comment il a été proposé de les résoudre. Ce "*how to do"* a pour vocation d'être adaptable à n'importe quelle discipline. Il a été conçu pour être accessible aux non spécialistes de la donnée et pourra être amélioré à chaque utilisation sous la forme d'un document type wiki[113] relié au prototype d'IndexMed. Un deuxième document "*how to do"* sur la visualisation de données environnementales sous forme de graphes est en cours de rédaction depuis la réalisation des ateliers sur la visualisation de données.

4. Discussion et perspectives concernant les méthodes de fouille de données environnementales basées sur les graphes

4.1 Intérêts des représentations visuelles

Les premières recherches concernant les communautés d'archéologues (ArkeoGIS) et d'écologues marins (CIGESMED et DEVOTES) ont permis de montrer que des techniques à base de graphes sont adaptées pour modéliser plus efficacement les composantes d'interactions spatiales (les facteurs de contextes) malgré l'hétérogénéité de ce type d'information (David et al, 2017) : l'utilisation de graphes rend possible la prise en compte des données malgré leur disparité et sans les hiérarchiser, et elle permet d'améliorer la précision des outils d'aide à la décision en utilisant des méthodes émergentes de fouille de données.

Les objets ayant le plus de liens en commun sont les plus proches, ceux ayant les liens les plus ténus (c'est-à-dire le moins de chemins possibles pour les relier à entre eux et beaucoup de nœuds intermédiaires) sont les plus éloignés dans la représentation. La représentation sous forme de graphe permet de représenter de nouveaux objets en combinant les valeurs de différents champs. On peut ainsi traiter les champs un à un ou bien en groupe de valeurs/

[113] Un wiki est une application web qui permet la création, la modification et l'illustration collaboratives de pages à l'intérieur d'un site web.

descripteurs : plusieurs descripteurs peuvent être assemblés pour former - selon la combinaison de leurs valeurs respectives - un motif, qui pour ces valeurs données sont appelés patrons (patterns en anglais).
Pour rappel, les champs de la base de données, que l'on considère comme des "contextes" ne participent pas à la topologie du graphe : la significativité des motifs peut ensuite testée par des méthodes comme le clustering de graphes (voir Chapitre 4 partie 1.4).
Cette agrégation de nœuds en classes est un élément-clé pour l'analyse de grands graphes. Une fois les groupes obtenus, on peut appliquer à nouveau l'opération pour obtenir un clustering hiérarchique (basé sur une autre variable par exemple). Cette décomposition hiérarchique (ou multi-échelle) permet de modifier la complexité des algorithmes de fouille, de faciliter l'exploration des données, et de proposer une visualisation paramétrable : on parle aussi de navigation multi-échelle (Lambert et al 2013).

Dans des graphes plus complexes où le nombre de combinaisons et de liens peut croître exponentiellement, l'étude de la corrélation entre fréquence de contextes et "clusters" de nœuds peut demander de paralléliser les calculs nécessaires à une investigation des parcours possibles. Selon la question scientifique sous-jacente aux objets représentés par le graphe, certains éléments dans les liens ou les nœuds peuvent être ignorés ou simplifiés.
Ces graphes sont paramétrables pour fouiller et visualiser ces données pluridisciplinaires en mettant sur le même plan des données de types écologiques, physico-chimiques, fonctionnelles (relations trophiques, traits fonctionnels), et socio-écologiques, économiques... Les questionnements scientifiques possibles concernent l'écologie des systèmes observés : bon état écologique, correspondance de patrons de contextes et de données concernant les abondances relatives d'espèces ou des systèmes d'observation (détection des biais dans la formation des observateurs, expertise partielle dans les jeux de données, définition de la puissance de l'échantillonnage nécessaire, gestion des coûts associés).

4.2 Éléments de discussion concernant les utilisations possibles des graphes générés par le prototype d'IndexMed et lors des ateliers

Les graphes générés par le prototype d'IndexMed présentés dans la partie 3.3 sont des graphes valués : les liens sont pondérés par la fréquence des taxons entre quadrats photo. Par exemple, si un quadrat photo contient 20% d'une espèce, et une autre 15%, le lien aura une pondération de 15, soit le nombre de points communs aux deux quadrats photo. Cela pourrait correspondre à 15 liens. Plus deux photos se ressemblent, plus cette pondération est importante, elle peut théoriquement, pour des quadrats photo identiques en terme de

fréquences respectives de leurs différents taxons, être de 100 %. Les deux nœuds correspondant à ces deux quadrats photo seront les plus proches l'un de l'autre dans le graphe. Pour permettre une visualisation plus pédagogique, une sélection aléatoire est opérée sur les relations nœuds-liens. Ces sélections ont été opérées par étapes successives en préservant les clusters évidents dans les graphes plus surchargés.

Les graphes permettent d'étudier le système d'observation et sa structure, même lorsqu'elle est complexifiée par une grande diversité de contextes : les graphes exposés dans la partie 3.3 montrent en premier lieu la répartition des échantillons en fonction de variables de contexte, et permettent de visualiser celle-ci très rapidement et sur un très grand nombre d'échantillon à la fois.

Les graphes permettent aussi d'étudier les systèmes observés et de visualiser les effets de combinaisons de contextes sur la structure des communautés (dans les exemples issus des programmes CIGESMED et DEVOTES).

Dans la figure 44, on voit différents regroupements de nœuds, et le graphe montre que l'orientation semble influer fortement sur la structure des communautés analysées à l'aide des quadrats photo. On constate notamment que les quadrats photo pris en exposition Est à CCA sont regroupés donc certainement très homogènes. L'interprétation que l'on peut en faire est que dans ce site, éloigné de la côte (CCA est le phare de Cassidaigne), les quadrats pris à l'est sont souvent et notablement plus exposés au courant ligurien que les autres sites, ce qui expliquerait que les compositions en fréquences relatives de taxons des quadrats photo soient plus homogènes et relativement différents des autres sites, tous relativement moins exposés à ce courant. *A contrario*, les quadrats photo pris à RMO sont répartis dans tous le graphe. Cet état de fait peut correspondre à la diversité des différents profils que l'on trouve sur le site en terme d'exposition (qui dépend de l'orientation de la paroie par rapport aux courants d'une part, et des éventuelles protections créées par d'autres amas rocheux ou îles en amont du courant), mais aussi en terme d'inclinaison et de rugosité, facteurs influant en parallèle les impacts du courant et les apports de lumière. Le deuxième cluster de sites à l'est regroupe des quadrats photo de sites différents (RMO, CCA et 1 FTF seulement, mélangé à d'autres expositions des sites RMO, LPD, MEJ, FTF). Le site FTF a peu de quadrats photo pris à l'est, et les quadrats photo de ce cluster peuvent correspondre à des sites moins exposés aux courants (car situés dans des renfoncements dans la paroi rocheuse par exemple). Sur ce graphe, on constate aussi que certains nœuds sont détachés du reste du graphe, notamment les deux quadrats photo de FTF, ce qui peut correspondre à la sélection d'une valeur rare ou une erreur de syntaxe commune dans un descripteur de ces quadrats. Lorsque des quadrats photo se regroupent en étant séparés du reste du graphe (les sept nœuds représentant des quadrats photo de FTF et de RMO), on met en évidence des assemblages d'espèces particuliers typiques (par exemple un recouvrement abondant

et localisé d'un taxon) que l'on peut expliquer par un regroupement de contextes particuliers. En ajoutant aussi expositions Nord en rouge (Figure 45), et en diminuant la taille des nœuds des quadrats photo situés à la profondeur D1 (dans la zone des 30 mètres de profondeur, on voit que des quadrats photo de FTF sont majoritairement à la profondeur D1, au Nord. Cela s'explique par la configuration du site, très homogène en terme d'orientation, et de par le fait que les transects d'inter-calibration ont tous été faits à FTF, et sont donc bien plus nombreux. De fait, ils forment un cluster dense car les populations qui s'y trouvent sont très homogènes en terme de structures (faciès à Paramuricea clavata avec des éponges et beaucoup d'autres cnidaires).

Les graphes aident à déterminer des sélections utiles pour affiner une analyse statistique : Dans la figure 45, on constate que la profondeur est un descripteur très structurant du jeu de données, notamment pour le site FTF pour lequel les clusters sont très marqués, alors qu'il l'est moins pour RMO. Ce constat est expliqué par la grande homogénéité des peuplements coralligènes notamment très densément recouverts de gorgones rouges (*Paramuricea clavata*), alors que la cartographie a montré une grande diversité de profils et de type de peuplements sur RMO : les grandes espèces dressées formant parfois des peuplements denses comme les gorgones rouges ou recouvrantes peuvent couvrir une partie importante des quadrats photo, et si leur population est importante sur le site, chaque quadrat photo contiendra une proportion importante de points attribués à ces espèces. Les nœuds représentant ces quadrats photo forment très facilement des clusters très marqués. D'autre part, lors de l'inter-calibration, nous avons constaté que selon leur état (polypes ouverts ou non) peut faire varier de manière significative les résultats de l'analyse des quadrats photo d'un même transect. Dans ce cas, l'analyse de la structure des communautés son suivi à long terme sera plus efficace en écartant certains taxons.

Les graphes permettent de détecter des éléments particuliers dans les jeux de données : dans le cas du regroupement de sept nœuds dans la figure 44, on peut voir que chacun des sept nœuds est relié à tous les autres (on appelle cela un graphe complet). Cet élément topologique est certainement généré par la sélection opérée sur les entités relations pour la visualisation, mais correspond de toute manière à un cluster lorsque l'on ne fait pas cette sélection. Le fait de considérer des éléments très particuliers en même temps que le reste du jeu de donnée peut diminuer la sensibilité des tests statistiques permettant de détecter un effet d'une conjonction de valeur de paramètres environnementaux. Dans le cas de clusters très évidents, et très détachés du reste, il peut donc être judicieux de considérer les deux parties du graphe séparément pour les analyses statistiques testant les effets des différents descripteurs de contexte.

Les graphes permettent de détecter des erreurs : dans la figure 48, la sélection des photos-quadrats par exclusion des autres catégories a normalement sélectionné uniquement les nœuds de quadrats photo pris sur une inclinaison intermédiaire et verticale : les nœuds sont colorés en bleu lorsque les photos-quadrats sont effectués sur une paroi verticale et en orange lorsque les photos quadrats sont faits sur une inclinaison de paroi intermédiaire. Il subsiste alors au milieu un nœud vert qui est une erreur dans les données (ici simplement une erreur de syntaxe), et qui peut facilement être retrouvée en cliquant sur le nœuds. Le numéro de l'enregistrement s'affiche au-dessus du graphe avec toutes ses données attributaires. Nous avons fait la même opération dans la figure 47, en colorant en rose les expositions Ouest (O). Les 4 expositions disposent d'une couleur différente qui n'est pas le vert. Cela signifie qu'il n'y a pas d'erreur de syntaxe dans les données concernant ce descripteur.

Ces graphes montrent que ces visualisations permettent d'analyser et de qualifier la qualité de l'information (les valeurs rares de descripteurs, ou les erreurs de syntaxe par exemple), et peuvent être utilisés pour estimer et prévoir les actions de curation des données. Les graphes peuvent être déconnectés comme les petits éléments de la figure 44, imposant alors le regroupement de deux valeurs d'un descripteur en une seule, ou la modification de valeurs non vraies.

Les graphes permettent de représenter les plans d'échantillonnage. Par exemple, nous avons effectué une représentation du plan d'échantillonnage des plaques des ARMS sous forme de graphe, mettant en évidence des éléments manquants ou des erreurs et oublis (Annexe 9.2).

Les graphes peuvent être animés et utilisés comme supports pédagogiques : ils permettent non seulement de représenter un grand nombre d'objets, mais aussi un grand nombre de contextes qui se superposent, sous une forme assez didactique. La figure 49 montre comment il est possible d'organiser par étapes successives la construction d'un graphe montrant plusieurs paramètres de contextes différents (toujours basés sur le graphe utilisé dans les figures 44 à 47). Ces étapes pourraient même constituer les différentes images d'une animation (par exemple sous forme de gif animé, augmentant ainsi leur potentiel pédagogique). La dernière de ces trois figures montre que l'on peut superposer une valeur de paramètre à une autre lorsque les valeurs du descripteur qu'on superpose remplacent uniquement une valeur du premier paramètre, ce qui permet de ne pas perdre d'information, même si l'interface a un nombre limité de configuration visuelle des noeuds : aucune information ne disparaît. Il faut alors commenter précisément le graphe lorsqu'on l'utilise afin d'éviter des interprétation erronées car la valeur de paramètre de contexte qui aura été partiellement remplacée aura deux représentations visuelles.

Les graphes permettent de comparer des bases de données et d'analyser les descripteurs communs et particuliers différents systèmes d'information : nous avons dans le cadre des premiers essais du prototype montré les éléments communs à différentes bases dans le cas de l'empilement des bases d'ArkeoGIS (David *et al.*, 2017), en visualisant les objets mobiliers et immobiliers communs ou très particuliers de plus de 1400 sites archéologiques (Figure 50 partie 3.4 de ce chapitre).

4.3 Les graphes utilisés par des environnementalistes : la nécessité absolue de travailler en ateliers

La découverte de ces approches puis leur première expérimentation par un néophyte demande des phases de sensibilisation et d'explication répétées pendant lesquelles le chercheur comprendra l'intérêt de ces méthodes pour ses propres recherches, puis une expérimentation sous forme d'ateliers d'abord basés sur des « cas résolus » explicites, puis une transposition sur ses propres problématiques de recherche au travers de la réalisation de premiers exemples sur ses propres données. Avec chaque participant, nous avons alterné les phases de découverte et de curation de données, puis la construction de graphes, ce qui permettait de montrer la complexité de ces prétraitements ainsi que leur intérêt dans le cadre de l'amélioration du « potentiel de découverte » de ces jeux de données.

Les graphes peuvent aider à organiser des travaux interdisciplinaires : l'atelier de travail organisé lors des rencontres MaDICS le 23 Juin 2017 à Marseille a permis la réalisation d'un premier « mind-mapping » autour des compétences en écologie et informatique des participants. La réalisation de cette carte de compétences, qui doit être enrichie, a permis de répartir chaque aspect du travail sur les graphes entre les spécialistes et laboratoires de différentes disciplines.

Deux ateliers préparatoires « cas d'études » puis un séminaire centré sur les algorithmes

Les questions scientifiques véritablement novatrices, pour l'instant peu explorées, émergent des données grâce aux graphes (cet état de fait a été une des conclusions majeures lors des deux derniers séminaires IndexMEED). Le consortium priorise les cas d'étude présentant les meilleurs niveaux d'accessibilité et d'utilisabilité des données (par exemple, données ArkeoGIS, GBIF, MNHN, ECOSCOPE ou IMBE), et prépare les outils d'audit de la donnée permettant d'organiser les étapes nécessaires au respect de ces principes FAIR pour les cas d'études. Cet objectif d'amélioration qualitative de jeux de données s'est traduit par l'organisation de deux ateliers qui permettront de faire une démonstration par l'exemple, basée sur plusieurs cas d'études précités. Le premier atelier a porté sur la curation de données, en vue de les intégrer dans des interfaces de

manipulation de graphes. Les ateliers de curation, organisés à Aix-en-Provence, Marseille et à Paris, ont permis d'identifier des verrous actuels et futurs concernant les données et l'organisation des compétences, pour permettre une amélioration significative du référencement, de l'accessibilité, de l'interopérabilité et de la réutilisabilité des données selon les principes FAIR, et ainsi définir les moyens nécessaires au respect de la nouvelle réglementation sur l'accessibilité numérique des données et travaux de la recherche (Loi pour une République numérique, ; Lambert *et al.*, 2013). Le second atelier, organisé mi-novembre à Marseille et à Paris, portait sur la visualisation des jeux de données, améliorés lors de l'atelier précédent.

Le cercle vertueux de la curation et de la visualisation de données promet de les rendre réutilisables et intégrables dans divers modèles, dans un objectif de compréhension et d'analyse des systèmes socio-écologiques.

4.4 Des défis à venir

Les futurs défis concernant le développement d'interfaces accessibles basées sur les graphes sont le passage à l'échelle (100 000 nœuds à 1 000 000 de nœuds ou plus) et l'augmentation des services aux usagers. La prochaine étape consisterait à développer les cas d'utilisation du prototype IndexMEED et à améliorer les fonctionnalités selon les retours utilisateurs. La temporalisation de ces graphes sous forme de graphes évolutifs, devra permettre de générer des scénarios dépendants de contextes mesurés et utilisables dans les domaines de l'aide à la décision.

Le traitement des problématiques de la qualité et de l'interopérabilité des données nécessite encore de prévoir de nombreux échanges entre les différents participants. Les verrous scientifiques à lever dans un projet de recherche interdisciplinaire dans le cadre d'IndexMEED ont été identifiés par les experts du domaine des S.T.I.C., et concernent notamment i) l'augmentation des fréquences et de la densité d'acquisition des observations (liées notamment au développement des méthodes de reconnaissance automatique et au déploiement d'outils d'acquisition moins onéreux), ii) la diversification des objets et des descripteurs d'objet intégrés dans les graphes, iii) la normalisation des descripteurs de la donnée et les méthodes permettant d'intégrer les différents niveaux de qualité des données. Une fois les méthodes développées, les bases de données de chaque participant (M.N.H.N., ECOSCOPE, GBIF et I.M.B.E. notamment), représentant parfois plusieurs Téraoctets de données, seront utilisées comme cas d'étude. Cet exercice doit permettre de rendre les modèles plus génériques et capables de s'adapter à l'automatisation future des systèmes d'acquisition de données (drones ou systèmes d'observation permanents par exemple).

Un effort devrait être porté sur les fonctions d'analyse de la qualité des données et d'exploration des possibilités d'interopération entre bases de données de différents champs disciplinaires (écologie, sciences humaines et sociales, économie). De nouvelles qualifications (apportées par chaque nouvel usager) permettront de lier des objets décrits par des attributs hétérogènes mais de même nature et de proposer de nouveaux concepts permettant des approches globales. Ce prototype est un élément majeur du projet porté par le consortium IndexMEED dont un des objectifs est d'utiliser les graphes et hypergraphes paramétrables basés sur ces données hétérogènes et distantes (de la molécule à l'écosystème, en passant par les traits de vie, jusqu'aux paysages et aux interactions Homme-milieux) pour alimenter des systèmes d'aide à la décision. Pour atteindre cet objectif, il sera incontournable de documenter une bibliothèque d'algorithmes déjà utilisés par d'autres disciplines ou développés en propre par les chercheurs intéressés dans le domaine des S.T.I.C..

Chapitre 5 : Recommandations, perspectives et conclusions

1. Discussion générale

1.1. Où en est-on ?

Le travail réalisé est la première étape d'une démonstration de ce qu'il est possible et souhaitable de faire dans le domaine environnemental avec des données hétérogènes en exploitant les méthodes liées à la théorie des graphes. Il est démontré qu'il est possible d'utiliser les graphes sans pour autant les formaliser comme dans le cas d'une ontologie, au moins dans un premier temps. Pour autant, ces procédés de visualisation de la donnée peuvent constituer une étape analytique pour développer des thésaurus plus robustes, voire même cerner le champ des possibles concernant un travail de développement d'ontologies. Les cas d'études explorés montrent que l'utilisation de graphes pour visualiser les données hétérogènes a un effet vertueux sur la donnée en elle-même : les besoins de curations deviennent une évidence pour les administrateurs de données (voir chapitre 4), même avec peu de compétences initiales, et les premiers graphes conçus dévoilent un potentiel des données beaucoup plus important que ce pour quoi elles ont été récoltées initialement (voir annexe 8.1 : protocole de curation des données V1.0).

Les graphes bipartites (mêlant deux objets différents sous forme de nœuds du graphe) ne peuvent être construits que sous la condition d'une interopération entre ces deux objets. Cette tentative d'interopération est toujours une bonne démonstration de l'importance du respect des standards et de la qualité des métadonnées. Les ateliers ont montré que cette importance des standards et des métadonnées devient encore plus cruciale lorsqu'on passe à une autre échelle, concernant deux aspects :

i) lorsqu'il s'agit d'agréger des données provenant de producteurs différents mais de même type comme c'est très souvent le cas à large échelle (par exemple dans le cas des fréquences relatives d'espèces entre deux pays comme pour les programmes CIGESMED ou DEVOTES), ou des sites archéologiques décrits dans des bases de données agrégées (ArkeoGIS) ou les carottes de sédiments contenant des pollens comme pour les bases de données d'E.P.D. (voir annexe 8.2 résultat des ateliers sur les cas d'étude IndexMEED),

ii) lorsqu'il s'agit de contextualiser un jeu de données avec des données provenant de systèmes de mesures préexistants ou produits dans un autre cadre.

Les travaux réalisés dans le cadre d'IndexMEED, notamment avec les données des programmes CIGESMED et DEVOTES ont permis de développer puis de tester un prototype. Celui-ci a été conçu avec des spécificités fonctionnelles développées partiellement, le temps de développement informatique disponible pendant la thèse étant insuffisant pour proposer toutes les fonctionnalités souhaitables pour obtenir un outil fiable et performant. Il sera nécessaire de le perfectionner pour que celui-ci devienne un outil didactique et pédagogique permettant de développer les recherches basées sur la théorie des graphes. Ce prototype permet actuellement de visualiser les données d'ArkeoGIS et de CIGESMED et de paramétrer et manipuler des graphes de l'ordre de 10 000 nœuds et quelques centaines de milliers de liens. Il permet surtout de faire émerger dans chaque cas d'étude des questionnements scientifiques concernant non seulement le système observé, mais aussi le système d'observation et la qualité/exploitabilité des paramètres surveillés, ou même la qualité de l'information et du système d'information (visualisation du niveau de renseignement des champs, des erreurs dans les données par exemple, ou comparaison avant et après curation). Des expérimentations ont aussi été réalisées lors des ateliers en utilisant d'autres outils libres (Gephy, Tulip ou Neo4j) avec différentes bases de données environnementales à large échelle (voir le rapport en annexe 8.2).

L'intérêt d'une démarche progressive et itérative d'expérimentation de la fouille de données basée sur les graphes par des publics non spécialistes a été démontré. La simplicité de l'interface a permis d'initier les participants à ces approches tout en proposant des visualisations d'un premier niveau de complexité mêlant par exemple un ou deux types d'objets différents, ou permettant de pondérer les liens (par des fréquences d'espèces par exemple), et de proposer un graphe "valué" (c'est à dire, avec une pondération des liens). Les possibilités offertes par ces premières analyses sont nombreuses, et la perspective de proposer des scénarios basés sur des graphes est aujourd'hui reconnue comme un débouché prometteur. Ces scénarios seront basés sur la classification de motifs de valeurs de contextes qui doivent devenir des indicateurs d'état ou l'alerte utiles pour l'aide à la décision où même comme nous l'avons présenté dans David *et al.* (2015), alimenter des I.D.S.S. (Intelligent Decision Support System). Pour développer ces nouvelles approches, un investissement simultané dans différents domaines de recherche est incontournable.

1.2. Besoins et perspectives générés par ce travail

Ces expérimentations ont mis en évidence des besoins de recherche dans les domaines S.T.I.C. et environnemental et un développement des approches interdisciplinaires :

Concernant les besoins des S.T.I.C., une adaptation des méthodes de clustering sera nécessaire pour pouvoir traiter des données de faible consistance (lorsqu'un paramètre est

peu ou mal renseigné). Les méthodes de clustering appliqué à nos graphes peuvent selon l'avis de spécialistes être parallélisées pour être réalisées avec la grille de calcul française voire européenne permettant une recherche fondatrice sur toutes les problématiques liées au passage à l'échelle (c'est à dire passer de l'ordre de 10 000 nœuds et quelques centaines de milliers de liens dans les expérimentations réalisées à plus d'un milliard de nœuds et plusieurs centaines de milliard de liens). Dans sa communication lors de l'atelier JDEV[114] 2017 animé par IndexMEED, Luc HOGIE présente les possibilités offertes par une nouvelle bibliothèque d'algorithmes de fouille de graphes utilisant la parallélisation. La conception d'un arbre de décision permettant une utilisation éclairée d'un algorithme en fonction des caractéristiques des données considérées est une des clefs pour permettre un usage élargi de ce genre de bibliothèque d'algorithmes.

L'exploration des fréquences des motifs de valeurs de variables de contextes est un autre challenge autant algorithmique que mathématique, mais celui-ci nécessitera aussi un travail de la part des spécialistes sur les valeurs des variables de contexte concernant leur véracité (c'est à dire la capacité à porter une information). La temporalisation[115] des graphes pourra permettre de comprendre comment l'évolution simultanée de certains paramètres contextuels impacte les clusters du graphes (créant ainsi une indication sur un changement dans le système observé) ou au contraire ne semble pas changer la situation environnementale (donnant ainsi des informations sur la capacité de résistance du système observé). Le nombre de modalités que peut prendre chaque variable considérée et l'impact de l'incertitude aux limites entre les modalités[116] devra être pris en compte autant par des approches mathématiques que via l'expertise des thématiciens pour valider les modèles de scénarios.

D'autres défis concernent les disciplines des S.T.I.C. notamment la conception de l'architecture nécessairement répartie des systèmes d'informations délivrant de l'information sur les contextes, notamment sur les équivalences ou compatibilités et terme d'échelle et de précision de variables contextuelles produites dans des contextes différents.

[114] 04-07 juillet : **JDEV** Journées nationales du DEVeloppement logiciel, Marseille

[115] J'appelle temporalisation d'un graphe le fait de reproduire le graphe à différents temps, avec les mêmes objets, les liens qui les relient, tout en restant de même type, différent à chaque temps. Ces graphes successifs permettent de voir évoluer les clusters où les motifs de contextes étaient significatifs, où ils le restent ou où ils le deviennent, permettant ensuite de regrouper ces motifs par type de situation. Cette perspective, évoquée dans tous nos travaux, n'a pas encore été mise en œuvre.

[116] Par exemple, pour une orientation, on peut considérer 4 orientations cardinales possibles (Nord, Sud, Est, Ouest) ou 8 (en ajoutant les intermédiaires). Il faut préciser celle qu'on choisit lorsqu'on est exactement entre deux valeurs et pourquoi (par exemple : "celle toujours le plus au Sud car la lumière est pressenti comme un facteur déterminant", en notant combien de fois on se retrouve dans ce cas).

Du point de vue des disciplines environnementales, et bien que ces travaux ne puissent être réalisés que dans un cadre interdisciplinaire soutenu, les défis pour augmenter la portée des résultats obtenus concernent l'augmentation de la complexité des graphes pour devenir de plus en plus réaliste. Cette augmentation de la complexité se traduit par plus de types de nœuds et de liens (sous la forme de graphes monocouches multipartites ou de graphes multicouches qui peuvent aussi être multipartites dans leurs couches), et des motifs de variables de contextes contenant plus de variables de contextes et/ou plus de possibilité de modalités. Ceci implique de compléter les contextes par de nouveaux types de données, et d'organiser à large échelle et par un plus grand nombre d'opérateurs, un système de validation des formats, des définitions des variables et de leurs modalités. L'usage des nouvelles techniques d'apprentissage automatisées (en premier lieu supervisé, mais aussi non supervisé) est envisageable pour enrichir les graphes, soit pour améliorer les processus de curation de la donnée (prédire la valeur manquante d'une variable), soit pour classer un motif de valeurs de contexte dans une catégorie de scénario. Maîtriser ces différentes méthodes et en développer les usages dans le domaine environnemental demandera de développer des accompagnements solides reposant sur un personnel dédié, pérenne et bien formé.

Ces nouveaux besoins d'accompagnement pour développer les résultats de ces premières expériences sont partagés par tous les participants, avec non seulement l'envie d'augmenter la complexité de ces graphes (comme dit précédemment, augmenter le nombre d'objets, leur type, mélanger différents types de liens, etc.), mais aussi mieux maîtriser les processus d'amélioration de la qualité de la donnée (curation, métadonnées).
Pour atteindre cet objectif d'intégration de plus de données et de plus de complexité, l'enjeu, qui est aussi politique, est de répartir et distribuer l'information, la rendre accessible avec des formats multiples et à un haut niveau de qualité (disponibilité, adaptabilité, réplicabilité, etc.). Nous avons préconisé dans notre proposition d'architecture le développement de la pratique de l'interopération via des systèmes de mirroring concernant la donnée de contexte. Ce mirroring, prenant modèle de la gestion des noms de domaines par les registrars pourraient être une étape de développement importante pour rationaliser la distribution de l'information de contexte. Une telle architecture repose sur une réorganisation des systèmes d'observation et des systèmes d'information dans le domaine de l'environnement.

1.3. Une nécessaire observation normalisée à long terme et à large échelle ...

Les systèmes d'information rendant accessible les données d'observations recueillies via des protocoles mis en œuvre à large échelle sont une nécessité aujourd'hui avérée, et leur coût, même rationalisé par des méthodes et outils moins onéreux, ne devraient pas arrêter de progresser ces prochaines décennies. La considération pour les enjeux liés à la préservation de l'environnement et de la biodiversité sont de plus en plus partagés, même si les méthodes pour rendre l'évaluation de l'état de préservation sont encore balbutiantes. Encore plus balbutiantes et pourtant autant nécessaires sont la co-construction de normes acceptées sur les **variables essentielles de biodiversité** (EBVs) et la mesure des variables de contextes ayant une inférence avec les EBVs pour rendre ces suivis comparables dans le temps et sur de larges espaces. Le développement de ces normes de descripteurs environnementaux relève aujourd'hui d'une prise de décision politique accompagné des moyens nécessaires et d'une meilleure valorisation de l'interdisciplinarité dans les carrières de la recherche. Le développement d'une recherche basée sur une observation sur le long terme nécessite de rendre les systèmes d'observations indépendants des effets de modes qui impactent les appels à projet de recherche, et d'imposer la gratuité (et de fait la non rentabilité) de l'information produite. Pour autant, la mise à disposition de l'information demande des moyens techniques et humains compatibles avec les enjeux. Assurer cette accessibilité ne sera donc possible qu'en rendant aussi les systèmes d'information indépendants des effets de modes qui impactent les appels à projet de recherche.

1.4. … Puis le système d'information : a-t-on mis la charrue avant les bœufs ?

Les processus et outils pour rendre les données accessibles, référencées et bien indexées, normées et/ou standardisées sont aujourd'hui nombreux, peut-être trop pour être bien identifiables en fonction des besoins et des usages de données, et sont eux-mêmes issus d'initiatives souvent déconnectées entre elles. Le défi est aujourd'hui de rendre l'existant (les systèmes d'observations et les données déjà récoltés) réutilisable pour des usages encore non imaginés aujourd'hui (en évaluant leur respect des principes FAIR). Ceci impose d'expliciter le potentiel de la donnée, et de documenter précisément le contexte de recueil de ces données et les limites à leur interprétation, ce qui est difficile à faire *a posteriori* de leur production.

Les problématiques de qualité et d'adaptabilité des systèmes d'information, autant que la problématique de la qualité des données sont devenues primordiales, pour permettre ce suivi

à long terme de la préservation des écosystèmes et de la biodiversité. Le peu de capacité de résilience des habitats et espèces rend indispensable la comparabilité dans le temps et l'espace de leur suivi. Cette notion de la qualité des systèmes d'information, prégnante dans la totalité et dans tous les aspects du cycle de vie de la donnée doit être considérée en même temps que lorsque la question « que veut-on savoir ?» est posée. Beaucoup de systèmes d'observation imposent leur format et leur logique au système d'information, produisant une donnée à visée unique et centrée sur un métier. Les systèmes d'informations, alors développés en parallèle produisent des données pour des systèmes et logiciels propriétaires et verrouillés, multipliant aussi les efforts nécessaires à une bonne qualification de la qualité de la donnée. L'obsolescence des données est alors prévisible. La logique voudrait pourtant que ce soient les systèmes d'information "intercommunautaires" qui imposent les formats et l'organisation des données aux différents systèmes d'observation pour chaque type de données, mais en prenant en compte les impératifs d'exploitabilité de cette donnée en dehors du secteur métier qui produit les données. Actuellement, au sein de R.D.A., des outils et méthodes de qualification de la qualité du respect des principes FAIR, animées par GeoBON, vont dans ce sens. Il y a néanmoins deux importants écueils pour les systèmes actuels : le temps de mise en place de ce genre de référentiel est relativement long, et les besoins sont immédiats, et pour l'instant, ce développement ne prévoit pas plusieurs niveaux d'évaluation, qui faciliterait une démarche progressive mais ordonnée des opérateurs concernant le respect des principes FAIR. Chaque gestionnaire de système d'observation construit donc à sa manière, selon ses moyens et ses compétences, et le respect des principes FAIR dans la construction des systèmes d'information se fait en ordre dispersé.

2. Recommandations [à ce stade]

2.1. Recommandations concernant les protocoles

Même si documenter la méthode de production des données est une règle généralement bien suivie, d'autres paramètres explicites sur le potentiel des données sont souvent moins bien décrits. Chaque protocole scientifique notamment en écologie, a pour objectif de pouvoir produire des données interprétables pour tester des hypothèses en contrôlant un certain nombre de paramètres. Aujourd'hui, de nouvelles approches permettent de formuler des hypothèses en se basant sur les données (modélisation, approches par les graphes). Ces approches nécessitent de connaître mieux la donnée et les critères permettant de vérifier sa véracité.

Le premier d'entre eux concerne les erreurs inévitables dans le cas de la mise en place d'une procédure, surtout si elle est manuelle : une donnée peut avoir une incertitude, ou être

complètement faussée dans certaines conditions (confusion de champs, synonymie, etc.), et ces erreurs, dès qu'elles sont repérées, doivent être documentées même si elles ne sont pas réglées. Cette documentation des aléas peut aussi être très utile en cas de projet de curation des données en vue d'une réutilisation comme une agrégation pour une méta analyse ou une visualisation sous forme de graphe par exemple.

Le producteur de données doit aussi mettre en garde contre les possibles erreurs d'interprétation qui peuvent être faites, par exemple en changeant d'échelle ou en agrégeant des données (par exemple, en utilisant des données issues d'un krigeage comme une valeur vraie).

Souvent, la précision de la donnée est indiquée, mais les raisons du choix du nombre de variables et leurs qualités et défauts sont peu décrits, ce qui nuit non seulement à la reproductibilité du protocole mais aussi à son adaptabilité et donc à l'interopération des données produites au cours du temps.

Ces quelques considérations peuvent aussi concerner la documentation des plans d'échantillonnage : Tenir un registre des coûts, un journal des difficultés rencontrées, des exemples de solutions pratiques ou des illustrations décrivant la structure de ce plan tel que la Figure 52 le montre sont autant d'éléments permettant une meilleure reproductibilité.

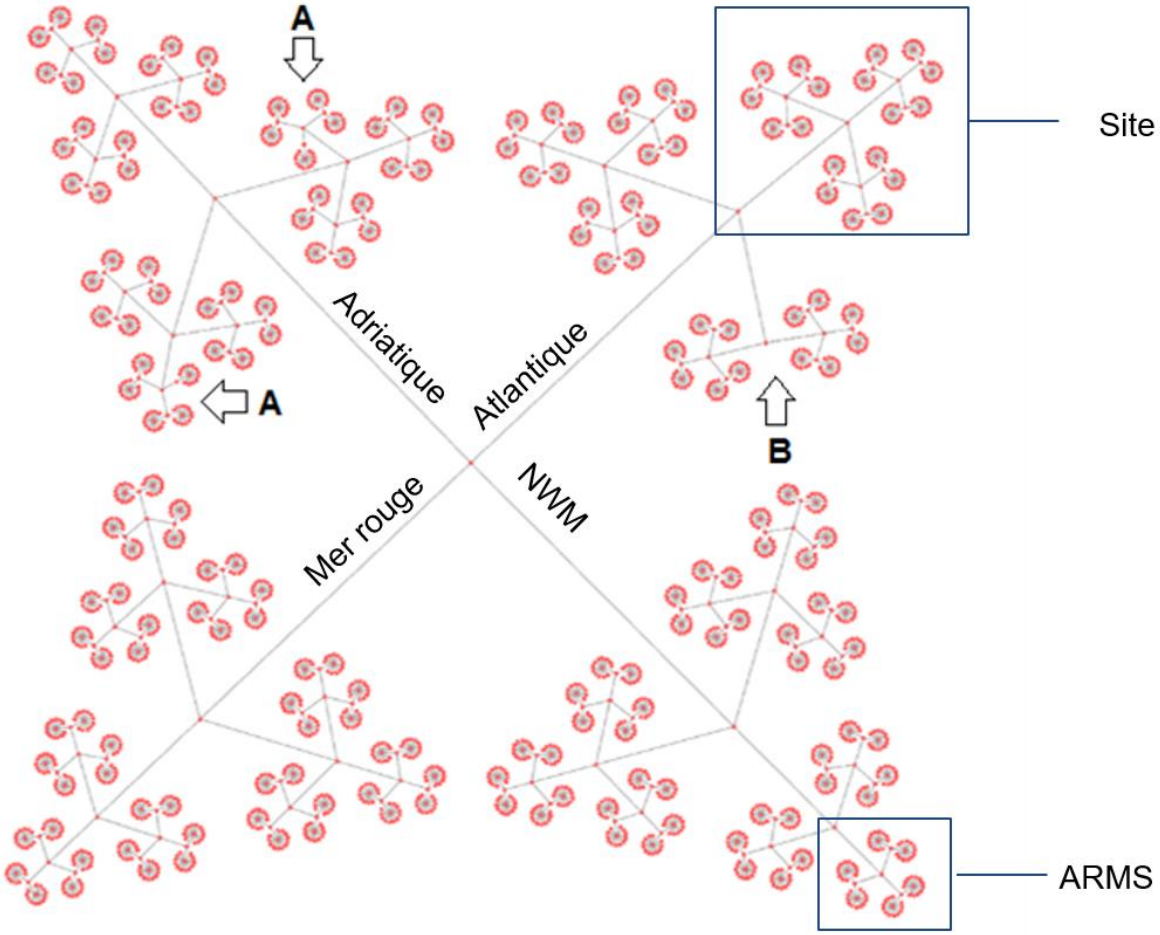


Figure 52 : représentation du plan d'échantillonnage des plaques des ARMS sous forme de graphe, mettant en évidence des éléments manquants ou des erreurs et oublis (des sites avec moins de plaques que d'autres par exemple). Pour cela, on représente les plaques en les reliant au récif, puis en reliant le récif au site, puis le site à la mer de manière à former un arbre (qui est une sorte de graphe). Les nœuds terminaux, en rouge, représentent les faces de plaques. Un simple algorithme de recherche de dissymétrie permet de mettre en évidence soit les pertes d'un dispositif, soit les disparités d'analyse (perte d'une photo d'une plaque, ou oubli de détermination d'une portion d'une plaque). Dans le graphe, les 4 branches de l'arbre représentent les 4 mers considérées (NWM : Nord West Mediterranean sea). Il manque 2 faces de plaques dans une mer (A : perte de 2 photos par l'opérateur), et une ARMS dans une autre mer (B : arrachée du substrat lors d'une tempête).

Enfin, la reproductibilité des protocoles rime avec la simplicité de mise en œuvre, et l'acceptation que la reproductibilité dans des contextes différents est approximée. Lorsque l'on reproduit un protocole, il est recommandable de s'assurer de la rétro compatibilité des nouvelles données avec les observations plus anciennes, et d'amender tous les documents

produits précédemment pour assurer cette reproductibilité, ce qui implique que ceux-ci soient aussi dans des formats accessibles et ouverts (par exemple, un format type Wiki est préférable à un format type PDF). Quand on parle d'accessibilité, il s'agit aussi de rendre cette documentation pratique à l'usage, en créant des fonctionnalités et des services autour de cette documentation en suivant la logique du moindre effort : comme les informations structurées des sites web, il peut être utile de prévoir plusieurs niveau d'accession qui peuvent permettre d'améliorer la maîtrise rapide, puis approfondie des protocoles. Enfin, ce n'est pas toujours une évidence, la bonne application d'un protocole demande parfois des essais et tâtonnements, cette étape peut être facilitée par des indications sur les facteurs humains et les coûts induits (formation, entraînement, inter-calibration, échelles de progression et temps de progression, exercices types avec résultats préparés).

2.2. Recommandations concernant l'inter-opération des systèmes d'informations

La centralisation de toutes les données est un objectif aujourd'hui abandonné car non efficient (les organismes à vocation environnementale ayant tenté l'aventure ont inéluctablement abouti à des modélisations de schémas relationnels labyrinthiques, impossibles à maintenir et faire évoluer). Les gestionnaires de données développent aujourd'hui des systèmes plus spécialisés, modulaires et réplicables, et l'utilisation de la même information à plusieurs échelles et pour plusieurs usages demande de travailler à leur interconnexion. Plusieurs systèmes, voire même un nombre de systèmes indéfini doivent coexister, puisque plusieurs systèmes coexistent pour chaque métier, et il s'en crée toujours de nouveaux.

L'interopérabilité est une qualité souvent définie comme un objectif à atteindre pour un ensemble de systèmes. On ne peut pas dire qu'un système est inter-opérable en lui-même, car cela dépend de l'ensemble de système que l'on considère comme étant à inter-opérer.

A fortiori dans le domaine de l'environnement, plus le nombre de systèmes est grand, et plus ils sont différents, plus il sera difficile de les rendre complètement interopérables. L'interopérabilité dépend donc du domaine concerné, et est plus difficile dans le cadre de travaux interdisciplinaires. Cela signifie que l'interopérabilité est un concept relatif et n'est jamais complètement atteint. Tant que tout ne sera pas standardisé, l'interopérabilité complète de tous les systèmes est impossible. Tout standardiser est impossible à réaliser, d'autant qu'au fur et à mesure du temps, de nouvelles recherches inventent de nouveaux paramètres issus de nouveaux protocoles non standardisés. La quête de l'interopérabilité est donc en fait synonyme de priorisation de certaines interopérations par rapport à d'autres, et

elle doit être qualifiée et évaluée en fonction de la complexité des systèmes à interopérer et en fonction des objectifs de cette interopération.

L'interopération de deux systèmes peut être définie par la possibilité pour chaque système d'utiliser tout ou partie de l'information de l'autre système. Chaque système d'information dit "métier" contient des données qui n'auront pas d'intérêt pour l'autre système. Ces données spécialisées peuvent être "taguées" pour être identifiées par des usagers potentiels comme difficilement interopérables. *A contrario*, dans le cas des systèmes d'information agrégateurs (souvent à des échelles géographiques régionales, nationales ou internationales), l'interopération est faite sur une petite partie des données représentant un dénominateur de données commun à tous les systèmes "métiers" d'intérêt pour l'agrégateur. Ces données "dénominateur commun" sont les premières à identifier et à standardiser dans un système métier pour augmenter son interopérabilité.

L'adoption commune d'un format pour un type de données de mêmes caractéristiques (définition, précision, fréquence, échelle, modalités et/ou gamme de valeurs possibles, unités etc.) par deux systèmes d'information leur permet d'intégrer réciproquement les données de l'autre système ; En considérant chaque descripteur un par un (soit champ par champ), le premier travail pour interopérer consiste donc à rendre possible l'évaluation de l'équivalence des caractéristiques de ce type de données par d'autres systèmes. Pour cela, il faut les qualifier dans un système de métadonnées intelligible et explicite, si possible exploitable par une machine. Cette qualification est plus facilement explicite si elle s'appuie sur des standards. Pour l'instant, ceux-ci sont peu développés dans le domaine de la biodiversité. Les standards adoptés par les systèmes agrégateurs internationaux doivent être considérés en priorité pour améliorer l'interopérabilité "globale[117]" d'un système d'information métier.

D'un point de vue interdisciplinaire, le risque est l'adoption de standards contradictoires. La mise en place d'une politique de l'interopérabilité, s'appuyant sur des moyens et une gouvernance éclairée par des spécialistes de la donnée environnementale permettrait d'éviter cet écueil. Pour autant, les systèmes de "*mirroring*" et de services web autour de la donnée permettent de la rendre accessible à de multiples formats, en les conservant initialement à un seul endroit (administrable par le producteur de données ou son délégataire). Plus la donnée est accessible à de multiples formats (sous forme de flux ouverts et documentés, en X.M.L., J.SON, C.S.V.), plus on augmente son potentiel d''interopération car plus on permet à des interfaces techniques différentes de s'y connecter. Nous avons testé avec le prototype d'IndexMed la possibilité de paramétrer ces flux. *Via* un service de ce type, chaque utilisateur peut non seulement paramétrer les formats des flux, mais aussi faire des requêtes en fonction de valeurs d'intérêt. Au-delà de son pouvoir analytique (que peut-

[117] Ici, par interopérabilité "globale", on entend interopérabilité tous domaines confondus

on faire comme hypothèses en voyant un graphe), la visualisation des sélections faites sous forme de graphe a montré son intérêt pour un utilisateur non expert, et pour évaluer les qualités et la consistance d'un jeu de données, puis les faire évoluer.

Ce type de service pourrait être développé par les agrégateurs avec un plus grand nombre de fonctionnalités (notamment adapter les unités, fréquences, échelles, modalités et/ou gamme de valeurs possibles en fonction du dénominateur commun entre différentes sources de données, ou faire des correspondances entre définitions des données de deux disciplines différentes), permettant ainsi une prospection plus complexe et poussée des outils interdisciplinaires comme le prototype d'IndexMEED, voire l'alimentation de futurs systèmes d'aides à la décision comme les I.D.S.S..

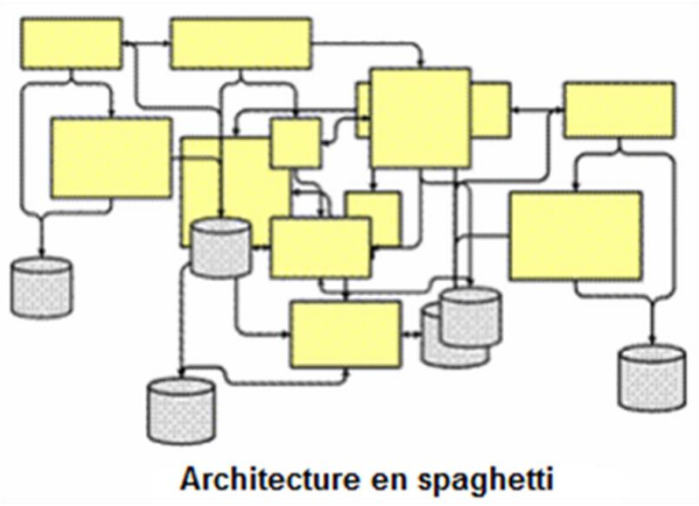


Figure 53 : Une interconnexion anarchique entre systèmes d'information échangeant des données peut aboutir à une architecture en spaghetti, multipliant les mouvements de données et nuisant à la traçabilité et donc à la mise à jour de celle-ci. Une répartition des rôles entre les différents acteurs et leurs plateformes est nécessaire.

D'un point de vue global, l'écueil d'un développement désorganisé des flux de données dans une architecture forcément décentralisée est d'obtenir une architecture en "spaghetti" (figure 53). La mise en place d'une politique de l'interopérabilité, s'appuyant sur des moyens et une gouvernance éclairée par des spécialistes de la donnée environnementale permettrait d'éviter cet écueil (il est possible de faire une analogie avec le développement du web : les liens entre les sites web ont suivi le même type de développement avant d'être rationalisé par une gouvernance, celle des grands moteurs de recherche, qui ont indexé et qualifié les contenus).

La gouvernance globale de ces systèmes doit reposer sur l'indexation et la traçabilité de la donnée, et donc du producteur de la donnée, et de toutes ses transformations et donc de tous les transformateurs de la donnée. Cette traçabilité implique la création d'un gestionnaire d'autorités piloté par cette gouvernance et organisée par points nodaux (disciplinaires et/ou géographiques - là encore, à l'image des registrars des noms de domaines du web). Seul l'utilisateur final n'a pas à être tracé : tracer un utilisateur sans son consentement est une attitude contre productrice et néanmoins pourtant souvent priorisé par les producteurs de données (sans compter que ce traçage est souvent non conforme au respect du R.G.P.D.).
Finalement, du point de vue du producteur, deux grands principes permettent d'augmenter l'interopérabilité de son propre système : i) les normes et standards et leurs évolutions sont à suivre scrupuleusement, avec des modèles de données les plus simples possible (un des écueils les plus courant est la sur qualité, qui induit une difficulté à faire évoluer le système à long terme), ii) le producteur de données doit anticiper l'évolution des usages et développer les accès aux données[118] en améliorant la qualité et la souplesse des services autour de la donnée existante, ce qui demande de travailler sur des formats ouverts et gratuits avec un personnel dédié, compétent (c'est à dire spécialiste de la donnée) et pérenne.
Enfin, l'interopérabilité interdisciplinaire (ou globale) est plus conditionnée par le temps passé par des spécialistes pour travailler, comprendre et construire avec d'autres spécialistes de disciplines différentes (notamment sur la sémantique) que par les capacités de deux systèmes à se connecter.

2.3. Recommandations concernant l'utilisabilité, l'utilisation / la réutilisation des données

<u>Développer la culture des données et leur « réutilisabilité »</u>

D'une manière générale les données, lorsqu'elles sont rassemblées, sont souvent au mieux « empilées » et n'utilisent pas les mêmes standards. Les typologies de champs ne sont la plupart du temps pas uniformisées lorsque ceux-ci contiennent le même type d'information (géographiques, temporelles, noms d'auteurs, objets, constructions humaines…) ; cependant certains référentiels sont petit à petit institutionnalisés. *De facto*, les correspondances entre les données générées par ces études de différentes disciplines, portant cependant sur les mêmes territoires, sont encore peu aisées à réaliser, surtout sur

[118] L'évolutivité des composantes du système d'information fait aujourd'hui partie de la définition de l'interopérabilité selon l'AFUL : "L'interopérabilité est la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs et ce sans restriction d'accès ou de mise en œuvre."

un temps long. Améliorer le potentiel de ces données (et donc ainsi leur valeur) nécessite de mettre en place une stratégie de curation des données, d'organiser la gestion de leur cycle de vie et leur accès (*via* des plans de gestion des données notamment), selon les grands principes FAIR (*Trouvable*/Findable, Accessible, Interoperable, and *Réutilisable*/Reusable).

Le lendemain du séminaire du GRAAL, deux ateliers ont permis de présenter les outils pour gérer et publier ces métadonnées (« Présentation du portail d'ECOSCOPE » et « Comment générer son data paper avec l'IPT du GBIF »), et de souligner l'importance i) d'un bon niveau de qualité des métadonnées ii) de l'interopérabilité des portails de métadonnées iii) des plans de gestion des données pourtant quasiment systématiquement absents « pour le moins » dans les systèmes présentés lors des journées du GRAAL.

Mieux connaître le potentiel des données pour mieux inter-opérer

Les métadonnées[119] favorisent la réutilisation des données et les collaborations scientifiques à des fins de recherche ou d'expertise. La documentation des données est une bonne pratique de gestion et une étape souvent requise dans les projets, mais sa réalisation effective est mal contrôlée. Les métadonnées peuvent être importées et exportées d'un système d'information à l'autre, limitant ainsi les efforts des chercheurs pour valoriser leurs travaux sur les plans national et international, voire à travers des "data papers". En France, l'amélioration des synergies entre bases de données sur la biodiversité (qui induit de fait une amélioration des synergies entre acteurs) s'appuie jusqu'à aujourd'hui sur l'infrastructure ECOSCOPE. Celle-ci déploie un portail de métadonnées avec les observatoires de recherche sur la biodiversité qui travaillent à différents niveaux d'organisation, de l'infraspécifique aux écosystèmes, des ressources génétiques au fonctionnement des milieux. L'objectif premier de ce portail est de porter à connaissance les jeux de données existants, leurs contextes d'acquisition, la qualité ainsi que les conditions d'accès et d'utilisation. Dans un contexte d'hétérogénéité et de dispersion des sources de données, ce portail catalogue les jeux de données et les ressources biologiques sous des formats standards (ISO/INSPIRE) ou en usage (EML, NCD) parmi les chercheurs en écologie. Les métadonnées permettent également d'analyser le paysage de la recherche - ses forces et lacunes - et de contribuer aux initiatives globales, comme le développement des Variables Essentielles de Biodiversité (EBV), proposées par GEO BON (Group on Earth Observation – Biodiversity Observation Network). Ces bases de métadonnées, dont l'exploitation sous forme de graphes a été récemment testée (Muñoz *et al.*, 2016), permettront d'estimer les opportunités d'interopération entre systèmes d'information (Loi "Pour une République numérique", 2016). Pour l'instant, le nombre de producteurs alimentant ces "méta bases" reste encore très limité et ceux-ci ne disposent pas ou peu de moyens pérennes, même si les producteurs les plus importants commencent à investir dans de nouvelles solutions. Les démarches d'animation scientifique visant à augmenter le potentiel de la donnée (et donc sa valeur) est pourtant une (voire la) clef pour inciter la création d'une métadonnée de qualité.

[119] Les métadonnées sont des informations sur les données et les jeux de données, permettant de les décrire et les cataloguer, ainsi que les dispositifs et structures dans le cadre desquels les données sont produites.

2.4. Recommandations concernant les traitements de données hétérogènes, la fouille de données et les approches par les graphes

L'utilisation de la théorie des graphes pour exploiter des données environnementales se précise

Les journées du GRAAL[120] et les propositions qui en ont découlé en préfigurent aussi des nouvelles ; elles s'insèrent dans un projet à long terme véritablement interdisciplinaire qui organise des compétences et l'investissement de personnel de l'INEE (INstitut Ecologie et Environnement du CNRS) au-delà du spectre déjà large de cet institut, et qui favorise enfin, grâce à une préparation de 3 ans, une implication réelle et sur le long terme de scientifiques des S.T.I.C. en sciences de l'environnement, tout en leur permettant de résoudre des questions scientifiques sur les graphes qui leur sont propres. Le support de la mise en place de ces méthodes (accès aux bases de données des partenaires) est encore dans un état de développement préliminaire, mais le travail annoncé devrait avoir des effets vertueux car respectant les très reconnus mais si mal appliqués principes FAIR, principes d'ailleurs pourtant désignés (comme au séminaire ministériel de l'Alliance nationale de recherche pour l'environnement – AllEnvi en 2014) comme incontournables pour les projets d'observatoire, surtout dans une dimension telle que les relations Homme-milieux. Ceux-ci exploitent les bases de métadonnées (répertoires) dans lesquelles différents acteurs aux périmètres propres (dont ECOSCOPE, le SINP ou le GBIF) ont investi ces dernières années. Répertoires qui pourraient rester déconnectés et incomplets, sans cette analyse scientifique de l'existant et du potentiel des métadonnées qui leur sont propres. Cette analyse a le mérite de permettre un état des lieux, et surtout d'être généralisable à tous les champs disciplinaires présents en écologie et environnement.

Quelques recommandations concernant l'initiation aux traitements de données hétérogènes et aux approches par les graphes

Les ateliers sur la visualisation des données environnementales sous forme de graphe et les tests faits avec le prototype IndexMed sur des données de CIGESMED et d'ArkeoGIS ont montré que la structuration des données sous forme de nœuds (ou entité ou objet) et de liens (attribut, et ou valeurs / modalités possibles de chaque attribut) étaient loin d'être des taches évidentes pour les participants.

[120] https://indexmed2016.sciencesconf.org/

Afin d'améliorer l'efficacité de la démonstration du potentiel de ces approches, la première recommandation consiste à commencer l'expérimentation des graphes sur une partie simple et assez complète de la base de données, en utilisant une entité (un objet) avec un nombre d'enregistrement suffisant pour obtenir un graphe un peu structuré (c'est à dire ayant des clusters bien visibles). Il est nécessaire de débuter par des graphes simples, c'est à dire avec un seul type de nœuds et des liens bidirectionnels (versus les graphes orientés[121] qui sont plus complexes à conceptualiser). Ensuite, il faut choisir dans un premier temps des conditions "idéales" pour une démonstration : i) un petit nombre de descripteurs bien renseignés, trois étant l'idéal (chaque objet n'ayant pas de valeurs pour cette modalité n'étant pas relié à d'autres objets se trouvera déconnecté du graphe) ii) des descripteurs dont le nombre de modalité n'est pas trop grand (sinon, le graphe peut être fragmenté en plein de petits graphes, et l'interprétation en est plus difficile), iii) des descripteurs dont les valeurs sont assez bien répartis entre les objets (sinon, dans le cas d'un descripteur prenant 90% la même valeur, 90% des objets sont tous reliés entre eux, ce qui fait un gros cluster très prévisible et affaiblit la démonstration). Si ces conditions ne sont pas réunies, il faudra recommencer le processus de curation pour les remplir (par exemple en regroupant deux valeurs de descripteurs en une seule), et ainsi aboutir à un graphe simple mais parlant. Ensuite, les autres descripteurs et la répartition des valeurs de ces descripteurs peuvent être "affichés" dans le graphe sans participer à sa topologie comme cela a été expliqué dans la partie 1.4 du chapitre 4. L'enjeu est de visualiser des combinaisons de valeurs de facteurs (ou des combinaisons de groupes de valeurs) ayant une fréquence plus grande dans un cluster que dans d'autres. Après visualisation, des algorithmes d'analyse de la significativité de ces fréquences sont utilisés.

Quelques précautions doivent être prises dans le choix des descripteurs de contexte pour rendre ces résultats robustes : Chaque descripteur doit être le plus indépendant possible des autres (par exemple, étudier la répartition de la lumière et de la profondeur sur le même graphe a de grande chance de donner la même répartition de ces deux facteurs de contextes).

La qualité des graphes est plus facile si on s'appuie sur des données catégorielles, même si des données numériques peuvent aussi être utilisées comme celles utilisées pour les graphes où ce sont des fréquences d'espèces qui relient les quadrats photo de CIGESMED et les photos d'ARMS de DEVOTES. Nous avons vu qu'il était possible de placer plusieurs objets en tant que nœuds et que les liens étaient alors issus soit d'une appartenance de l'un à l'autre (comme le graphe du plan d'échantillonnage de DEVOTES ou une Mer, contient des

[121] Un graphe est dit orienté si ses arêtes ne peuvent être parcourues que dans un sens. Par exemple, deux nœuds représentant des personnes sont dits "amis", la relation est bijective et non orientée. Mais si l'attribut est "aime bien", la relation n'est pas forcément réciproque, et est donc dite "orientée".

sites qui contiennent des ARMS qui contiennent des plaques qui contiennent des faces), soit de propriétés communes (un thésaurus de thématiques de données par exemple, comme dans le cas d'étude des métadonnées d'ECOSCOPE voir en annexe 9.1, ou des mobiliers et immobiliers communs à des bases de données et/ou à des sites archéologiques pour ArkeoGIS voir en annexe 9.1). Ces approches permettant de construire des graphes bipartites et/ou multipartites permettent de développer des représentations intégratives et multimodales, et les perspectives données par les graphes lorsque l'on les temporalise (ou qu'on y intègre des animations pour comparer des changements de motifs de contextes comme proposé dans la figure 49 (A, B et C) de quadrats photo de CIGESMED) montrent que les graphes contextualisés peuvent être des outils puissants pour construire des modèles prévisionnels).

Les graphes "modèles prévisionnels" ne sont aujourd'hui pas encore opérationnels à large échelle dans le domaine de l'environnement car les mathématiques combinatoire et les données catégorielles (notamment de contextes) sont encore peu utilisées dans les suivis (les mesures et traitements sont souvent numériques et la considération des inférences est surtout basée sur des distances euclidiennes). A cela il faut ajouter que i) l'établissement de pondération qui décrivent les effets relatifs et les interactions entre facteurs suivis est difficiles ii) les mesures des variables de contextes les plus importantes sont encore mal maîtrisées (exemple : particulièrement en milieu marin, la lumière, le courant…) iii) la variabilité naturelle de ces variables est mal documentée iv) l'état initial possible à mesurer est déjà un état perturbé.

2.5. Recommandations concernant les facteurs humains

La qualité de l'information, une problématique liée aux facteurs humains

Travailler sur la qualité de l'information, de son accessibilité et de sa réutilisabilité demande de s'attaquer à différentes problématiques avec un focus sur l'aspect qualification de la qualité, lui-même fortement lié à leur interprétation par l'homme.

-> **Concernant le modèle de production de la donnée**, les systèmes se devenant de plus en plus complexes, et pour pouvoir développer des systèmes à large échelle, il est nécessaire de développer :

- Une évaluation des besoins antagonistes (i) de normalisation ou de suivi des normes existantes, (ii) de diversification de la donnée sur la biodiversité, de ses méthodes d'acquisition et de ses formats et de leurs conséquences respectives sur l'efficacité des types de recherche associées,

- Un suivi du rapport entre (i) données très homogènes, notamment les données « calculées » ou mesurées *versus* les données « d'interprétation », plus sujettes à variation liée à l'observateur, ii) les méthodes de valorisation des producteurs de données *versus* les moyens de pérenniser des systèmes d'observation de données « interprétées » (et donc coûteuses en temps / homme),
- Des méthodes d'agrégation et de fouille de données pour des usages secondaires et les scénarios envisageables pour leur exploitation à différentes échelles en fonction du type de donnée.

Nous l'avons vu avec les graphes développés dans le cadre de CIGESMED, tous ces aspects concernant l'efficacité du modèle de production des données sont confrontés à des besoins humains souvent antagonistes (données toujours différentes pour les nouveaux protocoles, variabilité des observations à large échelle, temps en moyens nécessaires à l'inter-calibration, choix d'un vocabulaire contrôlé au-delà de sa discipline. L'amélioration des modèles de données demandera donc forcément de les prendre en compte.

-> **Concernant le développement des services et usages de la donnée**, différents aspects vont devenir stratégiques :

- La problématique du cycle de vie de la donnée, de sa véracité à long terme et de son obsolescence (ou durée de validité ?),
- La valeur de la donnée et les indicateurs d'utilisation effective de cette donnée, comme un nouveau critère d'évaluation des chercheurs travaillant sur la biodiversité,
- Le contexte juridique de la donnée sur la biodiversité issue de la recherche, au sein des organismes de recherche, mais aussi pour toute utilisation secondaire, externe au contexte « recherche »,
- Les outils nécessaires à une véritable traçabilité de la donnée, quelles que soient les transformations subies par les données brutes (agrégation, moyenne, division, requalification, etc.).

Nous avons vu lors de l'atelier de visualisation de données hétérogènes sous forme de graphe que les développements de nouveaux services doivent être réalisés non seulement en tenant compte de ces aspects méthodologiques et techniques mais aussi et surtout sans oublier qu'ils peuvent devenir un frein (par leur complexité et la rigueur nécessaire à leur réalisation), à la prise en compte efficace des besoins humains, et que le caractère progressif de leur mise en œuvre doit être un principe prioritaire.

Développer tous ces aspects demande de faire adopter par l'ensemble des communautés intéressées des définitions communes de critères d'investigation de la donnée et de l'évaluation de l'efficacité de leur usage. Ce facteur « évaluation » et le développement d'une sémantique nécessaire sur l'évaluation de la qualité de la donnée ne sont pour l'instant pas

pris en compte. Ils sont pourtant une autre clef incontournable d'une meilleure interopération des systèmes d'information environnementaux. Les moyens qui sont alloués au développement d'infrastructures de gestion de services autour de la donnée de biodiversité ne servent pour l'instant pas à développer et à diffuser des outils intuitifs et appropriés sur la gestion de la qualité de la donnée, ce qui est pourtant une étape *sine qua non* d'une organisation pérenne d'un « système de systèmes d'information » rationalisé où la démarche qualité, comme dans d'autres domaines, est centrale.

Une science mieux partagée

Les niveaux de réutilisation des données, de retour d'information et leurs impacts directs sur la volonté des producteurs de mieux partager des données sont en pleine évolution. Même si le besoin de principes partagés comme FAIR est ancien, leur publication est somme toute récente, et même l'application des principes évidents peine à se diffuser. Il semble qu'à ce stade le véritable frein soit généré par la concurrence entre les structures produisant le même type de données. En complément, les coûts de l'archivage public des données pourraient être sous-estimés, en particulier en ce qui concerne les études à long terme (Mills *et al.*, 2015) pour lesquelles la donnée a par exemple de grandes chances d'évoluer, de changer de format, ou de devoir s'adapter à un nouveau standard ou règlement (au plan national et international), sans que les moyens ne soient initialement prévus pour le faire. D'autre part, les données utilisables par les scientifiques seront de moins en moins produites en interne, et de plus en plus mutualisables. A cela s'ajoute le développement des sciences participatives, qui, un peu à l'image de l'arrivée du numérique dans le secteur de la photographie, va permettre à de plus en plus d'amateurs de concurrencer les spécialistes à de hauts niveaux dans les domaines de la biodiversité (déjà développée dans les domaines naturalistes). Pour autant, et vu que ces amateurs n'auront pas de salaire, la principale reconnaissance qu'auront ces nouveaux passionnés se constitue de la reconnaissance par leur pair et de notoriété. Dans ce sens, une évolution des 4 principes FAIR autour de la donnée vers 5 principes FAIRc[122] en y ajoutant le terme "citable" devrait concourir à cette notoriété. Il faut espérer que ce cinquième principe n'aura pas les mêmes effets pernicieux que la course à la publication, contre-productive concernant la qualité de chaque publication et bien connue dans de nombreux domaines scientifiques (et le fameux publish or perish).
Les centres de recherche qui effectuent la synthèse et l'analyse des données à large échelle (par exemple, CESAB (Centre d'Etude et de Synthèse et d'Analyse sur la Biodiversité) en France, ACEAS en Australie (Australian center making data synthesis and analysis) ont besoin d'accéder aux données à large échelle et ont réalisé une enquête montrant que les

[122] Le c de FAIRc est minuscule car le terme est en cours d'adoption

mécanismes d'attribution de crédits sont essentiels pour que les chercheurs partagent leurs données. Les verrous relevés par ACEAS sont le manque de métadonnées et donc le manque de capacité à évaluer la pertinence des données et bien sûr l'accès limité aux données. Les défis importants liés à l'acquisition de données qui sont mis en avant par cette étude sont la complexité et la difficulté dans l'obtention de l'autorisation d'accéder aux données, le temps et les contraintes financières associées à cette fouille /recherche de données pertinentes et "l'assurance qualité" liée aux jeux de données obtenus (Specht et al., 2015).

Les mécanismes de récompense des chercheurs vertueux sont essentiels, mais les ressources doivent être suffisamment expliquées pour éviter les hypothèses et les interprétations erronées qui, lorsqu'elles sont réutilisées et publiées, sont difficiles à éliminer (Mills *et al.*, 2015). Cette condition d'utilisabilité nécessite la prise en compte de deux besoins corollaires : des systèmes de qualification du partage de la donnée et de la qualité de ce partage standardisés, et une gestion des mécanismes d'évaluation de ce partage et de récompense orchestrée avec les moyens nécessaires.

Des objectifs mieux compris et partagés pour l'aide à la décision

Le domaine des systèmes d'aide à la décision (DSS Decision Support Systems) se concentre sur le développement de logiciels interactifs capables d'analyser des données d'un système et de fournir des réponses aux questions décisionnelles des utilisateurs, censées aider une personne ou un groupe à prendre de meilleures décisions. Early D.S.S. (Little, 1970) a basé son système sur une surveillance simple ; Plus tard, la simulation basée sur un modèle a introduit des analyses de type «what-if[123]» sous forme d'Intelligent DSS (I.D.S.S.) (Marakas, 1999) comprenait des connaissances de domaines spécifiques et des capacités de raisonnement automatique. Jusqu'à présent, des efforts importants sont nécessaires pour développer des (I.)D.S.S. dédiés pour chaque application particulière (Varanon *et al.*, 2007, Power *et al.*, 2007) et certaines expériences peuvent être considérées comme réussies dans plusieurs domaines, comme la sécurité des personnes (Marschollek, 2012), la gestion de l'eau (Pallottino *et al.*, 2005), les écosystèmes forestiers (Nute *et al.*, 2004) ou la pollution atmosphérique (Oprea *et al.*, 2005). Cependant, la mise à niveau de ces plateformes pour

[123] Une analyse de type *what if* est un processus de détermination des effets de la variation de valeurs de paramètres de contextes sur les résultats dans un modèle statistique ou calcul de tableur par des changements systématiques dans les facteurs rentrés (*input*). Par exemple, dans un processus financier, de nombreux facteurs, tels que les taux d'imposition futurs, les taux d'intérêt, les taux d'inflation, le nombre d'employés, les dépenses, sont variables dans la mesure où ils peuvent s'écarter des valeurs attendues. Cette approche est aussi appelée analyse de sensibilité. Ce type d'analyse peut être faite avec des facteurs environnementaux associés à des clusters de graphes, eux-mêmes construits à partir de fréquences d'espèces.

intégrer de nouveaux facteurs de risque de contrôle, de nouvelles connexions de capteurs ou pour tenir compte de nouveaux modèles prédictifs devient coûteuse en temps et en moyens.

La nouvelle génération I.D.S.S. offre une intégration suffisante pour réaliser une approche vraiment holistique (tenant compte non seulement de la surveillance des paramètres isolés, mais aussi des informations provenant des différentes sources de données disponibles, des activités développées dans la communauté et de tous les types d'informations disponibles (images, mesures qualitatives, données explicites ou non traitées [informations implicites dans un texte, un bruit de fond, une vidéo], les connaissances, les documents, les tweets, etc.) et d'obtenir une architecture de système de décision suffisamment flexible pour faciliter l'adaptation du système aux progrès de l'état de l'art. De nouvelles architectures doivent être conçues pour permettre des mises à niveau flexibles, ou des changements de domaine de ce genre de plateformes d'une manière plus aisée (Poch *et al.*, 2004, Rajasekaram et Nandalal, 2007, Gibert *et al.*, 2006) l'I.D.S.S. doit combiner des modèles informatisés, analytiques et basés sur le savoir (y compris les connaissances produites à dire d'experts), ainsi que certains raisonnements standardisés (S. Koch et M. Hagglund, 2010, Helmer *et al.*, 2010, Sànchez-Marrè et Gibert, 2015) pour fournir un soutien approprié aux gestionnaires, même s'il n'y a pas encore beaucoup d'expériences opérationnelles en cours sur cette approche. Les expériences effectuées dans ce travail concernant la construction de graphes et la récupération de motifs de contextes issus de clustering de graphe nous semblent prometteurs pour alimenter les raisonnements standardisés. Ils placent une marche intermédiaire entre les analyses multivariées et statistiques classiques (qui sont souvent uniquement des approches globales de la donnée) et les systèmes de classifications (qui rendent possible une qualification commune de données hétérogènes multi-sources) d'une part, et d'autre part les approches s'appuyant sur des ontologies, qui sont difficiles à mettre en œuvre dans des cas de données hétérogènes et multi sources.

L'écologie appartient à un ensemble de domaines critiques où les mauvaises décisions peuvent avoir des conséquences tragiques. La prise de décisions par les I.D.S.S. devrait être collaborative et non contradictoire, non seulement pour trouver des solutions optimales ou sous-optimales, mais pour rendre l'ensemble du processus plus ouvert et plus transparent. Le système devra faire face à l'incertitude inhérente aux décisions et les décisions doivent informer et impliquer ceux qui doivent vivre avec les (conséquences des) décisions.

Quelle stratégie pour encourager le partage des données et l'augmentation de sa qualité (Rewarding recommendation for Data Sharing) ?

Le partage de la donnée est aujourd'hui une gageure qui n'est pas considérée comme prioritaire par la plupart des producteurs de données dans le domaine de la recherche en écologie. Pourtant, il s'agit là théoriquement d'une condition *sine qua non* pour obtenir de futurs financements de programmes de recherche, notamment européens. Parmi les principaux blocage à cette prise en considération, le manque de moyens nécessaires à ce partage et l'accessibilité des formations à mettre en œuvre pour le faire de manière correcte sont souvent cités. Pourtant, certains producteurs de données, même avec des moyens limités, parviennent à mettre en œuvre une politique de partage de données efficace et sur le long terme. Il s'avère que les blocages pour la mise en œuvre du partage de données sont souvent liés à des facteurs humains surtout d'ordre social ou psychologique. Lors du développement de systèmes d'information distribués, la mise en place d'une stratégie, de processus et d'outils efficaces pour encourager le partage de données et le faire reconnaître doivent être une priorité. Ces processus d'encouragement pour un meilleur partage de données appelé aussi "rewarding for data sharing" doivent être mises en place à plusieurs niveaux en même temps (échelles temporelles, publics visés, type de rewarding[124]... etc.) et adaptés aux contraintes rencontrées dans différentes disciplines pour créer des synergies capables de créer de véritables changements de comportements. Ces adaptations peuvent être faites à partir de travaux interdisciplinaires qui commencent à être mis en œuvre dans la communauté R.D.A. (voir focus sur Sharing Rewards and Credit [SHARC] I.G.)

[124] Il n'existe pas vraiment de traduction française exacte du concept de rewarding, car on peut parler à la fois de reconnaissance, remerciement ou gratification, donc, pour éviter toute ambiguïté, je continuerais à utiliser le terme anglais.

Focus sur SHARC, nouvel "interest group (I.G.)"

Dans le cadre de la communauté internationale et interdisciplinaire R.D.A.

SHARC est aujourd'hui un groupe d'intérêt reconnu et approuvé au sein de R.D.A. (Research Data Alliance) qui cherche à analyser et à améliorer les mécanismes de crédit et de récompense dans le processus de partage de données / ressources[125]. Actuellement, sept communautés différentes sont représentées dans le groupe (Biologie et biomédecine (7 personnes), Sciences et technologies de l'information (3 personnes), Données géospatiales (1 personne), Biologie marine (1 personne), Biodiversité (4 personnes), Écologie industrielle (1 personne), Bioéthique (4 personnes).

Les questions qui y sont abordées sont transversales à de nombreux domaines académiques et à plusieurs communautés. Elles doivent aussi être déclinées selon le secteur et le périmètre académique pour lesquelles, partant d'un socle commun, des recommandations sensiblement différentes peuvent être appliquées. Sachant que certaines communautés sont beaucoup plus avancées que d'autres, un autre intérêt de ce groupe est de favoriser la diffusion et l'adaptation de bonnes pratiques, d'outils, de méthodes et surtout d'organisation du rewarding entre humains en s'appuyant sur les systèmes d'information de manière à rendre plus efficace le partage de données.

Le groupe SHARC IG a quatre objectifs principaux :

- Faire un bilan des mécanismes de "rewarding" existants dans diverses communautés, ainsi que leurs limites et identifier les facteurs qui pourraient améliorer le processus et optimiser le partage des ressources biologiques ; c'est-à-dire des données et des échantillons physiques (ex : outils, incitations, exigences ...),
- Utiliser cette analyse pour encourager l'inclusion de critères liés aux biblio ressources concernant le partage des données dans le processus d'évaluation de la recherche au niveau institutionnel européen (c'est-à-dire sans rendre cette activité obligatoire, augmenter la cohérence entre l'évaluation et la pratique réelle).
- Diffuser l'information et les résultats aux diverses communautés d'intervenants,

[125] Réunion précédente dans le cadre de Research Data Alliance : https://www.rd-alliance.org/how-give-credit-scientists-their-involvement-making-data-samples-available-sharing-rda-9th-plenary).

- Développer un processus d'adoption par étapes de principes et de mesures de mise en œuvre adaptés aux contextes nationaux, locaux et institutionnels.

Des exemples pris dans le cadre de CIGESMED et de DEVOTES sont donnés plus bas.
Source : https://www.rd-alliance.org/group/short-presentation-sharing-rewards-and-credit-sharc-ig/case-statement/sharc-sharing-reward

Une des définitions en cours de discussion du rewarding for data sharing s'inspirera sans doute de Murlis et al., 2004 pour lesquels :
"Reward management is concerned with the formulation and implementation of strategies and policies that aim to reward people fairly, equitably and consistently in accordance with their value to the organization."

En pratique, qu'entend-t-on par rewarding recommandation ? On peut le comprendre de différentes manières : augmenter soit leur utilisabilité, soit leur quantité, soit leurs qualités ou soit leur impact (ou des combinaisons de ces aspects). Chacun de ces aspect mérite d'être exploré, à la lumière des expérimentations et des contraintes opérationnelles comme celles que nous avons mis en évidence dans l'exploitation des données issues des protocoles des programmes CIGESMED et DEVOTES.

Concernant l'utilisabilité du rewarding, il faudrait lister les qualités qu'on en attend pour le mettre en œuvre. On pourrait ensuite lister / concevoir les outils qui permettent d'atteindre ces objectifs de qualité, puis les classer en fonction de leur niveau de difficulté et/ou utilité. Les qualités de processus de rewarding correspondent bien sûr à la qualité du partage de données mis en œuvre.
Je pense qu'un processus de rewarding reconnu est l'étape qui précède sa transposition en dispositif réglementaire s'appuyant sur un standard. Je recommande donc de prévoir et de discuter toutes les étapes de développement nécessaires et réalisables pour atteindre l'établissement d'un processus de rewarding reconnu et utilisé.
La conception du premier niveau doit être faite de manière à ce qu'un néophyte ou un non spécialiste puisse le faire valider sans que celui-ci soit trop "coûteux" à obtenir. Le rapport coût-avantage de ce premier niveau doit être maximal (beaucoup d'avantages pour pas trop d'engagement). Puis la différence de difficulté entre les niveaux doit être graduelle, en développant en parallèle les niveaux d'avantages. La difficulté ici est de ne pas insérer de marches de progression trop grandes, en prenant en compte la diversité des acteurs et des moyens selon chaque type d'acteur. Cela correspond à construire une stratégie de progression basée sur des méthodes, des outils mais aussi une démarche itérative du type

« sensibilisation, formation, contrôle qualité, évaluation, rétribution, explication des enjeux de l'étape suivante ». Dans cette construction, il faudrait tenir compte des effets de potentialisation de certaines composantes du rewarding, et du gain possible à les coupler avec les bonnes stratégies de sollicitation pour l'étape suivante. Toutes ces étapes demandent des moyens humains dédiés et une "évangélisation" des communautés de producteurs de données qui sera différente selon leur degré de maturité dans le partage de la donnée.

Figure 54 : Ce message montre que des méthodes pédagogiques simples peuvent être percutantes pour certains types d'individus et redondantes. Pour certains, il faut imaginer d'autres méthodes comme les contrôles et les sanctions. Il ne reste plus qu'à imaginer l'équivalent pour les données.

L'information sur les enjeux doit être percutante, permanente, et répétée à chaque étape et dans chaque endroit pertinent, à l'image de cette plaque d'égout (permettez-moi cette métaphore figure 54) qui porte un message montrant clairement les enjeux sous-jacents que chacun devrait avoir sous les yeux pour comprendre l'impact généré par la confusion entre une poubelle et les égouts (qui, malgré ce que certains imaginent, ne sont pas forcément filtrés et peuvent véhiculer tout ce qui est jeté dedans [cotons tiges, filtres de cigarette, etc.] jusqu'à la mer).

Les étapes « évaluation, rétribution, explication des enjeux de l'étape suivante » doivent aussi tenir compte des freins possibles (pourquoi je ne partage pas mes données voir chapitre 3 section 1.1) typiques d'une communauté. Dans le cadre de CIGESMED[126], par exemple, les Turcs ont produits des données compatibles avec les données des autres pays participants, mais ils n'ont pas le droit de les diffuser librement. Dans ce cas, le verrou empêchant la mise en place de ces processus de rewarding est institutionnel. Concernant les verrous empêchant la mise en place de ces processus de rewarding des autres données de CIGESMED, notamment concernant les données de sciences participatives, le problème principal a été le peu de disponibilité des ingénieurs responsables du développement des accès aux données et a abouti en fin de programme à un développement incomplet des services d'accès aux données.

Concernant la mise en place de ces processus de partage des autres données de DEVOTES, il y a eu un véritable investissement dans les plateformes et la démarche de mise en ligne s'est appuyée sur un personnel dédié. Les jeux de données d'une grande partie des suivis ont été décrits et partagés entre membres du programme mais paradoxalement ne sont pas considérés comme un "output" (à part le catalogue des keystone species auquel nous avons contribué). Il en découle que peu de processus de remerciements ou de gratification ont été mis en place pour les plus volontaires (à part d'être co-auteurs dans certaines publications). Une illustration marquante de cette moindre importance de la donnée par rapport à la publication est que le menu "output" contient "Deliverables and Milestones", "Publications", "Software and tools", "Catalogues", "Calendar 2016"[127]. Aucun accès n'est évident malgré toutes les données produites (qui ont alimentées 135 publications (mai 2018)!).

Les données ont été partagées, parfois *via* des entrepôts liés à chacun des laboratoires ou *via* des data papers, mais là encore il semble que ce soit la moindre importance du rewarding autour du partage de la donnée par rapport à l'intérêt de la publication pour la carrière du chercheur qui soit à l'origine du manque de développement et de l'organisation des accès aux données.

2.6. Conclusions sur ces recommandations

Finalement, une des erreurs principales est de considérer le système d'observation et le système d'information comme deux entités différentes. Pour certains plus expérimentés, ceux-ci sont obligatoirement couplés. J'irais un peu plus loin en affirmant que le système

[126] Le site de CIGESMED permet un téléchargement des données prétraitées de France sous les formats ZIP, TXT, XML et JSON.
[127] http://www.devotes-project.eu

d'observation et les qualités intrinsèques des variables mesurées (évolutivité, comparabilité dans le temps et l'espace, interopérabilité des variables de contextes avec d'autres systèmes) sont des composantes à part entière du système d'information. Cette erreur de les considérer comme deux entités différentes conduit souvent à un échafaudage en deux temps dans la conception des relations entre système d'observation et système d'information (beaucoup de systèmes souffrent du "syndrome de la mise en boîte" : que veut-on mesurer – pour quoi faire ?, puis seulement dans un second temps, comment je les « met en boîte » ?). De plus, l'effet programme, particulièrement développé dans le domaine de la recherche, conduit à une seule itération de ces deux phases (la phase de conception est unique, et les moments et processus d'évaluation « intégratifs » sont *quasi* inexistants).
Dans le cadre de systèmes plus évolués, qui n'existent pas encore dans le domaine de la biodiversité, la conception de l'un et de l'autre doit se faire non seulement de manière simultanée mais aussi intégrée et itérative, jalonnée d'évaluations concernant non seulement le système d'observation et le système d'information mais aussi l'intégration du système d'observation dans les systèmes d'information préexistants dans des périmètres plus larges. Cela requiert de la part non seulement des architectes, mais aussi de la part de tous les producteurs et utilisateurs que l'observation soit considérée dès sa fabrication comme un bien commun et inaliénable (ce qu'il doit être facile de concevoir, surtout quand il s'agit d'argent public). Chaque usage répliquera l'information, et l'enjeu principal est de ne jamais en perdre la source, voire même seulement de la déconnecter de la source (c'est à dire d'être en capacité de la faire mettre à jour par son auteur ou son autorité). Cela implique que pour chaque auteur ou autorité disparaissant, il y ait une reprise de cette fonction de mise à jour par un organisme habilité, spécialisé dans le type de données produites (à l'image des "couches de références" produites par le M.N.H.N.).

L'émergence des "Data Management Plan" est une avancée significative dans le domaine du partage de la donnée. Un groupe de travail de R.D.A. travaille même sur un D.M.P accessible aux ordinateurs (Simms *et al.*, 2017) (Interprétables car très formalisés). Néanmoins, chez la plupart des producteurs de données, leur développement n'est pas encore accompagné, et chaque producteur de données peut mettre en oeuvre un "Data Management Plan" qui n'en a pas les qualités, et que personne ne contrôlera. A l'image de la récente application du R.G.D.P.[128] (Règlement Général sur les Données Personnelles, entré en vigueur le 25 Mai 2018), il est aujourd'hui nécessaire de créer un cadre juridique unifié pour l'ensemble de l'Union Européenne pour les "Data Management Plan" dans le

[128] https://www.cnil.fr/fr/reglement-europeen-sur-la-protection-des-donnees-ce-qui-change-pour-les-professionnels

domaine environnemental. Ce “Règlement Général sur les Données Environnementales” pourrait reposer sur des principes similaires au R.G.D.P. à savoir imposer une portabilité des données d’un système à un autre, responsabiliser les acteurs traitant des données (responsables de traitement et sous-traitants, c’est à dire tous les usagers) et crédibiliser la régulation grâce à une coopération renforcée entre les autorités de gestion des données (ce qui implique des évaluations, rétributions en sanctions en fonction de la qualité des D.M.P. mis en place).

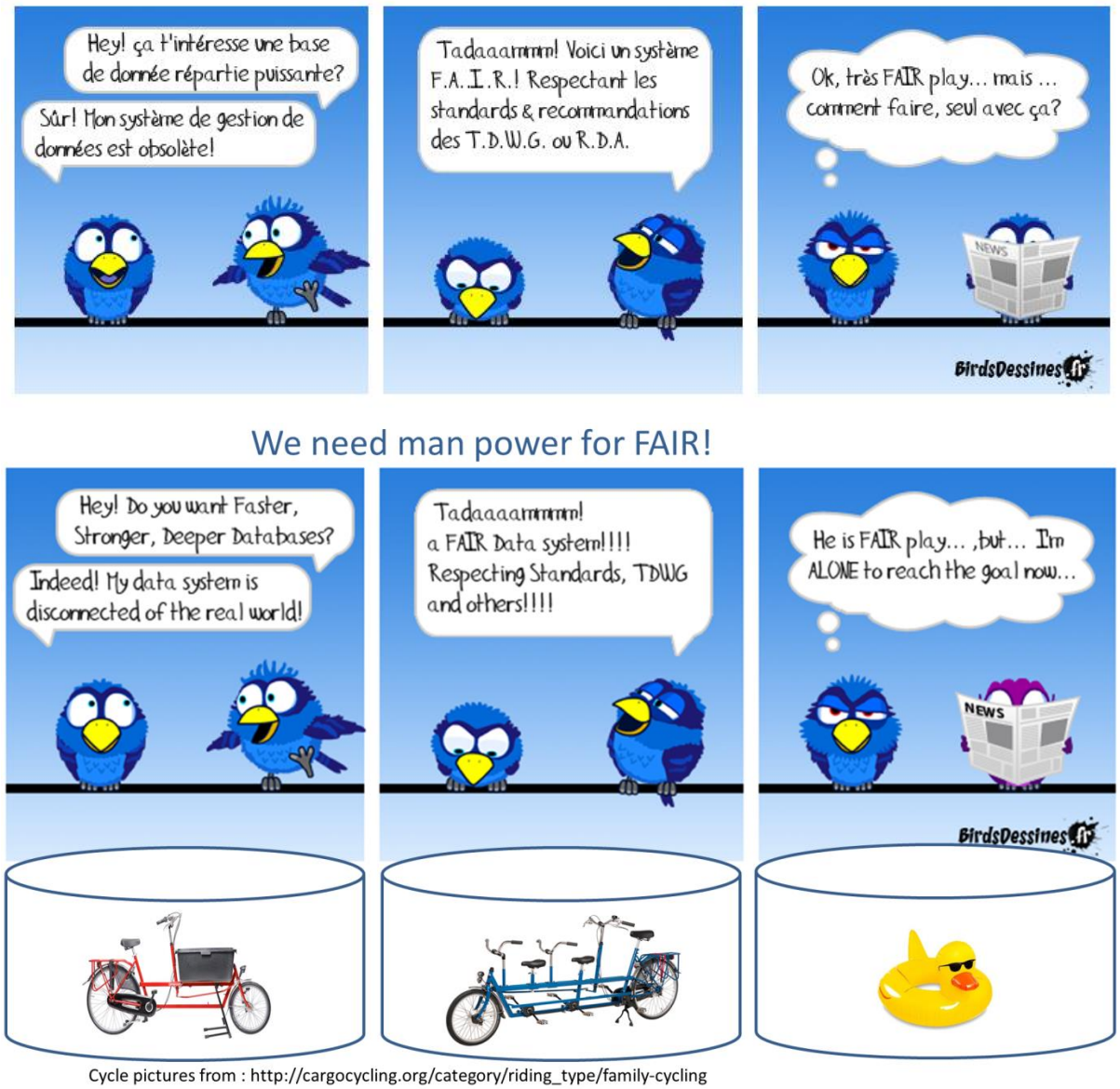


Figure 55 : quels moyens humains pour l’application des principes FAIR (FAIR, donc a minima bilingue…) ?

L'émergence d'outils toujours plus évolués pour gérer et traiter les données ne doit pas faire oublier que cette gestion sera forcément coûteuse en moyens, notamment humains (Figure 55).

Ces questions de coût sont centrales, mais est-ce une question que l'on se pose lorsque l'on construit un grand équipement comme une tour, une autoroute, un rond-point ou un avion ? A l'heure où l'on parle de l'automatisation et du remplacement de l'humain par les robots, l'utilisation "transdisciplinaire[129]" de la donnée est en mesure de faire émerger de nouvelles activités et de nouveaux métiers, à condition de ne négliger ni les investissements utiles, ni l'institution de nouvelles règles autour de la donnée. A ce titre, le principe de propriété de la donnée environnementale n'est ni éthique, ni productif, il empêche toute considération de l'enjeu stratégique d'un partage imposé (et organisé).
Pour autant, l'application de l'*open data* est une condition nécessaire mais pas suffisante au développement de l'utilisation "transdisciplinaire" de la donnée : même si un producteur met tout en œuvre pour rendre des données "FAIR", seront-elles vraiment prises en compte et utilisées ? Contrairement aux enjeux liés aux changements climatiques, la crise actuelle de biodiversité ne touche pas le grand public. La couverture médiatique du changement climatique est jusqu'à huit fois supérieure à celle de la perte de biodiversité, et ce, malgré une faible différence dans la production de littérature scientifique et le financement de la recherche (Legagneux et al., 2018). J'interprète cette différence par notre incapacité dans les différents domaines de l'écologie, *a contrario* des climatologues, par exemple, à construire des modèles réalistes et intégratifs permettant de tester des scénarios et d'évaluer les impacts des différentes modifications de notre environnement sur la biodiversité et le coût réel qui en résulte pour l'humanité. J'espère que ce travail contribuera au futur développement de ces scénarios, à leur diffusion et à leur compréhension par le plus grand nombre ; ceci nécessitera le développement d'une recherche opérationnelle productive, nécessairement sur le long terme, capitalisant sur le développement et le maintien des compétences. Ce projet de recherche sur l'aide à la décision dans le domaine environnemental et d'alimentation et de développement d'I.D.S.S. (qui doivent nécessairement être contrôlés par des humains), à partir de motifs de contextes issus de graphes hétérogènes et multi-sources nécessite un travail interdisciplinaire dans toute sa durée et doit s'appuyer sur des ingénieurs ayant un niveau suffisant de compétences simultanément en S.T.I.C. et en sciences environnementales.

[129] Le mot « transdisciplinaire » est souvent défini comme un synonyme de « interdisciplinaire » (Qui traverse les frontières entre les disciplines). La nuance proposée par certains et à laquelle je souscris est que c'est à cette frontière que se développent de nouvelles disciplines, on peut alors parler non plus d'interdisciplinarité mais de transdisciplinarité.

Références bibliographiques

- Aggarwal C., Wang H., (2010) - Managing and Mining Graph Data, ***Springer***, 1st Edition., 2010, XXII, 600 pp.
- Altman S., Whitlach R.B., (2007) - Effect of small-scale disturbance on invasion success in marine communities. ***Journal of Experimental Marine Biology and Ecology*** 342: 15-29.
- Amanqui F.K., Serique K.J., Cardoso S.D., dos Santos J.L., Albuquerque A. and Moreira D.A., (2014) - Improving Biodiversity Data Retrieval through Semantic Search and Ontologies, in ***Web Intelligence*** (WI) and ***Intelligent Agent Technologies*** (IAT), 2014 ***IEEE/WIC/ACM International Conference on Web Intelligence***, 11-14 August 2014, Warsaw, Poland, vol.1 (WI), 274-281, doi: 10.1109/WI-IAT.2014.44.
- Amorim D.S., Santos C.M.D., Krell F.-T., Dubois A., Nihei S.S., Oliveira O.M.P., Pont A., Song H., Verdade V.K., Fachin D.A., Klassa B., Lamas C.J.E., Oliveira S.S., Carvalho C.J.B., De Mello-Patiu C.A., Hajdu E., Couri M.S., Silva V.C., Capellari R.S., Falaschi R.L., Feitosa R.M., Prendini L., Pombal J.P.J., Fernández F., Rocha R.M., Lattke J.E., Caramaschi U., Duarte M., Marques A.C., Reis R.E., Kurina O., Takiya D.M., Tavares M., Fernandes D.S., Franco F.L., Cuezzo F., Paulson D., Guénard B., Schlick-Steiner B.C., Arthofer W., Steiner F.M.,Fisher B.L., Johnson R.A., Delsinne T.D., Donoso D.A., Mulieri P.R., Patitucci L.D., Carpenter J.M., Herman L. & Grimaldi D., (2016) - Timeless standards for species delimitation. ***Zootaxa***, 4137 (1), 121–128.http://dx.doi.org/10.11646/zootaxa.4137.1.9.
- Armstrong M., Murlis H., (2004) - Reward management: a handbook of remuneration strategy and practice (5th ed.). London [u.a.]: ***Kogan Page***. 704 pp. ISBN 978-0749439842.
- Athanasiadis A., (1999) - The taxonomic status of Lithophyllum stictaeforme (Rhodophyta, Corallinales) and its generic position in light of phylogenetic considerations. ***Nordic Journal of Botany***, 19(6), 735–746.
- Auber D., Archambault D., Bourqui R., Delest M., Dubois J., Pinaud B., Lambert A., Mary P., Mathiaut M., Melançon G., (2014) - Tulip III, ***Encyclopedia of Social Network Analysis and Mining***, 2216-2240, doi: 10.1007/978-1-4614-6170-8_315. <hal-01096759>
- Ballesteros E., (2006) - Mediterranean coralligenous assemblages: a syntheses of present knowledge, ***Oceanogr. Mar. Biol.***, Annu. Rev. 44, 123 – 195.
- Bianchi C. N., Cattaneo-Vietti R., Morri C., Navone A., Panzalis P., Orru P., (2007) - Coralligenous formations in the Marine Protected Area of Tavolara Punta Coda Cavallo (N-E Sardinia, Italy). ***Biologia marina mediterranea***, 14(2), 148-149.

- Borja Á., Franco J., Pérez V., (2000) - A marine biotic index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. ***Marine Pollution Bulletin*** 40: 1100-1114.
- Borja A., Ranasinghe A., Weisberg S., (2009) - Assessing ecological integrity in marine waters, using multiple indices and ecosystem components: challenges for the future. ***Marine Pollution Bulletin*** 59, 1–4.
- Borja, Á., Marín S.L., Muxika I., Pino L., Rodríguez J.G., (2015) - Is there a possibility of ranking benthic quality assessment indices to select the most responsive to different human pressures? ***Marine Pollution Bulletin***, 97: 85-94.
- Borja Á., Elliott M., Andersen J. H., Berg T., Carstensen J., Halpern B.S. , Heiskanen A.S., Korpinen S., Lowndes J.S.S., Martin G., Rodriguez-Ezpeleta N., (2016) - Overview of integrative assessment of marine systems: the Ecosystem Approach in practice. ***Frontiers in Marine Science***, 3: doi: 10.3389/fmars.2016.00020.
- Borja Á., (2017) - Final Report Summary - DEVOTES (DEVelopment Of innovative Tools for understanding marine biodiversity and assessing good Environmental Status), ***Cordis***, 1-11. https://cordis.europa.eu/result/rcn/195921_en.html
- Boudouresque, C.F., (2004) - Marine biodiversity in the Mediterranean: status of species, populations and communities. ***Scientific reports of Port-Cros national park,*** 20, 97–146.
- Boudouresque C.F., (1971) - Méthodes d'étude qualitative et quantitative du benthos (en particulier du phytobenthos). ***Téthys***, 3, 79-104.
- Bourcier M., (1988) - Macrobenthos de substrat meuble circalittoral autour de l'île de Port-Cros (Méditerranée, France). ***Scientific reports of Port-Cros national park***, 14: 41-63
- Bowden D.A., Clarke A., Peck L.S., Barnes D.K.A., (2006) - Antarctic sessile marine benthos: colonisation and growth on artificial substrata over three years. ***Marine Ecology Progress Series*** 316: 1-16.
- Brainard R.E., Asher J., Blyth-Skyrme V., Coccagna E.F., Dennis K., Donovan M.K., Gove J.M., Kenyon J., Looney E.E., Miller J.E., Timmers M.A., Vargas-Angel B., Vroom P.S., Vetter O., Zgliczynski B., Acoba T., DesRochers A., Dunlap M.J., Franklin E.C., Fisher-Pool P.I., Braun C.L., Richards B.L., Schopmeyer S.A., Schroeder R.E., Toperoff A., Weijerman M., Williams I., Withall R.D., (2012) - Coral reef ecosystem monitoring report of the Mariana Archipelago: a 2003 – 2007 Pacific Islands Fisheries Science Center, ***PIFSC Special Publication***, SP-12-01, 1019 pp.
- Bray J.R. and J.T. Curtis. 1957. An ordination of upland forest communities of southern Wisconsin. ***Ecological Monographs*** 27:325-349.

- Cardoso S.D., Amanqui, K.J.A. Serique, dos Santos J.L.C., Moreira D.A., (2016) - SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data. ***Future Generation Computer Systems***, vol. 54, January 2016, 389-398, doi: 10.1016/j.future.2015. 05.006.
- Carugati L., Corinaldesi C., Dell'Anno A., Danovaro R., (2015) - Metagenetic tools for the census of marine meiofaunal biodiversity: An overview. ***Marine Genomics***, 2015;24: 11–20. doi:10.1016/j.margen.2015.04.010.
- Casas-Güell E., Cebrian E., Garrabou J., Ledoux J.B., Linares C., Teixidó N., (2016) - Structure and biodiversity of coralligenous assemblages dominated by the precious red coral Corallium rubrum over broad spatial scales. ***Scientific Report*** 6, 36535; doi: 10.1038/srep36535.
- Casoli E., Nicoletti L., Mastrantonio G., Jona-Lasinio G., Belluscio A., Ardizzone G.D., (2017) - Scuba diving damage on coralligenous builders: Bryozoan species as an indicator of stress, ***Ecological Indicators***, Volume 74, March 2017, 441-450, ISSN 1470-160X, http://dx.doi.org/10.1016/j.ecolind.2016.12.005.
- Cecchi E., Piazzi, (2010) - A new method for the assessment of the ecological status of coralligenous assemblages. ***41° Congresso della Società Italiana di Biologia Marina***, Rapallo (GE), 7-11 giugno 2010.
- Ceríaco L., Gutiérrez E., Dubois A., (2016) - Photography-based taxonomy is inadequate, unnecessary, and potentially harmful for biological sciences. ***Zootaxa***, 4196(3), 435–445. doi:http://dx.doi.org/10.11646/zootaxa.4196.3.9.
- Cianferoni F., Bartolozzi L., (2016) - Warning: potential problems for taxonomy on the horizon? ***Zootaxa***, 4139 (1), 128–130. http://dx.doi.org/10.11646/zootaxa.4139.1.8.
- Clarke K.R., Gorley R.N., Somerfield P.J., Warwick R.M. (2014) - Change in marine communities: an approach to statistical analysis and interpretation, 3rd edition. ***PRIMER-E, Plymouth***, 260 pp.
- Clarke K.R., Gorley R.N. (2015) - PRIMER v7: User Manual/Tutorial. ***PRIMER-E, Plymouth***, 296 pp.
- Claudet J., Pelletier D., (2004) - Marine protected areas and artificial reefs : A review of the interactions between management and scientific studies. ***Aquatic Living Resources***, 17(2), 129-138. Publisher's official version : http://doi.org/10.1051/alr:2004017 , Open Access version : http://archimer.ifremer.fr/doc/00000/397/
- Claudet J., Pelletier D., Jouvenel J.Y., Bachet F., Galzin R., (2006) - Assessing the effects of marine protected area (MPA) on a reef fish assemblage in a northwestern Mediterranean marine reserve: identifying community-based indicators. ***Biological Conservation*** 130,349–369.

- Claudet J., Fraschetti S. (2010) - Human-driven impacts on marine habitats: a regional meta-analysis in the Mediterranean Sea. ***Biological Conservation***, 143 (9): 2195–2206. DOI: 10.1016/j.biocon.2010.06.004.
- Coll M., Piroddi C., Albouy C., Ben Rais Lasram F., Cheung W.W.L., Christensen V., Karpouzi V.S., Guilhaumon F., Mouillot D., Paleczny M., Lourdes Palomares M., Steenbeek J., Trujillo P., Watson R., Pauly D., (2011) - The Mediterranean Sea under siege : spatial overlap between marine biodiversity, cumulative threats and marine reserve. ***Global Ecology and Biogeography***, 1-16. ISSN 1466-8238.
- Coll M., Piroddi C., Steenbeek J., Kaschner K., Ben Rais Lasram F., Aguzzi J., Ballesteros E., Bianchi C.N., Corbera J., Dailianis T., Danovaro R., Estrada M., Froglia C., Galil B.S., Gasol J.M., Gertwagen R., Gil J., Guilhaumon F., Kesner-Reyes K., Kitsos M.-S., Koukouras A., Lampadariou N., Laxamana E., López-Fé de la Cuadra C.M., Lotze H.K., Martin D., Mouillot D., Oro D., Raicevich S., Rius-Barile J., Saiz-Salinas J.I., San Vicente C., Somot S., Templado J., Turon X., Vafidis D., Villanueva R., Voultsiadou E. (2010) - The Biodiversity of the Mediterranean Sea: Estimates, Patterns, and Threats. ***PLoS One***, 5, e11842.doi:10.1371/journal.pone.0011842.
- Conruyt N., Sébastien D., Vignes-Lebbe R., Cosadia S., (2010) - Moving from biodiversity information systems to biodiversity information services. ***Information and Communication Technologies for Biodiversity Conservation and Agriculture***, 107-128.
- Cortez P., Embrechts M.J., (2013) - Using sensitivity analysis and visualization techniques to open black box data mining models. ***Information Sciences***, 225, 1-17.
- Courteau R., (2011) - La pollution de la Méditerranée : état et perspectives à l'horizon 2030 », ***Rapport du Sénat n° 652***, 188 pp.
- David R., Arvanitidis C., Çinar M.E., Sartoretto S., Doğan A., Dubois S., Erga Z., Guillemain D., Thierry de Ville d'Avray L., Zuberer F., Chenuil A., Féral J.P., (2014a) - CIGESMED habitat's characterization : a simple and reusable typology at the Mediterranean scale. ***RAC/SPA 2nd Mediterranean Symp. on the Conservation of coralligenous and other calcareous bioconcretions***, Portorož (SI), 29-30/10/2014, 211-212.
- David R., Arvanitidis C., Çinar M.E., Dubar J., Dubois S., Erga Z., Guillemain D., Sartoretto S., Thierry de Ville d'Avray L., Zuberer F., Chenuil A., Féral J.-P., (2014b), with contributors : Açik Çinar S., Andral B., Aurelle D., Aysel V., Bakir K., Bellan G., Bellan-Santini D., Bouchoucha M., Celik C., Chatzigeorgiou G., Chatzinikolaou E., Chenesseau S., Dağli E., Dailianis T., Dimitriadis C., D'Iribarne C., Doğan A., Dounas C., Egea E., Elguerrabi W., Emery E., Evcen A., Faulwetter S., Gatti G., Gerovasileiou V., Güçver S.M., Issaris Y., Katağan T., Keklikoglou K., Kirkim F., Koçak F.,

Koutsoubas D., Marschal C., Önen M., Önen S., Öztürk B., Panayiotidis P., Pavloudi C., Pergent G., Pergent-Martini C., Poursanidis D., Ravel C., Reizopoulou S., Rocher C., Ruiton S., Sakher S., Salomidi M., Sarropoulou E., Selva M., Sini M., Sourbes L., Simboura N., Taşkin E., Vacelet J., Valavanis V., Vasileiadou A., Verlaque M., (2014) - Protocols for monitoring of coralligenous habitats of mediterranean (Coralligenous based Indicators to Evaluate and Monitor the "good ecological status" of the MEDiterranean coastal waters) Protocoles de suivi du coralligène en méditerranée (Coralligenous based Indicators to Evaluate and Monitor the "good ecological status" of the MEDiterranean coastal waters).

- David R., Arvanitidis C., Çinar M.E., Sartoretto S., Dogan A., Dubois S., Erga Z., Guillemain D., Thierry de Ville d'avray L., Zuberer F., Chenuil A., Féral J.P., (2014c) - CIGESMED Protocols : how to implement a multidisciplinary approach on a large scale for coralligenous habitats surveys. ***RAC/SPA 2nd Mediterranean Symp. on the Conservation of coralligenous and other calcareous bioconcretions***, Portorož (SI), 29-30/10/2014, 66-71.
- David R., Dubois S., Erga Z., Guillemain D., Thierry de Ville d'Avray L., Arvanitidis C., Çinar M., Sartoretto S., Zuberer F., Chenuil A., Féral J.P., (2014d) - CIGESMED's protocol and network (Coralligenous basEd. Indicators to evaluate and monitor the "Good Environmental Status" of MEDiterranean coastal waters). Proc. ***5th Intn. Symp. Monitoring of Mediterranean coastal areas : problems and measurement techniques***, Livorno (Italy) 17-18-19 June 2014, F. Benincasa (Ed.), pp. 828-843 ; CNR-IBIMET : Florence (IT), ISBN 978-88-95597-19-5.
- David R., Féral J.P., Gachet S., Dias A., Blanpain C., Lecubin J., Diaconu C., Surace C., Gibert K., (2015) - A first prototype for indexing, visualizing and mining heterogeneous data in Mediterranean ecology within the IndexMed consortium interdisciplinary framework. In: SITIS 2015, ***11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)***, 23-27 nov. 2015, Bangkok, Thailand, 232-239, doi: 10.1109/SITIS.2015.119.
- David R., Féral J.P., Archambeau A.S., Bailly N., Blanpain C., Breton V., De Jode A., Delavaud A., Dias A., Gachet S., Guillemain D., Lecubin J.,Romier G., Surace C., Thierry de Ville d'Avray L., Arvanitidis C., Chenuil A., Çinar M.E., Koutsoubas D., Sartoretto S., Tatoni T., (2016a) - IndexMed projects : new tools using the CIGESMED DataBase on Coralligenous for indexing, visualizing and data mining based on graphs. In : Sauvage S., Sánchez-Pérez J.M., Rizzoli A.E., (Eds.), ***Proceedings of the 8th International Congress on Environmental Modelling and Software, Environmental modelling and software for supporting a sustainable future***, Vol. 3, 656-665, July 10-14, Toulouse, France. ISBN : 978-88-9035-745-9.

- David R., Féral J.P., Tatoni T., (2016b) - Architecture, concepts et services d'un système d'indexation de données distribuées pour l'observation à large échelle en écologie marine dans le cadre du consortium IndexMEED (Indexing for Mining Ecological and Environmental Data). ***36ème conférence sur la Gestion de Données - Principes, Technologies et Applications (BDA 2016)***, 15-18 novembre 2016, Poitiers.
- David R., Bernard L., Blanpain C., Dias A., Féral J.P., Gachet S., Lecubin J., Surace C., Tatoni T., (2017) - Visualisation de données sous forme de graphes en archéologie : rencontre opérationnelle des archéologues d'ArkeoGIS et des écologues d'IndexMed, in Costa, L., Giligny, F., Djindjian, F. (Eds.), *Actes des 5èmes Journées d'informatique et archéologie de Paris, JIAP 2016*, Paris, 7-10 juin 2016, ***Archéologies numériques***, vol.1 (https://www.openscience.fr/Archeologies-numeriques).
- David R., Féral J.-P., Blanpain C., (in press, 2018) - Ecological Data Preservation in the context of IndexMed, In: C. Diaconu. (Ed) PREDON.
- David R., Uyarra M.C., Carvalho S., Anlauf H., Borja A., Cahill A.E., Carugati L., Danovaro R., De Jode A., Féral J.-P., Guillemain D., Lo Martire M., Thierry de Ville d'Avray L., Pearman J.K., Chenuil A. (2018) - *submitted* - annexe 1 - Photo analyses of Autonomous Reef Monitoring Structures, as tools to detect geographical, spatial, and environmental effects: lessons from a pilot study. ***Marine Pollution Bulletin***
- De Jode A., (2018) - Etude de la biodiversité des habitats coralligènes et de l'influence des facteurs environnementaux par des approches génétiques : des populations d'espèces ingénieures aux communautés, ***Thèse 3ème cycle Ecologie Marine***, Université Aix-Marseille.
- De Jode A., David R., Haguenauer A., Cahill A.E., Erga Z., Guillemain D., Sartoretto S., Rocher C., Selva M., Le Gall L., Féral J.-P., Chenuil A. (2018 *submitted*) - Multiple cryptic species, spatial and ecological differentiation in a major builder of coralligenous habitats. ***Molecular Ecology.***
- Deter J., Descamp P., Ballesta L., Boissery P., Holon F., (2010) - A preliminary study toward an index based on coralligenous assemblages for the ecological status assessment of Mediterranean French coastal waters. ***Ecological Indicators***, 20: 345-352.3.
- Deter J., Descamp P., Ballesta L., Boissery P., & Holon F., (2012a) - A preliminary study toward an index based on coralligenous assemblages for the ecological status assessment of Mediterranean French coastal waters. ***Ecological indicators***, 20, 345-352.

- Deter J., Descamp P., Boissery P., Ballesta L., Holon F., (2012b) - A rapid photographic method detects depth gradient in coralligenous assemblages. ***Journal of Experimental Marine Biology and Ecology***, 418, 75-82.
- Donegan T.M., (2008) - New species and subspecies descriptions do not and should not always require a dead type specimen. ***Zootaxa***, 1761, 37–48.
- Donegan T.M., (2009) - Type specimens, samples of live individuals and the Galapagos Pink Land Iguana. ***Zootaxa***, 2021, 12–20.
- Dubois A., Nemésio A., (2007) - Does nomenclatural availability of nomina of new species or subspecies require the deposition of vouchers in collections? ***Zootaxa***, 1409, 1–22.
- Dubois A., (2009) - Endangered species and endangered knowledge. ***Zootaxa***, 2201, 26–29.
- El Guerrabi W. (2014) - Qualification de données de CIGESMED : Un système de prospection de données pour l'aide à la gestion des habitats coralligènes. *Rapport de stage d'école d'ingénieur.* 47 pp. + annexes.
- Ellison A. M., (2010) - Partitioning diversity. ***Ecology***, 91: 1962–1963. doi:10.1890/09-1692.1.
- Farré M., Lombarte A., Recasens L., Maynou F., Tuset V.M., (2015) - Habitat influence in the morphological diversity of coastal fish assemblages, ***Journal of Sea Research*** http://dx.doi.org/10.1016/j.seares.2015.03.002.
- Fayyad U., Piatetsky-Shapiro G., & Smyt P., (1996) - From data mining to knowledge discovery in databases. ***AI magazine***, 17(3), 37.
- Féral J.P, David R., (2013) - Zone côtière et développement durable, une équation à résoudre. In : ***Le développement durable à découvert L'environnement, un système global dynamique***, Euzen A., Eymard L., Gaill F. (Eds.) ***CNRS éditions***: Paris, September, 96-97, ISBN : 978-2-271-07896-4.
- Féral J.P., Arvanitidis C., Chenuil A., Çinar M.E., David R., Frémaux A., Koutsoubas D., Sartoretto S., (2014) - CIGESMED, Coralligenous based Indicators to evaluate and monitor the « Good Environmental Status » of the MEDiterranean coastal waters, a SeasEra project (www.cigesmed.eu). ***Proceedings RAC/SPA 2nd Mediterranean Symposium on the Conservation of coralligenous and other calcareous bio-concretions***, Portorož, Slovenia, October, 15-21.
- Féral J.-P., Arvanitidis C., Chenuil A., Çinar M.E., David R., Egea E., Sartoretto S. (2016) - ***CIGESMED : Coralligenous based Indicators to evaluate and monitor the « Good Environmental Status » of the Mediterranean coastal waters***. SeasEra project Towards Integrated Marine Research Strategy and Programmes. Final report, 179 pp. DOI: 10.13140/RG.2.2.35960.03848

- Forestier G., Wemmert C., P. Gançarski, (2008) - Multi-source Images Analysis Using Collaborative Clustering". ***EURASIP Journal on Advances in Signal Processing, Special issue on Machine Learning in Image Processing***, (2008) - Article ID 374095, 11 pp., doi:10.1155/2008/374095.
- Fredj G., Bellan-Santini D. & Meinard M., (1992) - Etat des connaissances sur la faune marine méditerranéenne. ***Bulletin de l'Institut Océanographique, Monaco***, no special 9:133-145.
- Gachet S., Véla E., Tatoni T., (2005) - BASECO: a floristic and ecological database of Mediterranean French flora. ***Biodiversity & Conservation***, 14(4), 1023-1034.
- Gasparini S., (2007) - PLANKTON IDENTIFIER : a software for automatic recognition of planktonic organisms. http://wwwobs-vlfrfr/~gaspari/Plankton_Identifier/indexphp.
- Gatti G., Bianchi C.N., Morri C., Montefalcone M., Sartoretto S., (2015) - Coralligenous reefs state along anthropized coasts: Application and validation of the COARSE index, based on a rapid visual assessment (RVA) approach. ***Ecological Indicators***, 52, 567-576.
- Gentile G., Snell H., (2009) - *Conolophus marthae* sp. nov. (Squamata: Iguanidae), a new species of land iguana from the Galápagos archipelago. ***Zootaxa***, 2201, 1–10.
- Gibert K., Conti D., Vrecko D., (2012) - Assisting the end-user in the interpretation of profiles for decision support. an application to wastewater treatment plants. ***Environmental Engineering and Management Journal***, 11(5), 931-944.
- Gibert K., Valls A., Batet M., (2014) - Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering. ***Knowledge and Information Systems***, 40(3), 559-593.
- Gili J.M., Ros J., (1985) - Study and cartography of the benthic communities of MedesIslands (NE Spain). PSZN I: ***Marine Ecology*** 6, 219–238.
- Gimenez O., Buckland S.T., Morgan B.J., Bez N., Bertrand S., Choquet R., Dray S., Etienne M.P., Fewster R., Gosselin F., Mérigot B., Monestiez P., Morales J.M., Mortier F., Munoz F., Ovaskainen O., Pavoine S., Pradel R., Schurr F.M., Thomas L., Thuiller W., Trenkel V., de Valpine P., Rexstad E., (2014) - Statistical ecology comes of age. ***Biology Letters*** 10: 20140698.
- Glasby T.M., (2000) - Surface composition and orientation interact to affect subtidal epibiota. ***Journal of Experimental Marine Biology and Ecology***, 248,177–190.
- González-Goñi L., Borja A., Uyarra M.C. Comparación de herramientas de análisis de imagen: eficiencia y uso en ecología bentónica de sustrato duro. ***Revista de Investigación Marina*** 2017;24: 1–12.

- Gorsky G., Ohman M.D., Picheral M., Gasparini S., Stemmann L., Romagnan J.B., Cawood A., Pesant S., Garcia-Comas C., Prejger F,. (2010) - Digital zooplankton image analysis using the ZooScan integrated system. ***Journal of Plankton Research*** March;32(3):285–303. WOS:000274339900003.
- Hajj-Hassan H., (2016) - Les bases de données environnementales : entre complexité et simplification : Mutualisation et intégration d'outils partagés et adaptés à l'observatoire O-LiFE.
- Harmelin-Vivien M.L., Harmelin. J.G., Chauvet C., Duval C.,. Galzin R., Lejeune P., Barnabé. G., Blanc F., Chevalier R.,. Duclerc J., Lasserre G., (1985) - Evaluation visuelle des peuplements et populations de poissons: méthodes et problèmes. ***Revue d'Écologie (Terre et Vie)*** 40: 467–539.
- Hayes K.R., Cannon R., Neil K., Inglis G. (2005) -. Sensitivity and cost considerations for the detection and eradication of marine pests in ports. ***Marine Pollution Bulletin*** 50: 823-834. doi:10.1016/j.marpolbul.2005.02.032.
- Helmer A., Song B., Ludwig W., Schulze M., Eichelberg M., Hein A., Tegtbur U., Kayser R., Haux R. and Marschollek M., (2006) - A sensor-enhanced health information system to support automatically controlled exercise training of COPD patients. In: ***4th International Conference on Pervasive Computing Technologies for Healthcare. Munich: IEEE***, 22-25 March 2010, 1-6, doi: 10.4108/ CSTPERVASIVEHEALTH2010.8827.
- Hodgson G., (2000) - Coral Reef Monitoring and Management Using Reef Check. Integrated Coastal Zone Management. 1: 169 - 179.
- Hong, J.S., (1980) - Etude faunistique d'un fond de concrétionnement de type coralligène soumis à un gradient de pollution en Méditerranée nord-occidentale (Golfe de Fos)., ***Thèse 3ème cycle Océanologie***, Université Aix-Marseille II, 137 pp.
- Hughes B., Burghardt T., (2016) - Automated Visual Fin Identification of Individual Great White Sharks eprint arXiv:1609.06323, Computer Science - ***Computer Vision and Pattern Recognition***, 17 p, 2016arXiv160906323H
- Humphries A.T., La Peyre M.K., Kimball M.E., Rozas L.P., (2011) - Testing the effect of habitat structure and complexity on nekton assemblages using experimental oyster reefs. ***Journal of Experimental Marine Biology and Ecology*** 409, 172–179.
- Judge M.L., Craig S.F., (1997) - Positive flow dependence in the initial colonization of a fouling community: results from in situ water current manipulations. ***Journal of Experimental Marine Biology and Ecology*** 210: 209-222.
- Kohler K.E., Gill S.M., (2006) - Coral Point Count with Excel extensions (CPCe): a visual basic program for the determination of coral and substrate coverage using random point count methodology. ***Computers & Geosciences***. 32, 1259–1269.

- Kipson S., Linares C., Čižmek H. ,Cebrián E., Ballesteros E., Bakran-Petricioli T. and Garrabou J., (2015) - Population structure and conservation status of the red gorgonian Paramuricea clavata (Risso, 1826) in the Eastern Adriatic Sea. ***Marine Ecology***, 36: 982-993. doi:10.1111/maec.12195
- Kipson S., Fourt M., Teixidó N., Cebrian E., Casas E., Ballesteros E., Zabala M., Garrabou J., (2011) - Rapid biodiversity assessment and monitoring method for highly diverse benthic communities: a case study of Mediterranean coralligenous outcrops. ***PLoS One***, 6(11), e27103.
- Kissling W.D., Hardisty A., García E.A., Santamaria M., De Leo F., Pesole G., Freyhof J., Wissel S., Konijn J., Los W., (2015) - Towards global interoperability for supporting biodiversity research on essential biodiversity variables (EBVs). ***Biodiversity***, 16(2-3), 99-107.
- Knowlton N., Brainard R.E., Fisher R., Moews M., Plaisance L., Caley M.J., (2010) - Coral reef biodiversity. In Life in the World's Oceans: Diversity, Distribution, and Abundance, McIntyre A., (ed). ***Oxford: Wiley-Blackwell***, 65-77.
- Koch S., Hagglund M., (2009) - Health informatics and the delivery of care to older people. ***Maturitas***, vol. 63(3), July 2009, 195-199.
- Kremen C., (1992) - Assessing the Indicator Properties of Species Assemblages for Natural Areas Monitoring, ***Ecological Applications***, vol.2(2), 203-217.
- Kulbicki M., Cornuet N., Vigliola L., Wantiez L., Moutham G., Chabanet P., (2010) - Counting coral reef fishes: Interaction between fish life-history traits and transect design. ***Journal of Experimental Marine Biology and Ecology*** 387: 15–23.
- Laborel J., (1987) - Marine biogenic constructions in the Mediterranean. A review. ***Scientific Reports of Port-Cros national park***, 13, 97-126.
- Laborel J., (1961) - Le concrétionnement algal "coralligène" et son importance géomorphologique en Méditerranée. ***Recueil des Travaux de la Station Marine d'Endoume***. 23, 37-60.
- Lambert A., Bourqui R., Auber D., (2013) - Graph Visualization For Geography, in Rozenblat C., Melançon G. (eds), Methods for Multilevel Analysis and Visualisation of Geographical Networks. ***Methodos Series (Methodological Prospects in the Social Sciences)***, vol 11. Springer, Dordrecht, doi :10.1007/978-94-007-6677-8_6.
- Lagadeuc Y., Chenorkian R., (2009) - Les systèmes socio-écologiques : vers une approche spatiale et temporelle. ***Natures Sciences Sociétés***, vol. 17,(2), 194-196. (http://www.cairn.info/revue-natures-sciences-societes-2009-2-page-194.htm).
- Laubier L. (1966) - Le coralligène des Albères. Monographie Biocénotique. ***Annales de l'Institut Océanographique***, Paris.

- Laporte M.A., Garnier E., (2012) - ThesauForm—Traits: A web based collaborative tool to develop a thesaurus for plant functional diversity research. ***Ecological Informatics***, 11, 34-44.
- Laporte M.A., Mougenot I., Garnier E., Stahl U., Maicher L., Kattge J., (2014) - A semantic web faceted search system for facilitating building of biodiversity and ecosystems services. In ***Data Integration in the Life Sciences,*** Springer International Publishing, 50-57.
- Legagneux P., Casajus N., Cazelles K., Chevallier C., Chevrinais M., Guéry L., Jacquet C., Jaffré M., Naud M.J., Noisette F., Ropars P., Vissault S., Archambault P., Bêty J., Berteaux D., Gravel D., (2018) - Our house is burning: discrepancy in climate change vs biodiversity coverage in the media as compared to scientific literature. ***Frontiers in Ecology and Evolution*** https://doi.org/10.3389/fevo.2017.00175.
- Leray M., Knowlton N., (2015) - DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. ***Proceeding of the National Academy of Sciences*** USA, 112 (7): 2076–2081, doi: 10.1073/pnas.1424997112
- Linares C., Coma R., Garrabou J., Díaz D. and Zabala M. (2008) - Size distribution, density and disturbance in two Mediterranean gorgonians: Paramuricea clavata and Eunicella singularis. ***Journal of Applied Ecology***, 45: 688-699. doi:10.1111/j.1365-2664.2007.01419.x
- Little J.D., (1970) - Models and Managers: The Concept of a Decision Calculus. ***Management Science***, vol. 16(8), B-466-B485.
- Liu J., Dietz T., Carpenter S.R., Alberti M., Folke C., Moran E., Pell A.N., Deadman P., Kratz T., Lubchenco J., Ostrom E., Ouyang Z., Provencher W., Redman C.L., Schneider S.H., Taylor W.W., (2007) - Complexity of coupled human and natural systems, ***Science***, 317, 5844, 1513-1516.
- ***Loi n° 2004-575*** du 21 juin 2004 pour la confiance dans l'économie numérique. NOR : ECOX0200175L Version consolidée au 09 janvier 2017.
- ***Loi n° 2016-1321*** du 7 octobre 2016 pour une République numérique. NOR : ECFI1524250L Version consolidée au 16 décembre 2016.
- MacNeil M.A., Tyler E.H.M., Fonnesbeck C.J., Rushton S.P., Polunin N.V.C., Conroy M.J., (2008) - Accounting for detectability in reef-fish biodiversity estimates. ***Marine Ecology Progress Series***, vol. 367, 249-260. doi: 10.3354/meps07580.
- Marschollek M., (2012) - Decision support at home (DS@ HOME)–system architectures and requirements. ***BMC medical informatics and decision making***, May 2012, 12/43, 8 pp., doi: 10.1186/1472-6947-12-43.

- Marraffini M.L., Ashton G.V., Brown C.W., Chang A.L., Ruiz G.M., (2017) - Settlement plates as monitoring devices for non-indigenous species in marine fouling communities. ***Management of Biological Invasions*** 8: published online.
- Marakas, G.M., (1999) - Decision support systems in the twenty-first century, ***Prentice Hall, Inc. Upper Saddle River, N.J.***, ISBN:0-13-744186-X
- Marion A.F. (1983) - Esquisse d'une topographie zoologique du Golfe de Marseille. ***Annales du Muséum d'Histoire Naturelle de Marseille***, Marseille.
- Martin S., Charnoz A., Gattuso J.P. (2013) - Photosynthesis, respiration and calcification of the Mediterranean crustose coralline alga Lithophyllum cabiochae (Corallinales, Rhodophyta). ***European Journal of Phycology*** 48, 163 – 172.
- Martins, G.M., Faria, J., Rubal, M., Neto, A.I., (2013) - Linkages between rocky reefs and soft-bottom habitats: effects of predation and granulometry on sandy macrofaunal assemblages. ***Journal of Sea Research***. 81, 1–9.
- Mellin C., Mouillot D., Kulbicki M., McClanahan T.R. , Vigliola L., Bradshaw C.J.A., Brainard R.E., Chabanet P., Edgar G.J., Fordham D.A., Friedlander A.M., Parravicini V., Sequeira A.M.M., Stuart-Smith R.D., Wantiez L., Caley M.J., (2016) - Humans and Seasonal climate variability threaten large-bodied coral reef fish with small ranges. ***Nature Communications***, 7, 10491, doi:10.1038/ncomms10491
- Michener W.K., Brunt J.W., Helly J.J., Kirchner T.B., Stafford, S.G., (1997) - Non geospacial metadata for the ecological sciences. ***Ecological Applications***, 7: 330–342. doi:10.1890/1051-0761(1997)007 [0330:NMFTES] 2.0.CO;2
- Michez N., Dirberg G., Bellan-Santini D., Verlaque M., Bellan G., Pergent G., Pergent-Martini C., Labrune C., Francour P., Sartoretto S. (2011) - Typologie des biocénoses benthiques de Méditerranée, Liste de référence française et correspondances. ***Rapport du Service du Patrimoine Naturel***, 13. MNHN, Paris. 48 pp.
- Mills J.A., Teplitsky C., Arroyo B., Charmantier A., H. Becker P., Birkhead T.R., Bize P., Blumstein D.T., Bonenfant C., Boutin S., Bushuev A., Cam E., Cockburn A., Côté S.D., Coulson J.C., Daunt F., Dingemanse N.J., Doligez B., Drummond H., Espie R.H.M., Festa-Bianchet M., Frentiu F., Fitzpatrick J.W., Furness R.W., Garant D., Gauthier G., Grant P.R., Griesser M., Gustafsson L., Hansson B., Harris M.P., Jiguet F., Kjellander P., Korpimäki E., Krebs C.J., Lens L., Linnell J.D.C., Low M., McAdam A., Margalida A., Merilä J., Møller A.P., Nakagawa S., Nilsson J.-Å., Nisbet I.C.T., van Noordwijk A.J., Oro D., Pärt T., Pelletier F., Potti J., Pujol B., Réale D., Rockwell R.F., Ropert-Coudert Y., Roulin A., Thébaud C., Sedinger J.S., Swenson J.E., Visser M.E., S.Wanless Westneat D.F., Wilson A.J., Zedrosser A., (2015) - Archiving Primary Data: Solutions for Long-Term Studies, ***Trends in Ecology & Evolution***, October 2015, Vol. 30, No. 10, 581-589.

- Minelli A., (2009) - Commentaries on Gentile & Snell (2009): an introduction. ***Zootaxa***, 2201, 11.
- Mooney H.A., (1999) - On the road to global ecology. Annual Review of Energy and the Environment. Vol. 24:1-31 https://doi.org/10.1146/annurev.energy.24.1.1.
- Muñoz V., Cohen Nabeiro A., Couvet D., David R., Delavaud A., Féral J.P., Ivars V.J., Nonell-Canals A., Senar M.A. and Tatoni T., (2017) - Analysis on the graph techniques for data-mining and visualization of heterogeneous biodiversity data sets, ***Proceedings of the 2nd International Conference on Complexity, Future Information Systems and Risk (COMPLEXIS)***. Apr 2017, Porto, Portugal, 144 - 151.
- Noisette F. (2013) - Impacts de l'acidification des océans sur les organismes benthiques calcifiants des milieux côtiers tempérés. ***Thèse de doctorat.*** Université Pierre et Marie Curie.
- Noss, R.F., (1990) - Indicators for monitoring biodiversity: a hierarchical approach. ***Conservation biology***, 4(4), 355-364.
- Nute D., Potter W.D., Maier F., Wang J., Twery M., Rauscher H.M., Knopp P., Thomasma S., Dass M., Uchiyama H., Glende A., (2004) - NED-2: an agent-based decision support system for forest ecosystem management". ***Environmental Modelling and Software***, vol. 19(9), September 2004, 831-843, doi: 10.1016/ j.envsoft.2003.03.002.
- Oksanen J., Blanchet F.G., Kindt R., Legendre P., Minchin P.R., O'Hara R.B., Simpson G.L., Solymos P., Stevens M.H.H., Wagner H., (2016) - vegan: ***Community Ecology Package***. R package version 2.3-4. <https://CRAN.R-project.org/package=vegan>
- Oprea M., Sanchez-Marré M., Wotawa F., (2005) - A case study of knowledge modelling in an air pollution control decision support system. ***AI Communications, Binding Environmental Sciences and AI***, vol. 18(4), December 2005, 293-303, ISSN:0921-7126.
- Pallottino S., Sechi G.M., Zuddas P., (2005) - A DSS for water resources management under uncertainty by scenario analysis, ***Environmental Modelling and Software***, vol. 20(8), August 2005, 1031-1042, doi: 10.1016/j.envsoft.2004.09.012.
- Patrício J., Little S., Mazik K., Papadopoulou N.J., Smith C., Teixeira H., Hoffmann H., Uyarra M.C., Solaun O., Zenetos A., Kaboğlu G., Kryvenko O., Churilova T., Moncheva S., Bucas M., Borja A., Hoepffner N., Elliott M., (2016) - European Marine Biodiversity Monitoring Networks: strengths, weaknesses, opportunities and threats. ***Frontiers in Marine Science*** 3.
- Pearman J.K., Anlauf H., Irigoien X., Carvalho S., (2016) - Please mind the gap - Visual census and cryptic biodiversity assessment at central Red Sea coral reefs. ***Marine Environmental Research*** 118: 20-30

- Pedel L., Fabri M.C., Menot L., Van Den Beld I., (2013) - Mesure de l'état écologique des habitats benthiques du domaine bathyal à partir de l'imagerie optique. (Sélection de métriques et proposition d'une stratégie de surveillance). Convention 13/1210491/NYF Convention MEDDE-Ifremer pour le Bon Etat Ecologique des habitats benthiques profonds.
- Pelletier D., Garcìa-Charton J.A., Ferrarris J., David G., Thébaud O., Letourneur Y., Claudet J., Amand M., Kulbicki M., Galzin R. (2005) - Designing indicators for assessing the effects of marine protected area on coral reef ecosystems : A multidisciplinary standpoint. ***Aquatic Living Resources,*** 18, 15-33.
- Pelletier D., Leleu K., (2008) - Utilisation de techniques vidéo pour l'observation et le suivi des ressources et des écosystèmes récifo lagonaires - ***Rapport d'opération ZONECO***. 81 pp.
- Pelletier D., Leleu K., Mallet D., Mou-Tham G., Hervé G., Boureau M., Guilpart N. (2012) - Remote High-Definition Rotating Video Enables Fast Spatial Survey of Marine Underwater Macrofauna and Habitats. ***PLoS One***, *7*(2), e30536. http://doi.org/10.1371/journal.pone.0030536
- Pennesi C., Danovaro R., (2017) - Assessing marine environmental status through microphytobenthos assemblages colonizing the Autonomous Reef Monitoring Structures (ARMS) and their potential in coastal marine restoration. *Marine Pollution Bulletin*.
- Pereira H.M., Ferrier S., Walters M., Geller G.N., Jongman R.H., Scholes R.J., Bruford M.W., Brummitt N., Butchart S.H., Cardoso A.C., Coops N.C., Dulloo E., Faith D.P., Freyhof J., Gregory R.D., Heip C., Höft R., Hurtt G., Jetz W., Karp D.S., McGeoch M.A., Obura D., Onoda Y., Pettorelli N., Reyers B., Sayre R., Scharlemann J.P., Stuart S.N., Turak E., Walpole M., Wegmann M., (2013) - Essential biodiversity variables. ***Science***, 339 (6117): 277-278. doi: 10.1126/science.1229931
- Pérès J.M., Picard J., (1964) - Nouveau manuel de bionomie benthique de la Méditerranée. ***Recueil des Travaux de la Station Marine d'Endoume***, 31(47) : 1-37.
- Pergent-Martini C., Alami S., Bonacorsi M., Clabaut P., Daniel B., Ruitton S., Sartoretto S., Pergent G., (2014) - New data concerning the coralligenous atolls of cap corse: an attempt to shed light on their origin. ***RAC/SPA 2nd Mediterranean Symp. on the Conservation of coralligenous and other calcareous bio-concretions***, Portorož (SIovenia), 29-30/10/2014, 129-134.
- Pesant S. et al. (2015) - Open science resources for the discovery and analysis of Tara Oceans data. **Sciences Data** 2:150023 doi: 10.1038/sdata.2015.23.

- Peters D.P., Havstad K.M., Cushing J., Tweedie C., Fuentes O., Villanueva-Rosales N., (2014) - Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. ***Ecosphere***, 5(6), 1-15.
- Piola R.F., Johnston E.L., (2008) - Pollution reduces native diversity and increases invader dominance in marine hard-substrate communities. ***Diversity and Distributions***, 14: 329-342.
- Poch M., Comas J., Roda I.R., Sànchez-Marrè M., Cortés U., (2004) - Designing and building real environmental decision support, ***Systems Environmental Modelling & Software***, vol. 19(9), September 2004, 857-873, doi: 10.1016/j.envsoft.2003.03.007.
- Pomeroy R.S., Parks J.E., Watson L.M., (2006) - Comment va votre AMP ? Guide sur les indicateurs naturels et sociaux destinées à évaluer l'efficade la gestion des aires marines protégées, Gland, Suisse et Cambridge, Royaume-Uni, ***Union mondiale pour la nature***, 232 p.
- Power D.J., (2007) - A Brief History of Decision Support Systems", DSSResources.COM (Editor), ***World Wide Web***, version 4.0, March 2007, <http://dssresources.com/history/dsshistory.html>.
- Prato G., Thiriet P., Di Franco A., Francour P., (2017) - Enhancing fish Underwater Visual Census to move forward assessment of fish assemblages: An application in three Mediterranean Marine Protected Areas. ***PLoS One*** 12(6): e0178511. https://doi.org/10.1371/journal.pone.0178511
- Puren M., (2016) - A l'´epreuve de l'hétérogénéité : données de recherche et interdisciplinarité : L'exemple du projet européen IPERION-CH. DHnord ***Humanités numériques: théories, débats, approches critiques***, Nov 2016, Lille, France. 2016, <https://www.meshs.fr/dhnord201160912174051W5,01440x900x1x1440x803x1frint1rwr2.php>. <hal-01408951>
- R Core Development Team, (2016) - R: A language and environment for statistical computing. ***R Foundation for Statistical Computing***, Vienna, Austria. URL <https://www.R-project.org/>.
- RAC/SPA UNEP – MAP (2006) - Classification des biocénoses benthiques marines de la région Méditerranéenne. ***CAR/ASP Tunis***, 13 pp.
- RAC/SPA UNEP – MAP (2009) - Proceedings of the 1st Mediterranean symposium on the conservation of the coralligenous and other calcareous bio-concretions, ***RAC/SPA Tabarka***, January. 1–278.
- Rajasekaram V., Nandalal K.D.W., (2005) - Decision Support system for Reservoir Water management conflict resolution". ***Journal of water resources planning and management***, vol. 131(6), November 2005, 1-10, doi: 10.1061/(ASCE)0733-9496(2005)131:6(410).

- Ransome E., Geller J.B., Timmers M., Leray M., Mahardini A., Sembiring A., Collins A.G., Meyer C.P., (2017) - The importance of standardization for biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. ***PLoS One***. 2017;12: e0175066. doi:10.1371/journal.pone.0175066
- Reichman O.J., Jones M.B., Schildhauer M.P., (2011) - Challenges and opportunities of open data in ecology. ***Science***, 331(6018).
- Ross J.D., Romero J., Ballesteros E., Gili J.M., (1984) - Diving in Blue Water. The benthos. In: Margalef R. (ed.), Western Mediterranean. ***Pergamon Press***, Oxford, 233-295.
- Rüegg J., Gries C., Bond-Lamberty B., Bowen G.J., Felzer B.S., McIntyre N.E., Soranno P.A., Vanderbilt K.L., Weathers K.C., (2014) - Completing the data life cycle: using information management in macrosystems ecology research. ***Frontiers in Ecology and the Environment***, 12: 24–30. doi:10.1890/120375
- Sainsbury K., Sumaila U.R., (2003) - Incorporating ecosystem objectives into management of sustainable marine fisheries, including 'Best Practice' reference points and use of marine protected areas. In Sinclair M., Valdimarsson G., Responsible Fisheries in the Marine Ecosystem, 343–361. Rome: ***FAO Cabi Publishing***.
- Sànchez-Marrè M., Gibert K., "Improving ontological knowledge with reinforcement in recommending the data mining method for real problems". In ***Proceedings of Conferencia de la Asociación Española para la Inteligencia Artificial*** (CAEPIA), Albacete, 9-12 November 2015,
- Sartoretto S. (1996) – Vitesse de croissance et bioérosion des concrétionnements "coralligènes" de Méditerranée nord-occidentale. Rapport avec les variations Holocènes du niveau marin. ***Thèse doctorat d'écologie***. Université d'Aix-Marseille II.. 194 pp.
- Sartoretto S., Verlaque M., Laborel J., (1996), Age of settlement and accumulation rate of submarine "coralligène" (−10 to −60 m) of the northwestern Mediterranean Sea; relation to Holocene rise in sea level. Marine Geology, 130 (3), 317–331.
- Sartoretto S., Schohn T., Bianchi C.N., Morri M.C., Garrabou J., Ballesteros E., Ruitton S., Verlaque M., Daniel B., Charbonnel E., Blouet S., David R., Féral J.P., Gatti G., (2017) - An integrated approach to evaluate and monitor the conservation state of coralligenous habitats: the Index-Cor approach. ***Marine Pollution Bulletin***. Elsevier, 2017, 〈10.1016/j.marpolbul.2017.05.020〉. 〈hal-01541141〉

- Selig E.R., Longo C., Halpern B.S., Best B.D., Hardy D., Elfes C.T. , Scarborough C., Kleisner K.M., Katona S.K., (2013) - Assessing Global Marine Biodiversity Status within a Coupled Socio-Ecological Perspective. ***PLoS One***, 8: e60284.
- Simms S., Jones S., Mietchen D., Miksa T., (2017) - Machine-actionable data management plans (maDMPs). ***Research Ideas and Outcomes*** 3: e13086. https://doi.org/10.3897/rio.3.e13086
- Sini M., Kipson S., Linares C., Koutsoubas D., Garrabou J., (2015) - The Yellow Gorgonian Eunicella cavolini: demography and disturbance levels across the Mediterranean Sea. ***PLoS One***, 10(5), e0126253.
- Sorte, C.J.B., Fuller A., Bracken M.E.S., (2010) - Impacts of a simulated heat wave on composition of a marine community. ***Oikos*** 119: 1909-1918.
- Spalding M.D., Fox H.E., Allen G.R., Davidson N., Ferdaña Z.A., Finlayson M.A.X., Halpern B.S., Jorge M.A., Lombana A., Lourie S.A., Martin K.D., McManus E., Molnar J., Recchia C.A., Robertson J., (2007) - Marine ecoregions of the world: a bioregionalization of coastal and shelf areas. ***BioScience***. 2007;57: 573–583.
- Specht A., Guru S., Houghton L., Keniger L., Driver P., Ritchie E.G., Lai K., Treloar A., (2015). Data management challenges in analysis and synthesis in the ecosystem sciences. ***Science of the Total Environment*** 534, Elsevier. 144-158. http://dx.doi.org/10.1016/j.scitotenv.2015.03.092
- Teixidó N., Casas E., Cebrián E., Linares C., Garrabou J., (2013) - Impacts on coralligenous outcrop biodiversity of a dramatic coastal storm. ***PLoS One*** 8, e53742.
- Teixidó N., Garrabou J., Harmelin J.-G. (2011a) - Low Dynamics, High Longevity and Persistence of Sessile Structural Species Dwelling on Mediterranean Coralligenous Outcrops. ***PLoS One***, 6 (8). e23744.
- Teixidó N., Albajes-Eizagirre A., Bolbo D., Le Hir E., Demestre M., Garrabou J., Guigues L., Gili J.M., Piera J., Prelot T., Soria-Frisch A., (2011b) - Hierarchical segmentation-based software for cover classification analyses of seabed images (Seascape). Mar. Ecol. Prog. Ser. 431, 45–53.
- The Royal Society, (2012) - Science as an open enterprise. ***Summary report. The Royal Society Science Policy Centre***, London
- Thierry de Ville d'Avray L., (2014) - Etude de la variabilité de la composition observée d'habitats coralligènes liée à l'application d'un protocole d'observation au moyen de quadra-photos. ***Mémoire de Master***, Aix Marseille Université
- Thierry de Ville d'Avray L., (2018) - Étude des services écosystémiques rendus par les habitats coralligènes de Méditerrannée et évaluation de leur valeur économique. ***Thèse de doctorat***, Aix-Marseille Université

- Thierry de Ville d'Avray L., Ami D., Chenuil A., David R., Féral J.P., (2018) - Application of the ecosystem service concept to a local-scale: the cases of coralligenous habitats in the North-western Mediterranean Sea. ***Marine Pollution Bulletin***
- Trygonis V., Sini M. (2012) - PhotoQuad: A dedicated seabed image processing software, and a comparative error analysis of four photo quadrat methods. ***Journal of Experimental Marine Biology and Ecology,*** 424, 99-108.
- van Hoytema N., Bednarz V.N., Cardini U., Naumann M.S., Al-Horani F.A., Wild C., (2016) - The influence of Seasonality on benthic primary production in a Red Sea coral reef. ***Marine Biology***, 163: 52, doi: 10.1007/s00227-015-2787-5
- Varanon U., Chan C.W., Tontiwachwuthikul P., (2007) - Artificial Intelligence for monitoring and supervisory control of process systems. ***Engineering applications of artificial intelligence***, 20(2): 115-131, doi: 10.1016/j.engappai.2006. 07.00
- Vassallo P., Bianchi C.N., Paoli C., Holon F., Navone A., Bavestrello G., Cattaneo Vietti R., Morri C., (2018) - A predictive approach to benthic marine habitat mapping: Efficacy and management implications, ***Marine Pollution Bulletin***, Volume 131, Part A, June 2018, Pages 218-232, ISSN 0025-326X, https://doi.org/10.1016/j.marpolbul.2018.04.016.
- Virgilio M., Airoldi L., Abbiati M., (2006) - Spatial and temporal variations of assemblages in a Mediterranean coralligenous reef and relationships with surface orientation. ***Coral Reefs***, 25, 265-272.
- Wang R., Strong D., (1996) - Beyond Accuracy: What Data Quality Means to Data Consumers. ***Journal of Management Information Systems***. 1996;12(4):30.
- Wieczorek J., Bloom D., Guralnick R., Blum S., Döring M., Giovanni R., Robertson T., Vieglais D., (2012) - Darwin Core: An evolving community-developed biodiversity data standard. ***PLoS One***, 7(1), e29715.
- Wilkinson, M. D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.W., da Silva Santos L.B., Bourne P.E., Bouwman J., Brookes A.J., Clark T., Crosas M., Dillo I., Dumon O., Edmunds S., Evelo C.T., Finkers R., Gonzalez-Beltran A., Gray A.J., Groth P., Goble C., Grethe J.S., Heringa J., 't Hoen P.A., Hooft R., Kuhn T., Kok R., Kok J., Lusher S.J., Martone M.E., Mons A., Packer A.L., Persson B., Rocca-Serra P., Roos M., van Schaik R., Sansone S.A., Schultes E., Sengstag T., Slater T., Strawn G., Swertz M.A., Thompson M., van der Lei J., van Mulligen E., Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K., Zhao J., Mons B., The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).

Glossaire

Ce glossaire est rédigé partiellement à partir des éléments traduits et adaptés de la publication du protocole CIGESMED d'une part, et d'éléments validés du micro-thésaurus CIGESMED d'autre part : David R., Arvanitidis C., Çinar M.E., Dubar J., Dubois S., Erga Z., Guillemain D., Sartoretto S., Thierry de Ville d'Avray L., Zuberer F., Chenuil A., Féral J.-P., (2014), with contributors : Açik Çinar S., Andral B., Aurelle D., Aysel V., Bakir K., Bellan G., Bellan-Santini D., Bouchoucha M., Celik C., Chatzigeorgiou G., Chatzinikolaou E., Chenesseau S., Dağli E., Dailianis T., Dimitriadis C., D'Iribarne C., Doğan A., Dounas C., Egea E., Elguerrabi W., Emery E., Evcen A., Faulwetter S., Gatti G., Gerovasileiou V., Güçver S.M., Issaris Y., Katağan T., Keklikoglou K., Kirkim F., Koçak F., Koutsoubas D., Marschal C., Önen M., Önen S., Öztürk B., Panayiotidis P., Pavloudi C., Pergent G., Pergent-Martini C., Poursanidis D., Ravel C., Reizopoulou S., Rocher C., Ruiton S., Sakher S., Salomidi M., Sarropoulou E., Selva M., Sini M., Sourbes L., Simboura N., Taşkin E., Vacelet J., Valavanis V., Vasileiadou A., Verlaque M. Protocols for monitoring of coralligenous habitats of Mediterranean (Coralligenous based Indicators to Evaluate and Monitor the "good ecological status" of the MEDiterranean coastal waters) Protocoles de suivi du coralligène en méditerranée (Coralligenous based Indicators to Evaluate and Monitor the "good ecological status" of the MEDiterranean coastal waters).

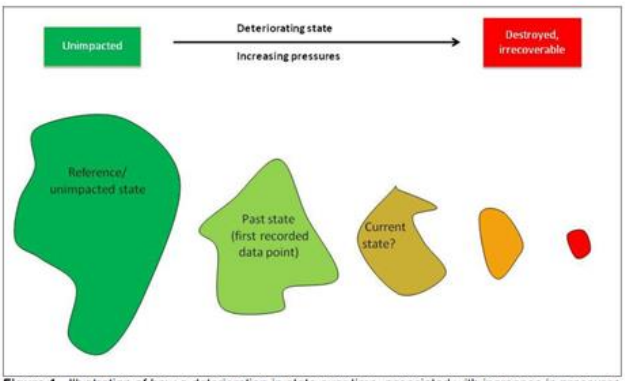
Ces définitions, lorsqu'elles n'ont pas de source, sont des propositions débattues dans le cadre de l'écriture des protocoles, et proposées dans le cadre du micro-thesaurus. Ces définitions, utiles pour comprendre leur sens dans le cadre de ce travail interdisciplinaire, pourront évoluer.

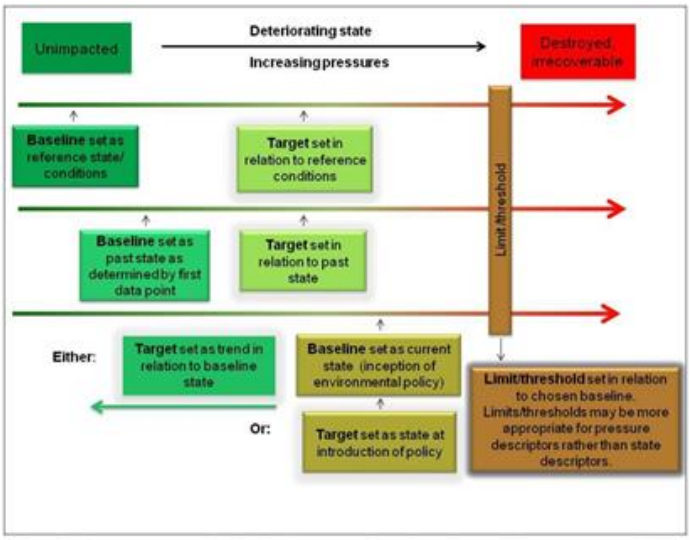
Accessibilité d'un dispositif d'observation	L'accessibilité d'un système d'observation est définie par le niveau de facilité avec lequel un observateur va pouvoir mettre en œuvre un dispositif d'observation (coût du matériel + compétences sollicitées + temps à impartir en formation, observation, traitement). Elle a une influence sur la taille du panel d'observateurs compétents à large échelle et sur les coûts.
Biocénose	La biocénose est un groupement d'organismes vivants, liés par des relations d'interdépendance dans un biotope dont les caractéristiques dominantes sont relativement

	homogènes ; chaque biocénose comprend notamment la phytocénose, limitée aux végétaux, et la zoocénose, limitée aux animaux (d'après PNUE, PAM, CAR/ASP, 2006).
Association (d'espèces)	L'association (d'espèces) est un aspect permanent d'une biocénose avec une dominance physionomique végétale dans laquelle les espèces sont liées par une compatibilité écologique et une affinité chorologique (d'après PNUE, PAM, CAR/ASP, 2006).
Bathyal (étage)	L'étage bathyal s'étend des marges du plateau continental (-200 mètres ; étage circalittoral) au début des plaines abyssales (-2.000 à -4.000 mètres ; étage abyssal).
Biotope	Le biotope est l'aire géographique de surface ou de volume variable soumise à des conditions écologiques où les dominantes sont homogènes. (D'après PNUE, PAM, CAR/ASP, 2006).
Cadre de positionnement	Le cadre de positionnement permet de positionner 9 photographies avec une variabilité minimisée et à partir d'un point fixe. Les coins du cadre sont en fait des croix autour desquelles se calent les quadrats photo.
Cartographie complète	La cartographie complète est un relevé complet des profils en détaillant tous les champs selon les recommandations du protocole.
Cartographie partielle	La cartographie partielle est un relevé limité aux profils intéressants à prélever (i.e. profils prédéterminés pour lesquels la fréquence du profil est attestée et la fréquence des populations de *Myriapora truncata* et *Lithophylum cabioche* sont suffisantes pour subir un prélèvement). Cette cartographie sera majoritairement mise en œuvre pour les prélèvements sur sites secondaires, lointains ou pour lesquels il n'y a pas de *continuum* de coralligène suffisant pour effectuer une cartographie complète. Cela peut aussi

	être le cas pour la deuxième profondeur, moins accessible, afin de raccourcir le temps d'intervention.
Circalittoral	L'étage circalittoral, correspond à la partie basse de la zone photique, la partie du littoral la plus profonde, presque totalement sombre. En Méditerranée, cet étage commence à la limite inférieure des herbiers de posidonies, jusqu'à la profondeur où les algues sciaphiles disparaissent.
Contextualisation (donnée de contexte)	La contextualisation consiste à donner un contexte à une donnée, attribuer un qualificatif correspondant aux conditions d'enregistrements ou aux conditions particulières biotiques ou abiotiques dans lesquelles la donnée a été mesurée. Dans le cadre de CIGESMED, elle correspond à une attribution d'un profil à un échantillon, afin de pouvoir étudier le jeu de donnée en ne s'intéressant qu'à un type de profils.
Cycle de vie des données de Recherche	Le cycle de vie des données de recherche décrit le processus d'utilisation des données de leur création à la publication et à leur réutilisation ultérieure.
Données de références	Les données de références sont des données partagées par l'ensemble des processus et partenaires d'une organisation. Elles sont à la base des prises de décision.
Dump	Un dump est une sauvegarde des bases de données, dans laquelle peut choisir de conserver la description de la structure des données.
Ecotone	L'écotone est la transition entre deux écosystèmes ou habitats ou zones aux conditions biotiques et/ou abiotiques différentes.
Efficacité (d'un indicateur)	Un indicateur est efficace s'il a une amplitude de variation correspondant à l'amplitude observable du phénomène qu'il indique, s'il varie de manière homogène quelque soit l'ordre de grandeur du phénomène indiqué, et si ses variations sont spécifiquement associées au phénomène qu'il indique. Il doit

	avoir la précision requise pour indiquer un changement précisément, c'est-à-dire sans retard.
Entité (en informatique)	Une entité est la représentation d'un élément matériel ou immatériel dans un système d'information.
Epibionte	L'épibionte est un organisme non parasite utilisant les surfaces externes d'animaux plus grands que lui comme support.
Epiphyte	L'épiphyte est un organisme autotrophe (capable de photosynthèse) se servant d'autres plantes comme support.
ESEI	L'ESEI est la métrique de l'indice IndexCor qui correspond au ratio du nombre d'espèces sensibles sur le nombre d'espèces indifférentes.
Espèces cryptiques	Les espèces cryptiques sont définies comme telles car isolées reproductivement et/ou dont la lignée génétique a une importante différenciation génétique, indiquant une divergence ancienne entre l'une et l'autre, mais qui n'est pas distinguable d'un point de vue morphologique.
Espèces endolithes	Les espèces endolithes vivent enfouies dans la roche ou les concrétions coralligènes en y creusant des cavités ou en utilisant celles d'autres espèces.
Espèce invasive	Une espèce invasive est une espèce colonisant une région non connectée à sa région d'origine et dont l'aire de répartition s'agrandit rapidement et durablement (souvent par le fait de l'Homme). Elle est capable de se reproduire sans l'aide de l'homme, ce qui pose des problèmes écologiques. Les phénomènes d'invasion biologique sont aujourd'hui considérés par l'ONU comme une des grandes causes de régression de la biodiversité, avec la pollution et la fragmentation écologique des écosystèmes.
Espèce saisonnière	Une espèce est saisonnière lorsqu'elle ne perdure pas pendant toutes les saisons. Sa présence sur un site ou un

	transect peut donc être ignorée, si la fréquence d'observation pour la prendre en compte n'est pas assez forte.
Etagement	L'étagement marin est fonction de facteurs ambiants. Ces facteurs sont : l'humectation, la lumière, comme facteurs principaux (facteurs climatiques) et l'hydrodynamisme, la salinité, la nature du substrat et la température, comme facteurs secondaires (facteurs édaphiques). (d'après Bellan-Santini *et al.*, 1994)
Etalonnage	L'étalonnage est la calibration d'un dispositif de mesure sur une échelle de valeur connue, en estimant l'incertitude associée à la mesure via un étalon de valeur connue.
Etat initial = Ligne de base ([en] : baseline)	L'état inititial est un état écologique qui correspond à l'état mesuré pour la première fois lors de l'étude de l'habitat (par opposition à l'état de référence), sans connaissance de son état antécédent (qui peut être meilleur ou moins bon). Définir un état initial permet de mettre en évidence des tendances et/ou de définir des objectifs (rester en l'état, améliorer telle ou telle métrique).  **Figure 1.** Illustration of how a deterioration in state over time, associated with increases in pressures and impacts, can include changes in both quality (e.g. of a habitat or population of a species) and quantity (e.g. habitat extent, population size) of a biodiversity component. Setting the baseline as 'current state' represents a very different scenario to using 'past state' or 'reference state'. Figure from Moffat *et al.*, 2011.
Etat de référence	L'état de référence est un état écologique qui correspond à l'état stable et n'ayant subi aucune pression autre que naturelle. On peut parler aussi d'un état optimum. Les

	conditions naturelles n'étant pas les mêmes pour un même habitat selon la région, l'état optimum est variable selon le contexte naturel.  **Figure 2.** The conceptual relationship between various baseline conditions, targets and limits. Figure from Moffat *et al.*, 2011.
Faciès	Le faciès est un aspect présenté par une biocénose coralligène lorsque la prédominance locale de certains facteurs biotiques et abiotiques entraîne l'exubérance d'une ou d'un très petit nombre d'espèces.
Facteurs de variabilité	Les facteurs de variabilité sont liés aux conditions de mesures, à l'observateur, à l'opérateur ou au système mesuré. Ils ont une influence sur l'exactitude de la mesure.
Fiabilité (d'un indicateur)	La fiabilité d'un indicateur est un paramètre dont l'évolution permet l'évaluation d'une situation sans biais. Un indicateur est fiable s'il est à la fois pertinent, efficace, sensible et robuste.
Génétique des populations	La génétique des populations est l'étude de la distribution et des changements de la fréquence des versions d'un gène (allèles) dans les populations d'une même espèce. La génétique des populations permet par exemple de répondre aux questions concernant la connectivité de deux populations d'une même espèce.

Graphe valué	Les graphes valués présentent des arêtes pondérées (par exemple par une fréquence d'espèce) dont la pondération agit sur les distances relatives entre les sommets.
Habitat	L'habitat est une zone se distinguant par ses caractéristiques géographiques, abiotiques et biotiques (définition de la directive 92/43 CEE – Anonyme, 1992). Cette définition peut correspondre à celle de biocénose, faciès et profil.
Infralittoral (étage)	En méditerranée, l'étage infralittoral est la partie du littoral constamment immergée dont la frange supérieure correspond à la ligne de base (le zéro des cartes). Sa limite inférieure est celle qui est compatible avec la vie des algues photophiles et des phanérogames marines.
Inter-calibration	L'inter-calibration consiste à confronter des techniques ne faisant varier qu'un paramètre (dans le meilleur des cas), afin de comparer l'amplitude des variations liées à ce paramètre. (Déf hrydrobio-dce.cemagref.fr : L'inter-calibration a pour objectif de valider la compatibilité des méthodes utilisées par différents observateurs/opérateurs en harmonisant le type de résultats qu'elles fournissent. Pour un même échantillon, les différentes méthodes doivent fournir un résultat comparable en terme d'efficacité et de pertinence).
Interopérabilité	L'interopérabilité est la capacité que possède un système informatique à fonctionner avec d'autres produits ou systèmes informatiques, existants ou futurs, sans restriction d'accès ou de mise en œuvre. Les deux termes sont normalisés par la CSA et la Commission électrotechnique internationale (ISO/IEC 2382-18:1999). La définition de l'interopérabilité a été élargie au-delà des dimensions informatiques comme la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants

	ou futurs et ce sans restriction d'accès ou de mise en œuvre. (Article 4 de la loi no 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique).
Isobathe	L'isobathe est la ligne imaginaire détourant un relief sous-marin à profondeur égale.
Large échelle (système d'observation à….)	Par système d'observation à large échelle, on entend un système d'observation pertinent pour effectuer des suivis à une échelle pan régionale, continentale ou intercontinentale.
Ligne de base	La ligne de base correspond à un état initial.
Localité	La localité est la région géographique où se regroupent des sites d'échantillonnage d'un même partenaire (exemple : localité de Marseille).
Métier (base de données ou système d'information)	"métier" se dit d'une base de données, d'une application ou d'un système d'information qui ont des caractéristiques spécifiques à un métier, et évoluant avec le métier qu'ils équipent.
Métrique	La métrique est un paramètre décrivant un phénomène, une variable biotique ou abiotique sous un format numérique.
Modèle Conceptuel de Communication (M.C.C.)	Le M.C.C. représente de manière normalisée les communications entre les différents acteurs dans le cadre d'un projet ou d'un métier.
Modèle Conceptuel de Données (M.C.D.)	Le M.C.D. est une représentation des données qui décrit de façon formelle les données utilisées par le système d'information sous forme d'entités.
Modèle Numérique de Terrain	Le modèle numérique de terrain est une cartographie en trois dimensions.
Observateur	L'observateur est responsable de l'acquisition des données sur le terrain.

Opérateur	L'opérateur exploite les observations ou données de terrain à l'aide de logiciels (souvent sous format numérique).
Palanquée	La palanquée est un groupe de plongeurs ayant les mêmes paramètres de plongée (direction, profondeur, durée et paliers).
Parallélisation des calculs	La parallélisation des calculs permet de raccourcir le temps de calcul en découpant ce calcul et en le répartissant sur différentes machines. La grille de calcul européenne permet de solliciter plusieurs dizaines de milliers de machines en même temps, ce qui permet par exemple de raccourcir un calcul qui durerait une année entière à quelques heures.
Parsing	Le parsing est une analyse syntaxique d'un flux de caractères (X.M.L., J.SON ou autre) qui permet soit de le segmenter en éléments plus petits, soit d'utiliser un motif, constitué d'un ou plusieurs modèles pour extraire du flux les données qui correspondent au motif, en vue de les manipuler.
Pertinence (d'un indicateur)	On dit d'un indicateur qu'il est pertinent s'il varie uniquement ou au moins *principalement* en fonction du paramètre étudié / surveillé.
Phylogéographie	La phylogéographie est l'étude des processus qui expliquent la distribution des lignées généalogiques au sein de la même espèce (processus allant éventuellement jusqu'à la spéciation).
Photophile	Un organisme photophile, animal ou végétal, se développe préférentiellement dans les zones suffisamment exposées à la lumière. Ex : Phanérogames marines.
Point nodal d'indexation	On appelle point nodal d'indexation un service web qui va moissonner les flux de données pour indexer les enregistrements en utilisant des descripteurs communs à ces flux de données. Un point nodal peut être spécifique d'un domaine et/ou d'une zone géographique, et peut

	sélectionner une partie d'un flux servi par un autre point nodal, ou regrouper des flux de plusieurs points nodaux.
Polysémie	"Polysémie" se dit d'un terme lorsqu'il peut prendre plusieurs sens selon les régions, disciplines ou matières considérées. Il s'agit d'un des freins principaux aux études transdisciplinaires.
Profil	Dans le cadre de CIGESMED, on appelle profil une conjoncture entre exposition (i.e. inclinaison + orientation + rugosité) et type de recouvrement. 2 profils seront ensuite choisis selon leur représentativité sur les 2 faces de l'île ainsi que deux profondeurs pour y réaliser les différents suivis (étude diversité interspécifique, étude diversité intraspécifique et connectivité, transects photos et vidéos, IndexCor, et possible ouverture à d'autres acteurs). Profil = recouvrement + orientation + inclinaison + rugosité.
Quadrat	Le quadrat est un cadre servant à cadrer la photo (prise en plongée dans notre protocole). Celui-ci peut être indépendant ou solidaire du dispositif de prise de vue (photo ou film).
Quadrat-photo	Un quadrat-photo est une photographie prise à travers le quadrat, délimitant une aire précise. C'est une unité d'échantillonnage photographique.
Qualification (de la donnée)	On entend par qualification de la donnée une métadonnée (description du site, des conditions d'observation, du système d'observation, de l'observateur) ou un enrichissement de cette donnée par une propriété issue d'une ontologie. Cette qualification peut se faire dans la base de données ou dans des index associés. Cette qualification permet de trier les données et/ou de les analyser avec des approches globales (multivariées) ou non statistiques (fouille de données).

Robustesse (d'un indicateur)	Un indicateur est robuste s'il conserve une valeur constante lors de la répétition d'événements en conditions identiques.
RVA (Rapid Visual Assessment)	Le RVA est une méthode d'appréciation de l'état des habitats coralligènes en plongée basée sur l'évaluation des peuplements sur une surface restreinte (Gatti *et al.*, 2015)
Sciaphile	L'adjectif sciaphile se dit d'un organisme, animal ou végétal, qui se développe préférentiellement dans les zones d'ombre, dans les milieux peu exposés à la lumière, adjectif opposé de photophile
Segment	Un segment est une longueur cartographiée (objet géographique) de 5 m à laquelle est attribué un profil. Un transect est donc composé de différents segments (que l'on considère comme les pixels de notre cartographie). C'est la plus petite unité cartographique utilisée pour la cartographie des profils dans le protocole CIGESMED.
Sensibilité (d'un indicateur)	Un indicateur est dit sensible s'il a un degré d'indication précis et simultané aux variations à observer. Il doit pour cela être constitué de catégories qualificatives ou quantitatives bien distinctes pour qu'il n'y ait pas d'incertitude lors de l'attribution de sa valeur.
Série de données	Une série de données est une "compilation identifiable de données géographiques", une donnée géographique étant « toute donnée faisant directement ou indirectement référence à un lieu ou une zone géographique spécifique ». le terme « identifiable » signifie que la série doit avoir un sens pour ses utilisateurs potentiels. En particulier ces derniers doivent pouvoir identifier facilement, parmi les thèmes des trois annexes de la directive, celui ou ceux qui sont concernés par la série de données géographiques.
Site	Dans le cadre de CIGESMED, quand on parle de site, il peut s'agir d'une île, d'un îlot, d'un sec, d'un bout de côte ou d'une langue de coralligène, tant que celui-ci possède un

	maximum d'orientations différentes et au moins deux orientations diamétralement opposées.
Spill-over	On appelle effet spillover ou effet de débordement dans une réserve, un transfert de la biomasse d'individus adultes et juvéniles vers les zones périphériques. L'effet spillover contribue à l'amélioration de la production des espèces pêchées à proximité d'une réserve, en raison de l'accroissement net de juvéniles et d'adultes dans celle-ci (source : http://www.aires-marines.fr/Glossaire/Spillover)
Strate inférieure ou Strate basale	Dans CIGESMED, la strate inférieure ou basale est la moins élevée correspondant aux espèces gazonnantes et encroûtantes. Dans cette strate, la compétition entre les espèces se fait essentiellement sur le substrat. NB : certaines espèces peuvent faire partie des deux strates.
Strate supérieure	Dans CIGESMED, la strate supérieure est la plus élevée correspondant aux espèces érigées. Dans cette strate, la compétition entre les espèces se fait essentiellement au-dessus du substrat. NB : certaines espèces peuvent faire partie des deux strates.
Substrat	Un substrat est un support naturel (en général le fond) sur lequel se développe un organisme ou un micro-organisme.
Topologie d'un graphe	Le terme topologie est utilisé pour décrire la forme d'un graphe, donné par les propriétés de ses composants (type de noeuds, nombre de noeuds, type de liens, propriété des liens, etc.).
Traits biologiques	Les « traits biologiques » appelés aussi « traits de vie » sont des descripteurs biologiques et comportementaux associés à un taxon, à une communauté ou à un habitat. Ils peuvent être quantitatifs (respiration, croissance, mode/rythme/stratégie de reproduction et alimentation) ou écologiques (préférendum de température, dureté, pH, etc.).

Transdisciplinarité	Le mot transdisciplinaire est souvent défini comme un synonyme de « interdisciplinaire » (qui traverse les frontières entre les disciplines). La nuance proposée par certains et à laquelle je souscris est que c'est à cette frontière que se développent de nouvelles disciplines, on peut alors parler non plus d'interdisciplinarité mais de transdisciplinarité.
Transect	Le transect est la trajectoire du plongeur sur laquelle des relevés (occurrence, abondance, biomasse...) peuvent être faits de manière systématique ou aléatoire. Dans notre cas, ces relevés seront faits à partir de quadrats photo de différentes tailles du substrat coralligène.
Transect aléatoire	Dans le cadre de CIGESMED, un transect aléatoire est 'un transect dont on ne prédétermine pas le début et la fin (mais on peut prédéterminer le nombre de photos ou la longueur).
Transect discontinu par opposition au transect continu	Dans le cadre de CIGESMED, un transect discontinu correspond à une trajectoire de transect où l'on prend des quadrats photo non bout à bout. Ces transects peuvent se faire à profondeur constante, et permettent de faire des suivis sur les sites ayant des habitats coralligènes discontinus.
Transect permanent	Dans le cadre de CIGESMED, un transect permanent correspond à une trajectoire ou l'on prend des photos de quadrats en partant d'un marquage ou un point de repère identifiable. Il permet à deux observateurs de s'inter calibrer.
Turbidité	La turbidité est une caractéristique quantifiable d'une eau dont la transparence est limitée par la présence de matières solides en suspension, entraînées par des courants et des tourbillons.
Universally Unique IDentifier (U.U.I.D.)	Universally Unique Identifier signifie littéralement « identifiant universel unique ». Il s'agit d'un numéro attribué à un objet permettant à des systèmes distribués d'identifier de façon unique une information sans coordination centrale

	importante. Dans ce contexte, le mot « unique » doit être pris au sens de « unicité très probable » plutôt que « garantie d'unicité », car celui-ci est généré aléatoirement sur un très grand nombre de possibilités (plus de 10^{38}).
Véracité des données	La véracité est la capacité d'un grand ensemble de données contenant certaines données incertaines à donner les mêmes résultats lorsqu'il est soumis à l'analyse que le même ensemble comportant uniquement des données « certifiées ».
World Geodetic System 1984 (WGS 84)	Le WGS 84 (système géodésique mondial, révision de 1984) est le système géodésique standard mondial, notamment utilisé par le système GPS.

Sigles, Acronymes et Abréviations

A.A.M.P : Agence des aires marines protégées

A.B.C.D : Access to Biological Collections Data

ACC : Algues Corallines Calcaires

A.C.O. : A Connected Ocean (conference)

A.F.B. : Agence Française de la Biodiversité

AFUL : Association Francophone des Utilisateurs de Logiciels Libres)

A.M.P : Aire Marine Protégée

AnaEE : Analyses et Expérimentations pour les Ecosystèmes

A.N.R : Agence Nationale de la Recherche

ANSES : Agence Nationale de SEcurité Sanitaire de l'alimentation, de l'environnement et du travail

A.Q. / C.Q. : Assurance Qualité / Contrôle Qualité

ARMS : Autonomous Reef Monitoring Structures

Ad.S: Adriatic Sea (Mer Adriatique)

ASUs : Artificial Substrate Units

AZTI Technalia :

B. D. A. : Bases de Données Avancées (conférence sur la « Gestion de Données — Principes, Technologies et Applications »)

B.E.E. : Bon Etat Ecologique

BoB: Golfe de Gascogne Bay of Biskay (Golfe de Gascogne)

B.R.G.M. : Bureau de Recherches Géologiques et Minières

CAR ASP : Centre d'Activité Régionale pour les Aires Spécialement Protégées

CARTHAM : CARtographie des Habitats Marins

C.A.T.D.S. (données du satellite S.M.O.S.) : Centre Aval de Traitement des Données SMOS

C.D.B : Convention sur la Diversité Biologique

C.E.R : conseil scientifique du Conseil Européen de la Recherche

CESAB : Centre de Synthèse et d'Analyse sur la Biodiversité

CHARLIEE : CHAnger de Regard En Liant dans IndexMed l'Environnement et les Etoiles

CIESM: Commission Internationale pour l'Exploration Scientifique de la Méditerranée

CIGESMED: Coralligenous based Indicators to evaluate and monitor the "Good Environmental Status" of the MEDiterranean coastal waters

CINES : Centre Informatique National de l'Enseignement supérieur

C.M.E.M.S. : Copernicus Marine Environment Monitoring Service

C.N.R.S. : Centre National de la Recherche Scientifique

COM : Centre d'Océanologie de Marseille

CONISMA : COnsorzio Nazionale Interuniversitario per le Scienze del MAre

C.M.S: Content Management System

C.N.R.S : Centre National de la Recherche Scientifique

CoL : Catalogue of Life

CRED : Division des REcifs Coralliens

Data-ONE : Data Observation Network for Earth

D.C.E : Directive Cadre sur l'Eau

D.C.S.M.M : Directive-Cadre sur la Stratégie pour le Milieu Marin

D.E.E : Données Élémentaires d'Échange

DEVOTES : DEVelopment Of innovative Tools for understanding marine biodiversity and assessing good Environmental Status

D.G.M.E : Direction générale de la modernisation de l'État

DIMAR : DIversité, évolution et écologie fonctionnelle MArine

D.M.P. : Data Management Plan

E.B.V : Variables Essentielles de la Biodiversité

E.C / D.H.F.F : Directive Habitat Faune Flore

EcoStat : ECOlogie STATistique (G.D.R.)

E.C.S.A. : European Citizen Science Association

E.G.I. : European Grid Infrastructure

E. M. B. R. C.: European Marine Biological Resources Centre

E.M.B.S. : European Marine Biology Symposium

E.M.L. : Ecological Metadata Language

EMODnet : European Marine Observation and Data Network

E.P.D. : European Pollen Database

E.P.H.E. : École Pratique des Hautes Études

ESEI : Espèces Sensibles Espèces Indifférentes

ERANET : European Research Area Net

EU-BON : European Biodiversity Observation Network

EVACOR et EVACOR2 EVAluation des services écosystémiques des habitats CORalligènes (1 et 2)

F.A.O : Organisation des Nations Unies pour l'alimentation et l'agriculture

FAIR : Findable, Accessible, Interoperable, Reusable

F.R.B. : Fondation pour la Recherche sur la Biodiversité

G.B.I.F. : Système mondial d'information sur la biodiversité / Global Biodiversity Information Facility

GEO-BON : Group On Earth Observations - Biodiversity Observation Network

G.E.S : Good Environmental Status

GIPREB : Groupement d'Intérêt Public pour la Réhabilitation de l'Etang de Berre
G.P.S : Global Positioning System
GRAMINÉES : GRAphe data Mining In Natural, Ecological and Environmental Sciences
HAL : Hyper Articles en Ligne
H.C.M.R.: Hellenic Center for Marine Research Institute of Marine Biology, Biotechnology and Aquaculture
I.C.S : Indice de Complexité Structurale
I.D.S.S. : Intelligent Decision Support System
I.E.E.E.: Institute of Electrical and Electronics Engineers
iEMSs : International Environmental Modelling and Software Society
IFREMER : Institut Français de Recherche pour l'Exploitation de la Mer
I.M.B.E : Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale
IndexCor : programme INDEX-COR visant à mettre au point un indice global de l'état de conservation des formations coralligènes, à destination des gestionnaires de milieu.
IndexMed : Indexer les données Méditerranéennes (consortium)
IndexMEED Indexing for Mining Ecological and Environmental Data
INIST : INstitut de l'Information Scientifique et Technique
I.N.P.N. : Institut National du Patrimoine Naturel
INRA : Institut National de la Recherche Agronomique
INSPIRE : Infrastructure for Spatial Information in Europe (EU directive)
I.P.B.E.S. : Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (Plateforme Intergouvernementale sur la Biodiversité et les Services Écosystémiques)
I.P.T. : Integrated Publishing Toolkit
LIRMM : Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
MaDICS : Masses de Données, Informations et Connaissances en Sciences (G.D.R.)
MASTODONS : grandes MASses DE DONnées Scientifiques
M.C.C. : Modèle Conceptuel de Communication
M.C.D. : Modèle Conceptuel de Données
M.I.O. : Institut Méditerranéen d'Océanographie
M.N.H.N. : Muséum National d'Histoire Naturelle
M.N.T. : Modèle Numérique de Terrain
mod. : modalités
M.S.F.D. : Marine Strategy Framework Directive
M.T.92 : tables de plongée du Ministère du Travail de 1992
N.G.S. : Next-Generation Sequencing
N.M.D.S. : Non-metric Multi Dimensional Scaling representations

N.M.P.Z. : National Marine Park of Zakynthos

N.O.A.A. : National Oceanic and Atmospheric Administration

N.W.M. : Nord-Ouest Méditerranée

OBIS : Ocean Biogeographic Information System

O.N.B. : Observatoire National de la Biodiversité

OSU : Observatoire des Sciences de l'Univers

PAMM : Plan d'Action pour le Milieu Marin

P.C.R. : Réaction en chaîne par polymérase

P.I.B. : Produit Intérieur Brut

P.D.F. : Portable Document Format

P.N.P.C. : Parc National de Port-Cros

P.N.U.E : Programme des Nations Unies pour l'Environnement

P.M.A : Plan d'Action pour la Méditerranée

P.M.E. : Petite et Moyenne Entreprise

PREDON : groupe de travail du C.N.R.S. sur la PREservation des DONnées

P.V.C. : Polychlorure de vinyle

q.P.C.R. : "Réaction en chaîne par polymérase" quantitative

RAC / S.P.A. : Regional Activity Center for Specially Protected Area

R.D.A.: Research Data Alliance

REBENT : RÉférentiel BEnthique Méditerranéen

R.G.I. : Référentiel Général d'Interopérabilité

R.L.M. : Réseau Littoral Méditerranéen

R.N.B.B. : Réserve Naturelle des Bouches de Bonifacio

R.S. : Red Sea

R.S.G. : Réseau Survie des Gorgones

R.S.L. : Réseau de Surveillance Lagunaire

R.S.O. : Richesse Spécifique Observable

R.S.P. : Réseau de suivi des Posidonies

R.V.A. : Rapid Visual Assessment

Sémandiv : SEMANtique de la bioDIVersité (G.D.R.)

S.E.S. : Systèmes Socio Écologiques

S.G.M.A.P. : Secrétariat Général pour la Modernisation de l'Action Publique

SHOM : Service Hydrographique et Océanographique de la Marine

S. I. C. B. : Society for Integrative and Comparative Ecology

S.I.G : Système d'Information Géographique

S.I.N.P. : Système d'Information sur la Nature et les Paysages

S.I.P. : Service Informatique de Pytheas

SMOS : Soil Moisture and Ocean Salinity

SOMLIT : Service d'Observation en Milieu LITtoral

S.P.E. : Systèmes pour l'Environnement, écosystèmes côtiers

S.P.N. : Service du Patrimoine Naturel MNHN

SUCCES (Journées) : rencontres Scientifiques des Utilisateurs de Calcul intensif, de Cloud Et de Stockage

T.D.W.G. : Taxonomic Databases Working Group [Biodiversity Information Standards]

U.M.R. : Unité Mixte de Recherche

U.M.S. : Unité Mixte de Service

U.M.T. : Unité Mixte Technologique

UNEP : voir P.N.U.E.

U.P.M.C. : Université Pierre et Marie Curie (Paris 6)

U.R. : Unité de Recherche

U.U.I.D. : Universally Unique IDentifier

U.V.C.: Underwater Visual Census

V.I.F.C. : Vertical, Incliné (Inclined), plat (Flat), en surplomb (Ceiling)

VIGI-GEEK : VIsualisation of Graph In transdisciplinary Global Ecology, Economy and Sociology data-Kernel

W.F.D. : Water Framework Directive

W.G.S. 84 : World Geodetic System 1984

Z.E.E : Zone Economique Exclusive

ZNIEFF : Zone naturelle d'Intérêt Ecologique, Faunistique et Floristique

Ressources

A.C.O. : http://aconnectedocean.sciencesconf.org/

A. F. B. : www.aires-marines.fr

AgroPortal LIRMM : http://agroportal.lirmm.fr/

Algaebase : http://www.algaebase.org/

ARCHIMER Marine protected areas and artificial reefs: A review of the interactions between management and scientific studies / http://archimer.ifremer.fr/doc/00000/397/

ARMS: https://pifsc-www.irc.noaa.gov/cred/survey_methods/arms/

AFUL : https://aful.org/gdt/interop

B. D. A. 2016 : https://bda2016.ensma.fr/index.html

Biodiversity Data Journal : http://bdj.pensoft.net/

Cargocycling : http://cargocycling.org/category/riding_type/family-cycling

CartOmer : http://cartographie.aires-marines.fr/?q=node/43

C.A.T.D.S. : https://www.catds.fr/Presentation

C.D.B. (fr) : https://www.cbd.int/doc/legal/cbd-fr.pdf

C.D.B. (en) : www.cbd.int/doc/decisions/cop-10/cop-10-dec-11-en.pdf

CESAB Thesaurus of ecological observations : http://thesaurus.cesab.org/ThesauformCesab/home

CIESM : 41st CIESM Congress : http://ciesm.org/marine/congresses/Kiel.htm

CIGESMED : www.cigesmed.eu

*CIGESMED Structural descriptors on species ans taxa : http://www.cigesmed.eu/Structural-descriptors-on-species

C. N. R. S. Restitution du défi Enviromics 2016 : http://www.cnrs.fr/mi/spip.php?article881

Commission européenne : https://ec.europa.eu/commission/index_fr

COMPLEXIX 2017 : http://www.complexis.org/?y=2017

Convention de Rio (1992) : https://www.cbd.int/convention/text/

Coralligène par l'INPN : https://inpn.mnhn.fr/habitat/recherche/libelle/corallig%C3%A8ne/

Decision adopted by the conference of the parties to the convention on biological diversity at its tenth meeting : www.cbd.int/doc/decisions/cop-10/cop-10-dec-11-en.pdf

Décret n° 2002-1187 du 12 septembre 2002 portant publication de la convention sur l'accès à l'information, la participation du public au processus décisionnel et l'accès à la justice en matière d'environnement (ensemble deux annexes), faite à Aarhus le 25 juin 1998 : https://www.legifrance.gouv.fr/eli/decret/2002/9/12/MAEJ0230045D/jo/texte

DEVOTES : http://www.devotes-project.eu/

DMP OPIDoR : https://dmp.opidor.fr/

E.C.S.A. : http://ecsa.citizen-science.net

E.G.I. : http://www.egi.eu/

E.G.I. “design your e-infrastructure” workshop : https://indico.egi.eu/indico/event/2895/

E. M. B. R. C. : http://www.embrc-france.fr/fr

E.M.B.S. : 51st European Marine Biology Symposium : http://www.embs51.org/

EMODnet : http://www.emodnet.eu/

E.P.D. : http://www.europeanpollendatabase.net/index.php

EUR-Lex : http://eur-lex.europa.eu/legal-content/FR/ALL/?uri=CELEX:32007L0002

Force 11 : https://www.force11.org/group/fairgroup/fairprinciples

GBIF : https://www.gbif.org

GBIF Introduction to sampling-event data : https://www.gbif.org/sampling-event-data

Gephi : https://gephi.org/

GEO : http://www.earthobservations.org/geo_community.php

GEO Data Management Principles Implementation Guidelines : https://www.earthobservations.org/documents/geo_xii/GEO-XII_10_Data%20Management%20Principles%20Implementation%20Guidelines.pdf

GRAMINÉES (site de MaDICS) : http://www.madics.fr/actions/actions-en-cours/graminees/

HAL : https://cv.archives-ouvertes.fr/

Hydrobio DCE : https://hrydrobio-dce.cemagref.fr

IEEE 6th Conference on Technologies for Sustainability Technologies that contribute to sustainability in all applications affecting human life : http://sites.ieee.org/sustech/

iEMSs 2016 : http://www.iemss.org/sites/iemss2016/

IndexMed premier séminaire "Interopérabilité des bases de données en écologie" : http://www.indexmed.eu/-Premier-seminaire-Interoperabilite-.html

IndexMed deuxième séminaire "Méthodes et outils pour la fouille de données hétérogènes et multi-sources en écologie" http://www.indexmed.eu/-Deuxieme-seminaire-Methodes-et-.html

IndexMed troisième séminaire Journées du GRAAL : “GRAphs and data mIning for environmentaL research - data, research questions and new hypotheses” https://indexmed2016.sciencesconf.org/

IndexMed quatrième séminaire “Sciences and Algorithms around Graphs in Environment and Societies” : https://indexmeed2017.sciencesconf.org

IndexMed : http://indexmed2016.sciencesconf.org/

IndexMEED prototype : http://data.imbe.fr/neo4j/

INEE : www.cnrs.fr/inee/

IndexMEED : www.indexmed.eu

I. N. P. N. habitat coralligène : https://inpn.mnhn.fr/habitat/recherche/libelle/corallig%C3%A8ne/

Instituts National de santé des États-Unis : http://rsb.info.nih.gov/ij/

INSU : http://www.insu.fr

INRA : http://www.inra.fr/

I.P.B.E.S. : https://www.ipbes.net/

ISTE Visualisation de données sous forme de graphes en archéologie. Rencontre opérationnelle des archéologues d'ArkeoGIS et des écologues d'IndexMed : https://www.openscience.fr/Data-visualisation-in-archaeology-based-on-graph-approach-Operational-meeting

jiap2016 : http://jiap2016.sciencesconf.org/

La Provence : http://sur.laprovence.com/bfmn-KDZB

Lignes directrices pour la gestion des données FAIR dans Horizon 2020 : http://www.donneesdelarecherche.fr/IMG/pdf/lignes-directrices_gestion-donnees-fair_horizon2020_version_3.0_tr-fr.pdf

LIRMM : http://www.lirmm.fr/

LITEAU : http://www1.liteau.net/index.php/agenda/colloque-liteau-janvier-2016-a-brest

LITEAU 3 (Programme) : http://www1.liteau.net/index.php/projet/liteau-iii

MaDICS : http://www.madics.fr/

*Makersite : https://global.makersite.net/login

MASTODONS : http://www.cnrs.fr/mi/spip.php?article53

Naissance du Web : https://home.cern/fr/topics/birth-web

Neo4j : https://neo4j.com/lp/try-neo4j-sandbox/

N.O.A.A. : http://www.noaa.gov/

PAMPA : https://wwz.ifremer.fr/pampa/

PATRINAT (U.M.S.) : http://patrinat.mnhn.fr/

PhotoQuad : http://www.mar.aegean.gr/sonarlab/photoquad/index.php

PhotoGrid : http://www2.hawaii.edu/~cbird/PhotoGrid/frames.htm

PREDON Workshop on Scientific Data Preservation : https://indico.cern.ch/event/338461/

PREDON website : https://www.cppm.in2p3.fr/~diaconu/concrete5.7.5.9/

Prototype d'IndexMEED : http://data.imbe.fr/neo4j/

RDA Sharing Rewards and Credit (SHARC) IG : https://www.rd-alliance.org/groups/sharing-rewards-and-credit-sharc-ig

RDA SHARC (SHAring Reward & Credit) IG Charter : https://www.rd-alliance.org/group/short-presentation-sharing-rewards-and-credit-sharc-ig/case-statement/sharc-sharing-reward

sfécologie 2016 : http://sfecologie2016.sciencesconf.org/

S. I. C. B. Annual Meeting 2016 : http://www.sicb.org/meetings/2016/index.php
SOMLIT : http://www.SOMLIT.INSU.fr
spillover (définition A.F.B.) : http://www.aires-marines.fr/Glossaire/Spillover
SUCCES (journées 2016) : https://succes2016.sciencesconf.org/resource/page/id/7
SUCCES (journées 2015) : https://succes2015.sciencesconf.org/resource/page/id/8
T. D. W. G. : www.tdwg.org
T. D. W. G. 2016 : http://www.tdwg.org/fileadmin/2016conference/documents/TDWG_Conference_program-en_US.pdf
TED conferences 2009, Tim Berners Lee https://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide?language=fr
Thesauform : http://thesaurus.cesab.org/ThesauformCesab/home
Tulip : http://tulip.labri.fr/TulipDrupal/
Wikipedia ISO 8601 : https://fr.wikipedia.org/wiki/ISO_8601
WoRMS : http://www.marinespecies.org/
Workshop: Design your e-Infrastructure : https://indico.egi.eu/indico/event/2895/
WP4 de CIGESMED : http://www.cigesmed.eu/-Module-de-travail-4-outils-

Annexes

Annexe 1 : Articles

Les références précédées du signe **(§)** sont annexées à la thèse

Articles issus du travail de thèse soumis à / acceptés dans des revues à comité de lecture

2018

(§9) Romain David, Maria C. Uyarra, Susana Carvalho, Holger Anlauf, Angel Borja, *et al.* Lessons from photo analyses of Autonomous Reef Monitoring Structures, as tools to detect geographical, spatial, and environmental effects" 2018 *submitted, Marine Pollution Bulletin.*

(§8) Romain David, Anna Cohen Nabeiro, Jean-Pierre Féral, Aurélie Delavaud, Anne-Sophie Archambeau, *et al.*, 2018 *accepted*, Bilan des journées du GRAAL 2016 et avenir de l'utilisation des graphes en écologie, *Nature, Science et Société, 18-3*.

2017

(§7) Víctor Méndez Muñoz, Anna Cohen-Nabeiro, **Romain David**, Vicente Ivars Camáñez, Alfons Nonell-Canals, *et al.*. Analysis on the Graph Techniques for Data-mining and Visualization of Heterogeneous Biodiversity Data Sets. Complexis 2017, Apr 2017, Porto, Portugal. pp.144 - 151, 2017, Proceedings of the 2nd International Conference on Complexity, Future Information Systems and Risk (COMPLEXIS 2017). 〈http://www.complexis.org/?y=2017〉. 〈10.5220/0006379701440151〉. 〈hal-01541140〉

(§6) Romain David, Loup Bernard, Cyrille Blanpain, Alrick Dias, Jean-Pierre Feral, *et al.*. Visualisation de données sous forme de graphes en archéologie. Rencontre opérationnelle des archéologues d'ArkeoGIS et des écologuesd'IndexMed . *Digital Archaeology*, iste open science, 2017, 17-1 (1), 〈https://www.openscience.fr/Data-visualisation-in-archaeology-based-on-graph-approach-Operational-meeting〉. 〈hal-01617580〉

2016

(§5) Romain David, Jean-Pierre Feral, Anne Archambeau, Nicolas Bailly, Cyrille Blanpain, *et al.*. IndexMed projects: new tools using the CIGESMED DataBase on Coralligenous for indexing, visualizing and data mining based on graphs. Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.). 8th International Congress on Environmental Modelling and Software, Jul 2016, Toulouse, France. Brigham Young University BYU Scholars Archive, International Environmental Modelling and Software Society (iEMSs) 8th International Congress on Environmental Modelling and Software Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.) http://www.iemss.org/society/index.php/iemss-2016-proceedings, 8th, pp.32, 2016, *International Congress on Environmental Modelling and Soft ware*. http://scholarsarchive.byu.edu/iemssconference/2016/?utm_source=scholarsarchive.byu.edu%2Fiemssconference%2F2016%2FStream-C%2F32&utm_medium=PDF&utm_campaign=PDFCoverPage. 〈hal-01425559〉

2015

(§4) Romain David, Jean-Pierre Feral, Sophie Gachet, Alrick Dias, Cyrille Blanpain, *et al.*. A first prototype for indexing, visualizing and mining heterogeneous data in Mediterranean ecology: within the IndexMed consortium interdisciplinary framework. 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Nov 2015, Bangkok, Thailand. IEEE Explore, 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 232-239, 2015, 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). 〈http://ieeexplore.ieee.org/document/7400571/〉. 〈10.1109/SITIS.2015.119〉. 〈hal-01433600〉

(§3) Romain David, Sophie Dubois, Zinovia Erga, Dorian Guillemain, Laure Thierry de Ville d'Avray, *et al.*, (2015) "CIGESMED*'s protocol and network (Coralligenous based Indicators to Evaluate and Monitor the « good ecological status » of the MEDiterranean coastal waters)", CIGESMED Project. Pages 828-843 ; CNR-IBIMET Florence (Italy), December 2014, ISBN : 978-88-95597-19-5 (Fifth Symposium Monitoring of Mediterranean coastal areas : problems and measurement techniques Livorno (Italy) 17-19 june 2014)

2014

(§2) Romain David, Christos Arvanitidis, Melih Ertan Çinar, Stéphane Sartoretto, A Doğan, *et al.*. CIGESMED habitat's characterization: a simple and reusable typology at the Mediterranean scale. 2nd Mediterranean Symposium on the conservation of Coralligenous & other Calcareous Bio-Concretions, Oct 2014, Portorož, Slovenia. Proceedings of the 2nd Mediterranean Symposium on the conservation of Coralligenous & other Calcareous Bio-Concretions. 〈hal-01620541〉

(§1) Romain David, Christos Arvanitidis, Melih Ertan Çinar, Stéphane Sartoretto, Sophie Dubois, *et al.*. CIGESMED protocols: CIGESMED Protocols : how to implement a multidisciplinary approach on a large scale for coralligenous habitats surveys. RAC/SPA 2nd Mediterranean Symp. on the Conservation of coralligenous and other calcareous bio-concretions, Oct 2014, Portorož, Slovenia. 〈10.13140/2.1.1895.0086〉. 〈hal-01620550〉

Jean-Pierre Féral, Christos Arvanitidis, Anne Chenuil-Maurel, Melih Ertan Çinar, **Romain David**, *et al.*. CIGESMED: Coralligenous based indicators to evaluate and monitor the "good environmental statut" of the Mediterranean coastal waters, a SeasEra project. 2nd Mediterranean Symposium on the conservation of Coralligenous & other Calcareous Bio-Concretions, Oct 2014, Portorož, Slovenia. Proceedings of the 2nd Mediterranean Symposium on the conservation of Coralligenous & other Calcareous Bio-Concretions. 〈hal-01620607〉

Autres publications en relation avec le travail de thèse

Aurélien De Jode, **Romain David**, Anne Haguenauer, Abigail E. Cahill, Zinovia Erga, Dorian Guillemain, Stéphane Sartoretto, Caroline Rocher, Marjorie Selva, Line Legall, Jean-Pierre Féral, Anne Chenuil, (2018, submitted), Multiple cryptic species, spatial and ecological differentiation in a major builder of coralligenous habitats, *Molecular Ecology*.

Laure Thierry de Ville D 'Avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Feral (2018 submission accepted), Application of the Ecosystem Service Concept to a Local-Scale: The Cases of Coralligenous Habitats in the North-Western Mediterranean Sea, *Marine Pollution Bulletin.*

Bénédicte Madon, **Romain David**, René Garello, Linwood Pendleton, Ronan Fablet. Strike-Alert: Towards Real-time, High Resolution Navigational Software for Whale Avoidance.

SUSTECH 2017 : 5th annual IEEE Conference on Technologies for Sustainability, Nov 2017, Phoenix, United States. 2017, 〈http://sites.ieee.org/sustech/〉. 〈hal-01623903〉

Romain David, Jean-Pierre Féral, Sophie Archambeau, Fanny Arnaud, David Auber, *et al.*. IndexMEED cases studies using "Omics" data with graph theory. Biodiversity Information Science and Standards, Sofia : Pensoft Publishers, 2017, *TDWG Proceedings 2017,* 1 (2), pp.340-361. 〈10.3897/tdwgproceedings.1.20740〉. 〈hal-01761535〉

Abigail E. Cahill, Aurelien De Jode, Sophie Dubois, Zoheir Bouzaza, Didier Aurelle, *et al.*. A multispecies approach reveals hot spots and cold spots of diversity and connectivity in invertebrate species with contrasting dispersal modes. *Molecular Ecology, Wiley,* 2017, 26 (23), pp.6563-6577. 〈10.1111/mec.14389〉. 〈hal-01681650〉

Stéphane Sartoretto, Thomas Schohn, Carlo Bianchi, Carla Morri, Joaquim Garrabou, *et al.*. An integrated method to evaluate and monitor the conservation state of coralligenous habitats: The INDEX-COR approach. *Marine Pollution Bulletin, Elsevier,* 2017, 〈10.1016/j.marpolbul.2017.05.020〉. 〈hal-01541141〉

Roberto Danovaro, Laura Carugati, Berzano Marco, Abigail E. Cahill, Susana De Carvalho Spinola, *et al.*. Implementing and Innovating Marine Monitoring Approaches for Assessing Marine Environmental Status. *Frontiers in Marine Science, Frontiers Media*, 2016, 3, pp.213. 〈10.3389/fmars.2016.00213〉. 〈hal-01448726〉

Vasilis Gerovasileiou, Thanos Dailianis, Emmanouela Panteri, Nikitas Michalakis, Giulia Gatti, *et al.*. CIGESMED for divers: Establishing a citizen science initiative for the mapping and monitoring of coralligenous assemblages in the Mediterranean Sea. *Biodiversity Data Journal, Pensoft*, 2016, 54 (e8692), 〈http://bdj.pensoft.net/〉. 〈10.3897/BDJ.4.e8692〉. 〈hal-01392025〉

Karina Sevastou, Papadopoulou N., Smith C., Teixeira H., Piroddi C., *et al.*, Defining keystone species in European regional seas: what are the candidates for the Mediterranean? In: *11th Panhellenic Symposium on Oceanography & Fisheries «Aquatic Horizons: Challenges & Perspectives*. Mytilene, Lesvos Island, Greece, 13-17 May 2015. Athens: H.C.M.R., 553-556.

Stéphane Sartoretto, **Romain David**, Didier Aurelle, Anne Chenuil-Maurel, Dorian Guillemain, *et al.*, An integrated approach to evaluate and monitor the conservation state of coralligenous bottoms: the INDEX-COR method. *Conference: 2nd Mediterranean Symposium on the conservation of Coralligenous & other Calcareous Bio-Concretions*, Oct 2014, Portorož, Slovenia. 2014, 〈10.13140/2.1.3180.6405〉. 〈hal-01620618〉

Romain DAVID, ARVANITIDIS C., ÇINAR M.E., SARTORETTO S., DOGANA., DUBOIS S., ERGA Z., GUILLEMAIN D., THIERRY DE VILLE D'AVRAY L., ZUBERER F., CHENUIL A., FERAL J.-P.
CNRS- IMBE: Mediterranean Institute of Biodiversity and marine and terrestrial Ecology, Station Marine d'Endoume, Marseille (CNRS, AMU, IRD, Avignon Univ.)
E-mail: romain.david@imbe.fr [www.cigesmed.eu]

CIGESMED PROTOCOLS: HOW TO IMPLEMENT A MULTIDISCIPLINARY APPROACH ON A LARGE SCALE FOR CORALLIGENOUS HABITATS SURVEYS

Abstract
The European program CIGESMED addresses the Good Environmental Status of the coralligenous habitats. Its implementation on the field is firstly attempted by 4 protocols to be applied in France, Greece and Turkey. They have been tested in Marseille's region, since early 2014. These protocols are the following: (i) cartography of chosen coralligenous sites, (ii) spatial variability analysis by means of photo-quadrats and image processing, (iii) population genetics study of two common biobuilding species that may be cryptic (the bryozoan Myriapora truncata, and the rhodophyta Lithophyllum cabiochiae), and (iv) metagenomic approach of benthic species. The ultimate aim of these protocols is to link the results from the population genetics analysis and the spatial variability analysis to the sites' features thanks to the cartography. First results suggest that different clades exist for both complex of the previous species. Cartography forshadows models of repartition for species assemblages; they will then be compared between regions in the second part of the project.

Key-words: Coralligenous habitats, monitoring protocols, cartography, photo-quadrats, population genetics.

Introduction
The term "coralligenous", meaning coral producer, was first used by Marion in 1883 to describe the hard bottoms called *broundo* by the fishermen from Marseilles. But the meaning of this term is nowadays different, including different types of hard bottoms communities in the Mediterranean Sea. In 2006, Ballesteros recommends to use the terms "coralligenous habitats" since there exist many distinct types of coralligenous habitats.
In the current European legislation context, coralligenous habitats are considered habitats of "community interest" (Habitats Directive 92/43/CEE, habitat code: 1170-14) and should be shortly promoted as "priority" habitat. It is currently considered as the second "hotspot" of biodiversity in the Mediterranean Sea (*Posidonia* meadow being the first one), with than 1,600 species hosted by these habitats (Ballesteros, 2006). Since few years, the EU Marine Strategy Framework Directive (MSFD) requires that each country develops a strategy and an action plan in order to reach and maintain a "Good Environmental Status" for its marine habitats.
There are currently a few programs and networks for the monitoring of coralligenous habitats. CIGESMED (Coralligenous Indicators based to Evaluate and Monitor the "Good Environmental Status" of the Mediterranean coastal waters) involves three countries (France, Greece and Turkey), on a collaborative effort from 2013 to 2016. CIGESMED objectives are (1) to fulfil the key gaps in the current scientific knowledge,

(2) to enhance the knowledge on coralligenous populations by deciding on reference states and setting up a network of Mediterranean experts (for the production of long term series), (3) to monitor networks, locally managed and coordinate them on a regional scale, (4) to test population genetics criteria as tools to monitor the GES of the coastal Mediterranean Sea, (5) to implement a "citizen science" network and (6) to use trees of knowledge as tools to sort, organize and illustrate the large heterogeneous sets of produced data. This work includes habitats cartography, population genetics studies to understand species relations and dispersal potential, and the setting up of a monitoring protocol.
A phase of inter-calibration of methods/material/operators has been implemented. This step is essential in order to evaluate the variability related to these experimental parameters. It allow comparability of results obtained by different underwater protocols. Moreover, this phase of test helps to select the best protocol to apply (the easiest, and most reliable), depending on the habitats types. The next phase will be the study of natural variability inter-site or intra-sites.

Material and methods

Observations and cartography of coralligenous habitats
Intercalibration methods
The French studied sites are located in Marseilles Bay. They are transects of 10 meters long at 28 meters depth. To date, three variables of the protocol implementation have been studied: the sampling method, the quality of the camera, and the level of knowledge of operators in charge to identify species. The protocol was implemented as follow. Divers made photo-quadrats using a frame of 50 cm by 50 cm. The pictures were analysed by operators using the software Photoquad® (Trygonis & Sini 2012). Hundred points were distributed by stratified randomization. Then the operator assigned each point to one category among these three: (i) higher taxa (such as phyla, orders), (ii) abiotic, (iii) indeterminate. In the first category (i) the sub-categories are lower taxa (such as genus or species). The second category (ii) is subdivided into four sub-categories: sediment, bare rock, organic detritus or debris. In the third one (iii), there are three sub-categories: fuzzy image, shadow/hole, and unidentified taxon.

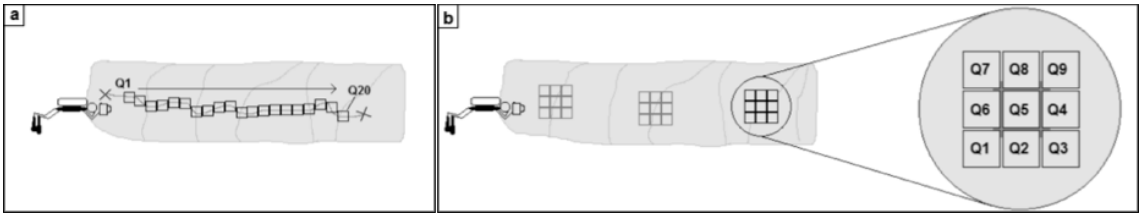


Fig. 1: The two types of transects tested. Linear transect (a): 20 photo-quadrats are taken at a given depth, located by permanent marks. Random patch transect (b): 3 groups of 9 photo-quadrats are taken, following the indicated numbers scheme from 1 to 9, several times at a same depth.

The two sampling methods compared were: (i) the permanent linear transect and (ii) the random patches transect (fig.1). To compare the sampling methods 20 photo-quadrats done by the permanent transect method were compared to 18 photo-quadrats (2 patches) done by the random patches method.
The second variable studied was the quality/performance of the camera. Two cameras have been compared: (i) a camera of medium quality and (ii) a camera of high quality. The models used are (i) GoPro®, and (ii) Nikon® D300s. To compare both cameras, two sets of 8 photo-quadrats done on a permanent transect at the exact same place, were used.
The third variable studied was the level of knowledge of the operators in charge to identify

the taxa on photo-quadrats. Two levels of operators were compared: (i) novice and (ii) experienced. The set of operators participating were: one novice and two experienced operators. Each of them analysed separately the first 5 photo-quadrats of the transect (series 1). Then they met to exchange their results (taxon identification) about this first set and re-do the identification work together to produce "validated data". Again, they studied separately the 2nd series of 5 photo-quadrats. Finally, they met again to exchange their knowledge and produce "validated" data on this 2nd series. Therefore, there is an iteration loop of 4 subsequent cycles to analyse the 20 photo-quadrats of the permanent transect.

Profile characterization and cartography

For each site, two depths were sampled around 28m deep (±1m), and around 45 m deep (±1m). Samples were collected along transects cut into segments of 5m long and 1m wide.

Population genetics studies

The aim here is to understand the population structure of the coralligenous species by studying the intraspecific diversity of demes and their connectivity. For the study of the target species, which are living throughout the Mediterranean Sea, we will use the barcoding method consisting in sequencing a part of the mitochondrial gene COI. We will eventually complete with other alternative or complementary markers. The objective is to test whether the previously-mentioned taxa consist of cryptic species in the Mediterranean.

The two chosen species are (i) the erect-like and tree-like bryozoan *Myriapora truncata*, and (ii) a complex of bioconstructing coralline algae *Lithophyllum stictaeforme /cabiochiae*. Both are identifiable *in situ*. They were selected since they have a widespread occurrence in the coralligenous communities, on all the facies and at all depths (even at very low irradiance). Standard PCR protocols is used to amplify COI fragments for both the bryozoan and the red alga, as well as another marker that is not from the mitochondrial genome, for each species (detailed methods to be published): an intron for *M. truncata* (Chenuil *et al.*, 2010; Gérard *et al*., 2013) and a chloroplast marker for *Lithophyllum sp.* (Broom *et al*., 2008). PCR products are sent to the industry for DNA sequencing, then after alignment, haplotype network reconstruction is made using the Median Joining Network software (Bandelt *et al*., 1999).

Results

Photo-quadrats inter calibration

The preliminary results are presented on figure 3. It shows that at the phylum level, the two methods give equivalent results. Differences of headcounts are significant only in the phylum *Porifera*. Thus results are comparable.

To compare both cameras, the pictures were not taken at the exact same time; the species *P. clavata* disturbed the observations as it had its polyps spread out or not, depending on the set. To release the experiment from this disturbance, all observations of *P. clavata* were removed from both sets. Figure 4 shows that at the species level both cameras give equivalent results. But the camera of high quality made it possible to reduce the number of "indeterminate" and these observations were assigned to other categories, in the majority at the genus level.

The preliminary results on certain categories are shown on the figure 5. After only one exchange between the three operators, the novice improves a lot his/her capacity of identification. For some categories, the level of knowledge of the three operators gets quickly homogenized: for instance for the categories *Cnidaria* and *Porifera*.

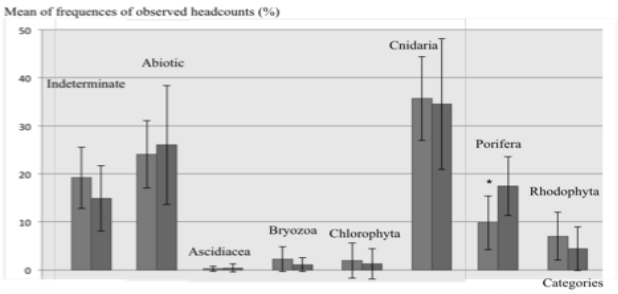


Fig. 3: Comparison of results given by the permanent transect method (light grey) and the random patches method (dark grey). *significant difference according to the Mann-Whitney- Wilcoxon test with a 5 % risk.

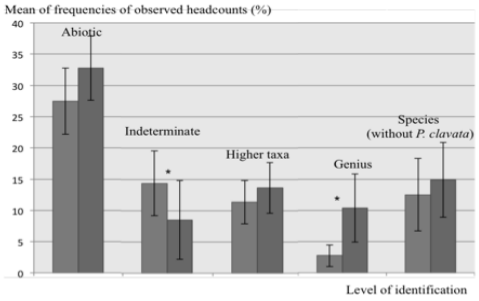


Fig. 4: Comparison of results obtained with a GoPro® (light grey) and a Nikon D300® (dark grey). The star marks a significant difference according to the Mann- Whitney-Wilcoxon test with a risk of 5 %.

They are in most case very hard to identify on photography. From this work, it appears that cnidarian species are easiest to identify on photography by beginners while poriferan species are mostly very hard to identify without specific training.

Analysis of the preliminary cartography results

Using Hierarchical Ascendant Classification (HAC) and Correspondence Factorial Analysis (CFA) on all processed data allowed: (i) to group species according to values of orientation, slope and roughness, and (ii) to pool profile parameters according to species observed per segment of transect. First results shows that cluster of species are better supported using combined factors of slopes and orientations. Roughness less explain the different groups. Using Factorial Correspondence Analysis and Ascending Hierarchical Classifications, preferential profiles of coralligenous species can be determined: bryozoans, encrusting and foliose red algae and foliose green algae occur preferentially on horizontal walls South-oriented with large roughness. *Eunicella cavolinii* is mainly present on inclined walls facing West/North-West with low or medium roughness. Porifera and *P. clavata* are preferentially present on vertical walls facing North. *Codium* genus and encrusting green algae are more present on walls facing South- East and North-East with tiny roughness.

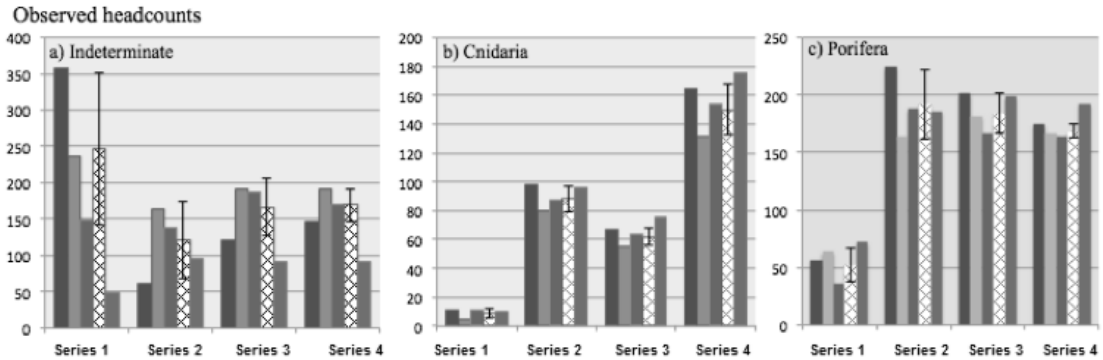


Fig. 5: Comparison of number of headcounts (example for 3phylum) observed by different operators: one novice (1st bar from left) and two experienced (2nd and 3rd bars from left). Mean and standard deviation between the three operators are in white cross-brace. The identifications validated by the three operators are represented by the bar on the right.

A "profile" is thus a combination of orientation parameters combined with inclination and roughness. Further detailed study will be conducted on associations of profile features to determine the preferential profiles of different coralligenous communities.
The cartography details the environmental profiles: depth, orientation, slope, roughness of the coralligenous wall, and main stands. The aim of this protocol is to identify the main species assemblages according to the profile type. The inter-calibration phase of the protocol in order to test the material and methods is done. The study of the variability of the results, linked to the observers and operators implementation, shows which metrics are the most robust. This phase was preliminary in order to make the results workable whatever the country of implementation. The running step permits to identify accessible metrics that are relevant, reliable and efficient to explain the "natural" spatial variability regarding the environmental context. The genetic study of *Myriapora* and *Lithophyllum* used data of sequences and molecular markers to answer taxonomic questions about these species complex and descriptions of connectivity patterns. The preliminary results are promising: the genetic study of *Myriapora* and *Lithophyllum* highlights different clades which strongly suggest that there are most probably mixtures of different species.

Discussion and conclusion

Assessment of the environmental status

The assessment consists of the analysis of the coralligenous megabenthic assemblages by means of direct observation, photographic/video surveys (which are directly influenced by competencies and experience of the operators).
The study of both sampling methods shows that observations made at a phylum level of identification are comparable if they are made according to the one or the other method. To compare Porifera observations from one site to another, it would be recommended to implement the same method in both sites, as it's shown that there might be significant difference of headcounts according to the applied method. As the random patch method is easier to perform, it should be recommended. Concerning the effect of the quality of the camera on the observations made, it has been proven that the medium quality camera is sufficient to identify as much species as the high quality camera. Difference between the two is observed for species difficult to differentiate at the genus level. Indeed, the high quality camera enables to give a taxonomic level at some individuals that were indeterminate with the medium quality camera. The study of the operator's knowledge shows that the discussion between operators enables novices to quickly improve their capacity of identification for a given set of taxa. Discussion is very useful at the beginning, and then operators reach a step, and would need a proper training to progress, if more precise identification is needed.

Population studies

The objective of the cartography goes beyond than mapping habitats; it provides information about environmental profiles that will be used to understand species preferences. Population genetics approach which will complete the analyses is essential to investigate species diversity, population structure and connectivity.
The first results about genetic differentiation of demes from distinct localities for each taxon illustrate the fact that gene flow (migration) is limited even at the small scale of Marseilles region for those important coralligenous builders. Genetic barrier were previously evidenced for other species as different as the mysid *Hemimysis margalefi* (Lejeusne *et al*., 2006) or the irregular sea urchin *Echinocardium cordatum* (Egea *et al*., 2011).

It may also be linked to ecological conditions (distribution of divergent groups of each species depending on currents and ecologic profiles to be carried out).
This preliminary work yet enables to better understand the importance of the skill and training of operators (human factor rarely taken into account) and of the implemented sampling. Indicators, from communities to infra-specific level, will be co-constructed by scientists and PMA managers, and through the implementation of a "citizen science" network. The outcome will be an integrative assessment of the GES within the MSFD. To build the network, the community is developing a metadata catalogue and shared typologies; we are working on (i) harmonization of data collection methods and normalization of data access (European standards) (ii) initiation/animation of thematic network about coralligenous habitats in Mediterranean Sea gathering all competent actors. This network is meant to be perennial, open, and fully decentralized (to allow for continuous update) at local, regional, national and international scales. This organisation will permit data diffusion and upper accessibility and will ensure continuous improvement.

Acknowledgements
We are grateful to the organizing committee who invites us to present the first results of the project. We thank A. Haguenauer, S. Chenesseau, F. Zuberer and all the underwater diving team for their participation to the sampling in Marseilles and T. Dailianis for samples from Crete. We thank M. Verlaque who helps us for the identification of Corallinales and A. Le Gall for helpful advices. A special thanks to the SIP (Computing Service of the OSU Pytheas) who provides all the hardware part and assistance for developments. This project was funded by the SeasEra program CIGESMED (CNRS- ANR convention n° 12-SEAS-0001-01).

Bibliography
BALLESTEROS E. (2006) - Mediterranean coralligenous assemblages: a synthesis of present knowledge. *Oceanogr. Mar. Biol.: Ann. Review*, 44: 123-195.

BANDELT H. J., FORSTER P., ROH*L A. (1999)* – Median-joining networks for inferring intraspecific phylogenies, *Mol. Biol. Evol*. 16, 37-48.

BROOM J. E. S., HART D. R., FARR T. J., NELSON W. A., NEILL K. F., HARVEY A. S., WOELKERLING W. J. (2008) - Utility of psbA and nSSU for phylogenetic reconstruction in the Corallinales based on New Zealand taxa, *Mol. Phylogenet . Evol*. 46, 958 – 973.

CHENUIL A. , HOAREAU T. B., EGEA E., PENANT G., ROCHER C., AURELLE D., MOKHTAR-JAMAI K., BISHOP J. D. D., BOISSIN E., DIAZ A., KRAKAU M., LUTTIKHUIZEN P. C., PATTI F. P. , BLAVET N., MOUSSET S. (2010) – An essicient method to find potentially universal population genetic markers, applied to metazoans, *BMC Evol. Biol.* 10, 276.

EGEA E (2011) - *H*istoire évolutive, structures génétique, morphologique et écologique comparées dans un complexe d'espèces jumelles: *Echinocardium cordatum* (Echinoidea, Irregularia). Aix-Marseille Université, Marseille.

GERARD K., GUILLOTON E., ARNAUD-HAOND S., AURELLE D., BASTROP R., CHEVALDONNE P., DERYCKE S., HANEL R., LAPEGUE S., LEJEUSNE C., MOUSSET S., RAMSAK A., REMERIE T., VIARD F., FERAL J.-P., CHENUIL A. (2013) – PCR survey of 50 introns in animals: Cross-amplification of homologous EPIC loci in eight non-bilaterian, protostome and deuterostome phyla, *Mar. Genomics.* 12, 1 – 8.

LEJEUSNE C., CHEVALDONNE P. (2006) - Brooding crustaceans in a highly fragmented habitat: the genetic structure of Mediterranean marine cave-dwelling mysid populations. *Mol. Ecol.* 15, 4123–40

MARION A. F. (1883) – Esquisse d'une topographie zoologique du Golfe de Marseille. *Ann. Mus. Hist.* Marseille, Marseille.

TRYGONIS V. & SINI M. (2012) – Photoquad: a dedicated seabed image processing software, and a comparative error analysis of four photoquadrat methods. J. Esp. Mar. Biol. Ecol. 424, 99-108.

Romain DAVID, ARVANITIDIS C., ÇINAR M.E., SARTORETTO S., DOĞAN A., DUBOIS S., ERGA Z., GUILLEMAIN D., THIERRY DE VILLE D'AVRAY L., ZUBERER F., CHENUIL A., FERAL J.-P.
CNRS-IMBE: Mediterranean Institute of Biodiversity and marine and terrestrial Ecology, Station Marine d'Endoume, Marseille (CNRS, AMU, IRD, Avignon Univ).
Email: romain.david@imbe.fr

CIGESMED HABITAT'S CHARACTERIZATION: A SIMPLE AND REUSABLE TYPOLOGY AT THE MEDITERRANEAN SCALE

Abstract
The so-called coralligenous makes Mediterranean marine habitats that are of the most important in terms of complexity and biodiversity. Coralligenous is formed by the development of several types of communities where bio-constructor, bio-erodor engeneer and "habitat" species interact to build complex structures. The European program CIGESMED studies the Good Environmental Status (G.E.S.) of these habitats. Several protocols are implemented, in particular the cartography of abiotic context, and species observation by means of photo-quadrats. The cartography inventories the profiles types of the coralligenous sites with as robust as possible categories: depth, orientation, slope, roughness, and main coralligenous stands.
The objective is to establish a link between the species occurrence features, and the profiles features in order to understand the "natural" spatial variability of coralligenous habitats.

Key-words: Coralligenous habitats, protocols, cartography, photo-quadrats, contextualization.

Introduction
Many studies on coralligenous habitats have been published (Marion, 1883, Laubier, 1966; Hong, 1980). The coralligenous milieu hosts complex assemblages of more than 1,600 species. (Ballesteros, 2006). A better understanding of the variability of this habitat demands estimation/measurement, at a large spatial scale, using simple methods and variables that can explain the species assemblages.

Material and methods
Due to their topography and to their complexity, there are only a few accurate maps of coralligenous habitats. CIGESMED program is experiencing a way to symbolise them by means of easily recognizable signs. Based on the estimation of the diversity and abundance of species observed during field surveys, coralligenous habitats cartography should also take into account the profiling parameters (orientation, slope, roughness, and major covers) that favour one or another taxon. This study of the variability of the coralligenous habitats structure is made on small islands of Marseilles' bay at a constant depth of 28 (± 1) meters. Observations were done on different types of sites. Either around small islands and shoals, or along coastline with all orientations represented. Samples were collected along transects cut into segments of 5m long and 1m wide. To define the profile of each segment, a common typology has been applied. This typology can be apply easily by the under-water diver with anatomic references (finger(s), fist, head, shoulders) completed by main species covers. For data processing, the frequency of species observation for each segment was calculated for each profile setting.

Results

Using Hierarchical Ascendant Classification (HAC) on all processed data allowed: (i) to group species according to values of orientation, slope and roughness, and (ii) to pool profile parameters according to species per segment. This HAC show four groups of profile parameters (Fig. 1).

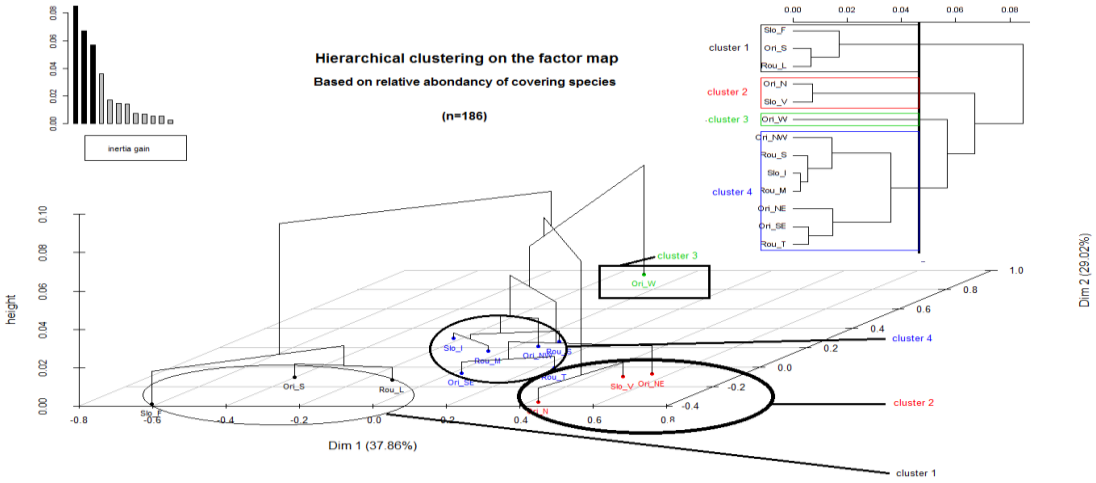


Fig.1: The two dimension HAC show a good repartition of species assemblages according to à first gradient (37,86% of the variability explanation) with two opposite clusters (cluster 1 and cluster 2), the second one for intermediates values of each categories.

Discussion and conclusion

For this first set of results, we can consider that the first axe of variability corresponds to the factor light, with two clusters grouping at opposite conditions of light. The second axis may pool factors according to light variability and currents (complementary studies are in progress). The preferences of species assemblages for different associations of parameters will permit to propose a coralligenous typology, and understand what the differences of preferences at the Mediterranean scale are. A "profile" is thus a combination of orientation, inclination and roughness parameters. Further detailed study will be conducted on associations of profile features to determine the preferential profiles of different coralligenous communities. The analysis should enable to identify metrics that are relevant, reliable, and efficient to explain this "natural" variability, taking in account the variability at the Mediterranean scale and a large panel of observers with different skills level.

Acknowledgements

Works are supported by: France - CNRS - ANR convention n°12-SEAS-0001–01 / LIGAMEN – ANR convention n°12-SEAS-0001-02 / IFREMER - ANR convention n°12-SEAS-0001-03, Greece - GSRT 12SEAS-2-C2, Turkey - Tübitak contract n°112Y393.

Bibliography

BALLESTEROS E. (2006) - Mediterranean coralligenous assemblages: a synthesis of present knowledge. Oceanogr. Mar. Biol.: Ann. Review, 44: 123-195.

HONG, J. -S. (1980) - Etude faunistique d'un fond de concrétionnement de type coralligène soumis à un gradient de pollution en Méditerranée nord-occidentale (Golfe de Fos). Thèse de doctorat. Aix- Marseille II.

LAUBIER L. (1966) - Le coralligène des Albères. Monographie Biocoenotique. Ann. Inst. Océan., Paris. MARION A. F. (1883) – Esquisse d'une topographie zoologique du Golfe de Marseille. Ann. Mus. Hist. Marseille, Marseille.

CIGESMED'S PROTOCOL AND NETWORK (CORALLIGENOUS BASED INDICATORS TO EVALUATE AND MONITOR THE "GOOD ENVIRONMENTAL STATUS" OF MEDITERRANEAN COASTAL WATERS)

Romain David[1], Sophie Dubois[1], Zinovia Erga[1,5], Dorian Guillemain[1], Laure Thierry de Ville d'Avray[1], Christos Arvanitidis [2], Melih Çinar[3], Stéphane Sartoretto[4], Frédéric Zuberer[1,*], Anne Chenuil[1], Jean-Pierre Féral[1] with other contributors**

[1]Partner CNRS- IMBE: Mediterranean Institute of Biodiversity and marine and terrestrial Ecology, Station Marine d'Endoume, Marseille (CNRS, AMU, IRD, Avignon University), MIO: Mediterranean Institut of Oceanography, Marseille (CNRS, AMU, IRD, Toulon University) and SPE: Systèmes Pour l'Environnement, écosystemes côtiers. University of Corsica, Corte. (*) Institut Pythéas

[2] Partner HCMR- Hellenic Centre for Marine Research - Institute of Marine Biology, Biotechnology and Aquaculture, Institute of Oceanography and Institute of Marine biological Resources, Thalassocosmos.

[3] Partner EGE- EGE University - Faculty of Fisheries - Dokuz Eylul University [Institute of Marine Sciences and Technology & Faculty of Science] - Cela Bayar University - Department of Biology and Ministry of Forestry and Water Affairs [General Directorate of Water Management], Izmir / Ankara.

[4]Partner IFREMER- (French Research Institute for Exploitation of the Sea), La Seyne sur Mer.

[5]National Marine Park of Zakinthos, Greece

Corresponding author: romain.david@imbe.fr

** Contributors in CIGESMED : Açik Çinar S., Andral B., Aurelle D., Aysel V., Bakir K., Bellan G., Bellan-Santini D., Bouchoucha M., Bricout R., Celik C., Chatzigeorgiou G., Chatzinikolaou E., Chenesseau S., Dağli E., Dailianis T., Dimitriadis C., Doğan A., Dounas C., Egea E., Emery E., Evcen A., Faulwetter S., Gatti G., Gerovasileiou V., Güçver S.M., Issaris Y., Katağan T., Keklikoglou K., Kirkim F., Koçak F., Koutsoubas D., Marschal C., Önen M., Önen S., Öztürk B., Panayiotidis P., Pavloudi C., Pergent G., Pergent-Martini C., Poursanidis D., Ravel C., Reizopoulou S., Rocher C., Ruiton S., Salomidi M., Sarropoulou E., Selva M., Sini M., Sourbes L., Simboura N., Taşkin E., Vacelet J., Valavanis V., Vasileiadou A., Verlaque M.

Abstract – Coralligenous habitats are part of the most important Mediterranean marine ecosystems in terms of complexity and biodiversity. They provide protection, feeding and reproduction areas for more than 1600 species. This biodiversity is essential for economic activities such as fishing and scuba diving. The European program CIGESMED (ERA-NET funding), involving France, Greece and Turkey, investigates the "Good Environmental Status" of these habitats in the framework of the MFSD (Marine Strategy Framework Directive). One major objective of CIGESMED is to propose an operational, long-term, large-scale protocol to monitor coralligenous habitats in the Mediterranean Sea. This protocol is currently tested in French sites: the effects of (sampling) methods, materials and operators are studied in order to evaluate the influence of protocol implementation, and to find the easiest and most reliable procedure that could be implemented by and for a large public, also consisting of non-scientists, and would provide workable data for long-term monitoring. This protocol is based on photo-quadrats observations. The analyses are done on occurrences, relative

abundances, species associations and species' favourite environmental profile. In parallel, complementary protocols are implemented: one concerns the cartography of coralligenous habitats, and two are about population genetics of significant habitat-forming species *Myriapora truncata* (Pallas, 1766) and *Lithophyllum cabiochiae* ((Areschoug) Hauck, 1877)). All the data and results are organized as a non-centralized information system, with configurable plugins that can be installed free by all new partners.

Introduction

The term "the coralligenous", meaning coral producer, was first used by Marion in 1883 [11] to describe the hard bottoms called *broundo* by the fishermen from Marseilles [1]. Indeed, Marion thought that the red coral (*Corallium rubrum*) was indivisible from these hard biogenic bottoms. Presently the word "coralligenous" sets off debate, because it is now known that the presence of red coral on this type of bottom is neither inevitable, nor exclusive. Yet, coralligenous habitats with high density of *Corallium rubrum* are just one of the possible type of these habitats [16]. In 2006, Ballesteros [1] recommends to use the terms "coralligenous habitats" since there are many type of coralligenous habitats, and not only one.

In the current European context, coralligenous habitats are considered habitats of "community interest" (Habitats Directive 92/43/CEE, habitat code: 1170-14) and should be shortly promoted as "priority" habitat. It is currently considered as the 2nd "hotspot" of biodiversity in the Mediterranean Sea (*Posidonia* meadow being the first one), and more than 1600 species constitute and live in these habitats (Ballesteros, 2006) [1]. They are also considered as high-value ecological zone since the Barcelona Convention. In 2008, this convention proposed a management plan for the coralligenous habitats, but there is currently no regulatory instrument for their protection. However, the EU Marine Strategy Framework Directive (DCSMM) requires that each state develop a strategy and an action plan in order to reach and maintain a "Good Environmental Status" [GES] for its marine habitats. This GES, as defined by the DCSMM, is assessed by 11 descriptors about the state and the pressures measured for each milieu. The first symposium on coralligenous habitats took place in 2009 [17], further to the Action Plan for the Conservation of Marine Vegetation started (adopted in 1999 by the stakeholders of the Barcelona Convention).

Since the publication from Marion in 1883, many studies on coralligenous habitats have been published and are now references including Laubier (1966) [9], Hong (1980) [7], Ballesteros (2006) [1]. Coralligenous habitats are assemblages of complex habitats with very low dynamic of construction that is not very documented. These habitats are not only "hotspots" of biodiversity, but furthermore they represent socio-economic stakes. Activities such as small-scale fishing and scuba diving highly depend on them. Fishermen look for species of high commercial value such as red coral, crustaceans, rock fishes, and other seafood. Divers look for the landscapes beauty, offered by coloured and erect species such as gorgonians, corals and bryozoans. Beyond these interests, other services provided by coralligenous habitats are suspected such as CO_2 sequestration [12] [13] or marine bottoms stabilization [14]. They are threatened by global change, and anthropogenic pressures. Only the GES can guarantee the maintaining of all these services provided by coralligenous habitats.

There are pretty few programs and networks for the monitoring of coralligenous habitats. CIGESMED (Coralligenous Indicators based to Evaluate and Monitor the "Good

Environmental Status" of the Mediterranean coastal waters) is one of them. It implies three countries (France, Greece and Turkey) from 2013 to 2016. CIGESMED objectives are (1) to fulfil the key gaps in the current scientific, knowledge of the coralligenous habitats that make it difficult to make recommendations for protecting them by developing barcoding to enhance reliable identification for conservation and protection purposes (invasive and cryptic species), and by studying genetic structuring and effective dispersal potential of keystone/habitat species, (2) to enhance the knowledge on coralligenous populations by deciding on reference states and setting up a network of Mediterranean experts (long term series), (3) to monitor networks, locally managed and coordinate them on a regional scale, standardizing protocols that could be applied to the entire Mediterranean and testing indices and indicators, specific to coralligenous, (4) to test population genetic criteria as tools to monitor the GES of the coastal Mediterranean Sea, (5) to implement a "citizen science" network and (6) to use trees of knowledge as tools to sort, organize and illustrate the large heterogeneous sets of produced data and as a tool of dissemination towards scientists, decision makers, environmental managers and general public.

This includes habitats cartography, population genetics studies to understand species relations and dispersal potential, and the setting up of a monitoring protocol.

A phase of inter-calibration of methods/material/operators is firstly implemented. This is done in order to evaluate the variability related to these experimental parameters. It enables to know in which situations results obtained by different underwater protocols are comparable. Moreover, this phase of test helps to select the best protocol to apply (the easiest, and most reliable) depending on the habitats types. The next phase will be the study of natural variability inter-site or intra-sites.

Materials and Methods

Observations and cartography of coralligenous habitats

Intercalibration methods

The studied sites are located in Marseilles Bay. They are transects of 10 meters long at 28 meters depth. The chosen coralligenous habitats are walls dominated by the red gorgonian *Paramuricea clavata*, with different grades of roughness, from anfractuosities to hollows or caves, and great species richness.

To date, three variables have been studied: the sampling method, the quality of the camera, and the level of knowledge of operators in charge to identify species. The protocol was implemented as follow. Divers made photo-quadrats using a frame of 50 cm by 50 cm. The pictures were analysed by operators using the software Photoquad® [16]. Hundred points were distributed by stratified randomization. Then the operator assigned each point to one category and one sub-category among these three: (i) higher taxa (such as phyla, orders), (ii) abiotic, (iii) indeterminate. In the first category (i) the sub-categories are lower taxa (such as genius or species). The second category (ii) is subdivided into four sub-categories: sediment, bare rock, organic detritus or debris. In the third one (iii), there are three sub-categories: fuzzy image, shadow/hole, and unidentified taxon.

The first variable studied was the sampling method. Two sampling methods were compared: (i) permanent linear transect and (ii) random patches transect. The implementation of method (i) consisted for the diver to start from a permanent point, and

make 20 photo-quadrats, in a continuous way, on the 10 meters long transect, following a virtual horizontal line at constant depth. The implementation of method (ii) consisted to make patches of 9 photo-quadrats randomly, at constant depth. To make a patch, the diver places a frame marking the centre of the patch. Then he makes the photo-quadrats around the frame starting by the left bottom corner, and finishing by the right upper corner. It should draw a patch of 3 by 3 photo-quadrats, as illustrated figure 1.

The second variable studied was the quality/performance of the camera. Two cameras have been compared: (i) a camera of medium quality and (ii) a camera of high quality. The models used are (i) GoPro®, and (ii) Nikon® D300s. The third variable studied was the level of knowledge of the operators in charge to identify the taxa on photo-quadrats. Two levels of operators were compared: (i) novice and (ii) experienced. The set of operators participating were: one novice and two experienced operators. Each of them analysed separately the first 5 photo-quadrats of the transect (series 1). Then they met to exchange their results (species identification) about this first set and re-do the identification work together to produce "validated data". Again they studied separately the 2nd series of 5 photo-quadrats. Then they met again to exchange their knowledge and produce "validated" data on this 2nd series. They proceeded like that 4 times to analyse the 20 photo-quadrats of the permanent transect.

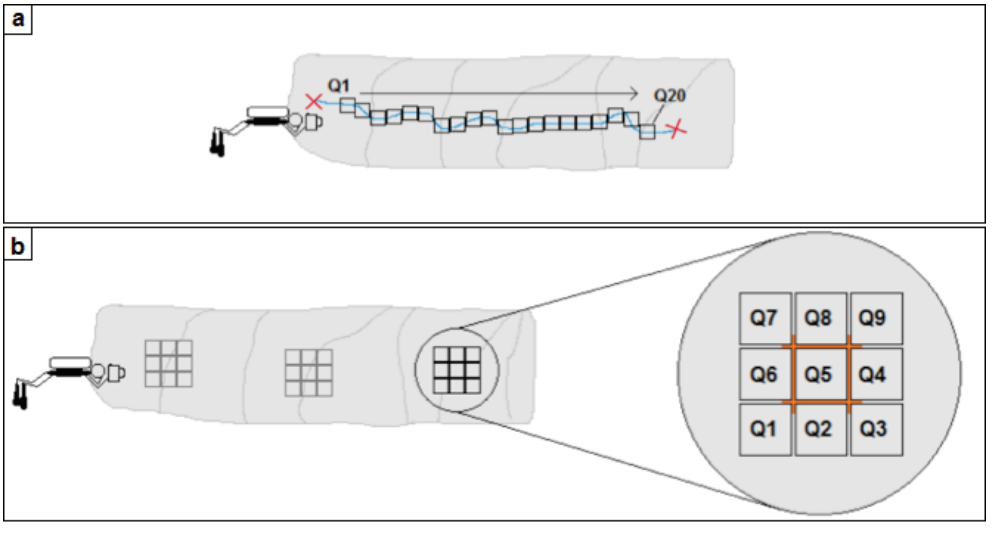


Figure 1 – The two types of transects tested. Linear transect (a): 20 photo-quadrats are taken at a given isobath, located by permanent marks. Random patch transect (b): 3 groups of 9 photo-quadrats are taken, following the indicated numbers scheme from 1 to 9, several times at a same depth.

Profile characterization and associate symbology

Due to their topography and to their complexity, there are few accurate maps of coralligenous habitats. CIGESMED program is experiencing a way to symbolise them by means of easily recognizable signs. Based on the estimation of the diversity and abundance of species observed during field surveys, coralligenous habitats cartography should also

take into account the profile parameters (orientation, slope, roughness, and major covers) that favour one or the other taxon. This study of the variability of the coralligenous habitats structure is made on small islands of Marseilles' bay at a constant depth of 28 (± 1) meters. Observations were done on different types of sites. Either around small islands and shoals, or along coastline with all orientations represented. For each site, two depths were sampled around 28 m deep (± 1 m), and around 45 m deep (± 1 m). Samples were collected along transects cut into segments of 5 m long and 1 m wide.

To define the profile of each segment, the following typology has been applied:

- *Orientation of the wall*: North, South, East, West and the four intermediates (Northeast, Northwest, Southeast, Southwest).
- *Slope of the wall*: the four categories are V, I, F, C = Vertical, Inclined, Flat, Ceiling (2a).
- *Roughness of the wall*: the size of holes observed on the entire segment was described as: « T » (Tiny) segments with holes are less than 10 cm. « S » (Small) segments with holes that measure between 10 and 30 cm. « M » (Medium): segments with holes between 30 cm and 1 m. « L » (Large): segments whose holes must be at least 1 m large. This typology can be apply easily by the under-water diver with anatomic references as shown in Figure 2b (finger(s), fist, head, and shoulders).
- *Main coralligenous species covers*: the 3 or 5 majoritarian taxa were recorded with indication of relative abundance of encrusting or erect species according to the code (Figure 2c).

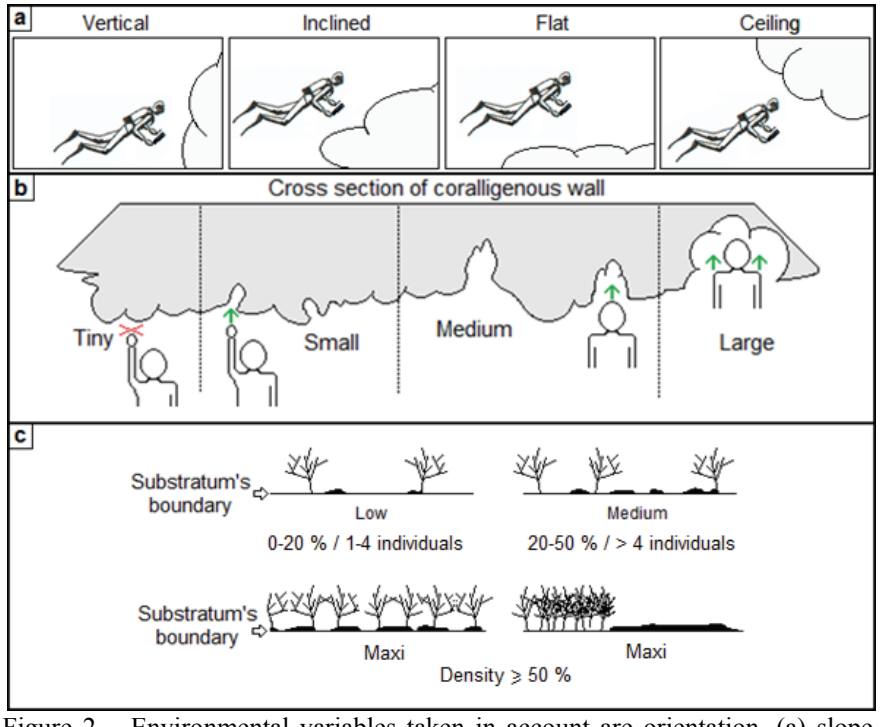


Figure 2 – Environmental variables taken in account are orientation, (a) slope, (b) roughness and (c) main coralligenous populations.

For this study, the cover "Low" and the lowest occurrences of species have not been taken into account in the data analysis. For data processing, the frequency of species observation for each segment was calculated for each profile setting.

Table 1 – List of the metrics and of their acronyms.

Metric	Possible values	Signification
Time	Digital positive integer	Time in minutes on the timer
Orientation	N, NE, E, SE, S, SW, W, NW	North, Northeast, East, Southeast, South, Southwest, West, Northwest
Inclination	V,I,F,C	Vertical, Inclined, Flat, Ceiling
Roughness	0, +, ++, +++	No roughness, Low roughness (fist), average roughness (head), large roughness (shoulders)
Upper stratum	CR, Esp, EC, ES, PC	*Corallium rubrum*, Erect sponge *Eunicella cavolinii, Eunicella singularis, Paramuricea clavata*
Basal stratum	EnRA, EnGA, ErRA, ErGA, Bryo, Cod, Sp, Pey, Turf	Encrusting red algae, Encrusting green algae, Erect red algae, Erect green algae, Bryozoan, *Codium* spp, Sponge, *Peysonnelia* spp, Turf
	HT, LP, LC, MA, PA	*Halimeda tuna, Leptosamia pruvotii, Lithophyllum cabiochae, Mesophyllum alternans, Parazooxanthus axinellae*
Remarkable species	According to diver's knowledge	The diver must specify his knowledge fields on the form
Remarkable stands	According to diver's knowledge	The diver must specify his knowledge fields on the form
Solid waste	Objects and sizes	The diver must precise the object's type and its size (50 cm, 1 m, several metres…)

Population genetics studies

This aspect is just starting. The aim is to understand the population structure of the coralligenous species by studying the intraspecific diversity of demes and their connectivity. The studied species are living throughout the Mediterranean Sea, giving opportunity of detecting cryptic species. For this reason, we will use the barcoding method consisting in sequencing a part of the mitochondrial gene Cytochrome c oxidase subunit I or COI. We will eventually complete with other alternative or complementary markers.

The chosen species are the erect and tree-like bryozoan *Myriapora truncata*, and one complex of bioconstructing corallinale algae *Lithophyllum stictaeforme/cabiochiae*, both of which are identifiable *in situ*. They were selected as they have a widespread occurrence in the coralligenous, on all the facies and at all depths (even at very low irradiance).

Myriapora truncata (Pallas, 1766) is a common bryozoan of the coralligenous habitats throughout their distribution area. Despite its reproduction pattern (low dispersal lecitotrophic and brooded larvae), it is widespread in the Mediterranean coasts. We wonder whether there is a single species in different Mediterranean basins, or whether there are cryptic species, even possibly in sympatry, as for other bryozoans. Red calcareous algae of the order

Corallinales are the main coralligenous builders [1] [8] [9] [18]. We chose to study the genetic diversity within the species complex of *Lithophyllum stictaeforme/cabiochae* (Areschoug) Hauck, 1877 / (Boudouresque & Verlaque, 1978) Athanasiadis). We compare the frequencies of different genetics variants, so it is important to have enough specimens in order to be able to conclude on the possible presence of significant differences of the genetic frequencies. About thirty individuals are required for each locality's sample [15] [19].

The sampling should be made on two sides of the study site, if possible opposite ones, and at depth of 28 ± 1 m, making sure that the collected samples would come from all possible orientations. In each side, we chose two different profiles (in terms of the inclination and the roughness of the substrate). These profiles would have been firstly determined by the results of the cartography.

Standard PCR protocols is used to amplify COI fragments for both the bryozoan and the red alga, as well as another marker, not from the mitochondrial genome, for each species (detailed methods to be published elsewhere): an intron for *Myriapora truncata* [4] [6], and a chloroplastic marker for *Lithophyllum* sp. [3]. PCR products are sent to the industry for DNA sequencing, then after alignement, haplotype network reconstruction is made using the Median Joining Network software [2].

Results

Photo-quadrats inter calibration

To compare the sampling methods 20 photo-quadrats done by the permanent transect method were compared to 18 photo-quadrats (2 patches) done by the random patches method. The preliminary results are presented on figure 3. It shows that at the phylum level, the two methods give equivalent results. Differences of headcounts are only significant in the phylum Porifera. Thus results are comparable.

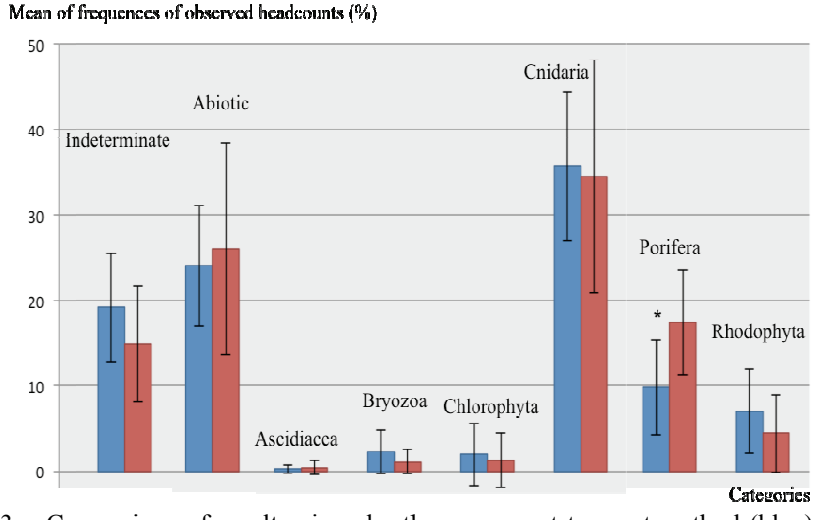


Figure 3 – Comparison of results given by the permanent transect method (blue) and the random patches method (red). The star marks a significant difference according to the Mann-Whitney-Wilcoxon test with a 5 % risk.

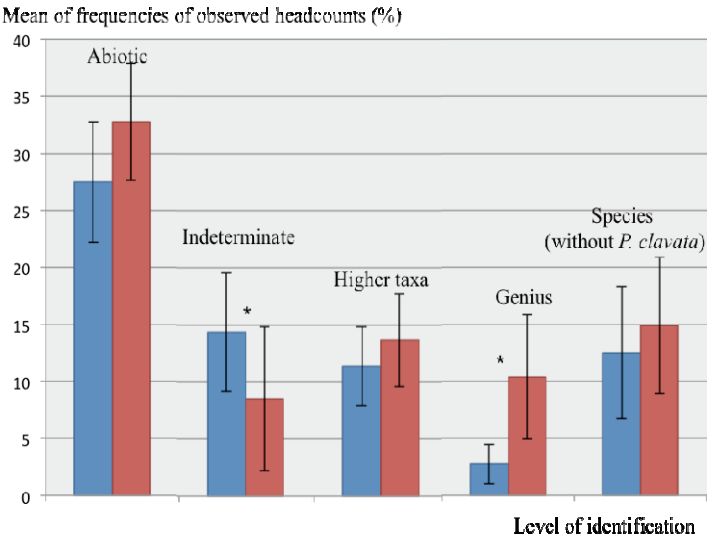


Figure 4 – Comparison of results obtained with a medium quality camera (blue) and a high quality camera (red). The star marks a significant difference according to the Mann-Whitney-Wilcoxon test with a risk of 5 %.

To compare both cameras, two sets of 8 photo-quadrats done on a permanent transect at the exact same place, were used. One was done with the medium quality camera, and the other one done with the high quality camera. But as the pictures were not taken at the exact same time, the species *Paramuricea clavata* disturbed the observations as it had its polyps spread out or not, depending on the set. To free ourselves from this disturbance, all observations of *Paramuricea clavata* were removed from both sets. The figure 4 shows that at the species level both cameras give equivalent results. But the camera of high quality enabled to reduce the number of "indeterminate" and these observations were assigned to other categories, in the majority at the genus taxa level.

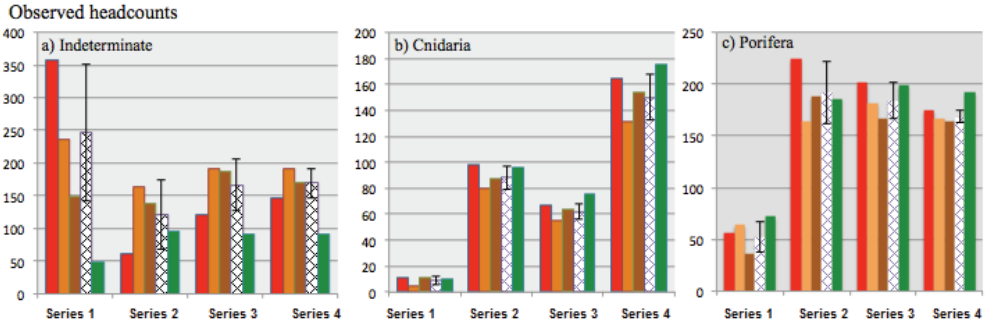


Figure 5 – Comparison of results observed by different operators: one novice (in red) and two experienced (orange and brown). Mean and standard deviation between the three operators are in white cross-brace. The identifications validated by the three operators are in green.

The preliminary results on certain categories are shown on the figure 5. It shows that after only one exchange between the three operators, the novice improve a lot his capacity of identification. For some categories, the level of knowledge of the three operators gets quickly homogenized: for instance for the categories Cnidaria and Porifera. They are in most case very hard to identify on photography. From this work, it appears that cnidarian species are easiest to identify on photography by beginners while poriferan species are mostly very hard to identify without specific training.

Photo-quadrats and genetic sampling contextualization

The figure 6 shows a specific symbology which has been developed in order to represent as clearly as possible on one map the orientation, slope (in black) and roughness (in red) of coralligenous habitats.

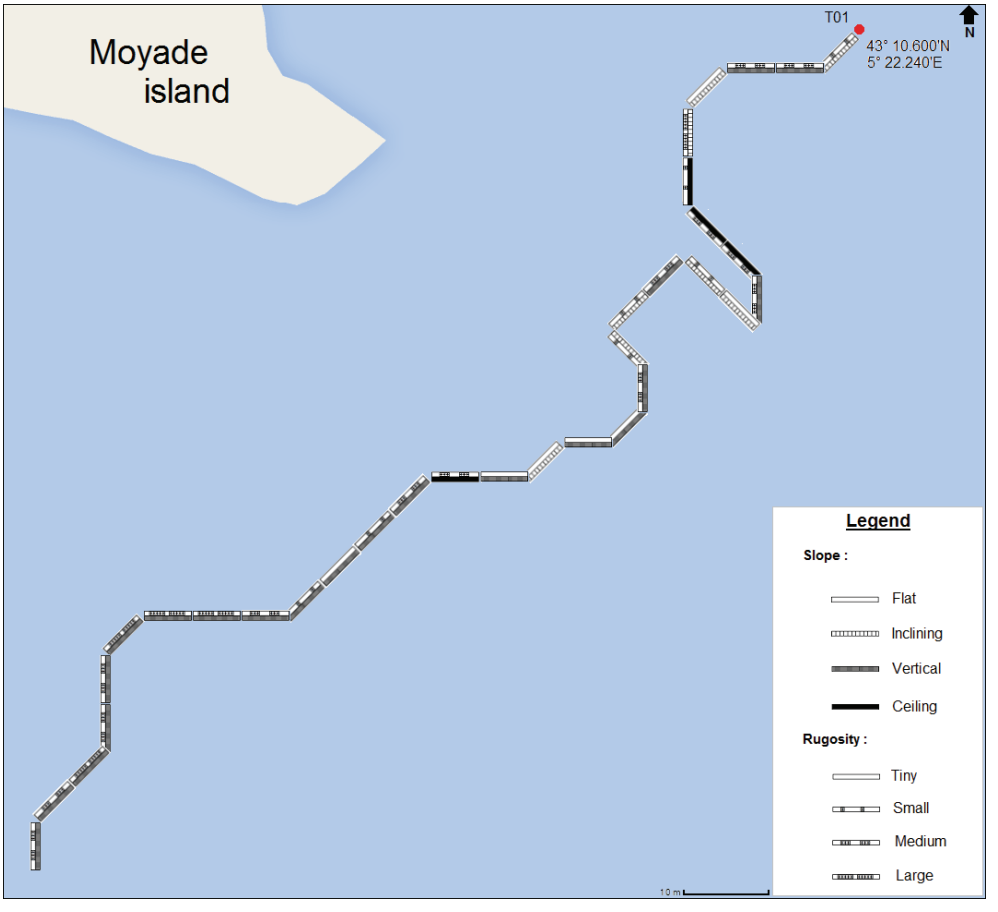


Figure 6 – Cartography of coralligenous habitats around Moyade Island at the isobath 28 ± 1 meters deep with first symbology (to be evaluated).

Analysis of the preliminary cartography results

Using Hierarchical Ascendant Classification (HAC) and Correspondence Factorial Analysis (CFA) on all processed data allowed (i) to group species according to values of orientation, slope and roughness, and (ii) to pool profile parameters according to species observed per segment. They HAC shows four groups of species (Figure 7a) and five groups of profil parameters (Figure 7b). The CFA was performed on the sum of the frequencies of each variable (species and profile parameters) show in Table 2.

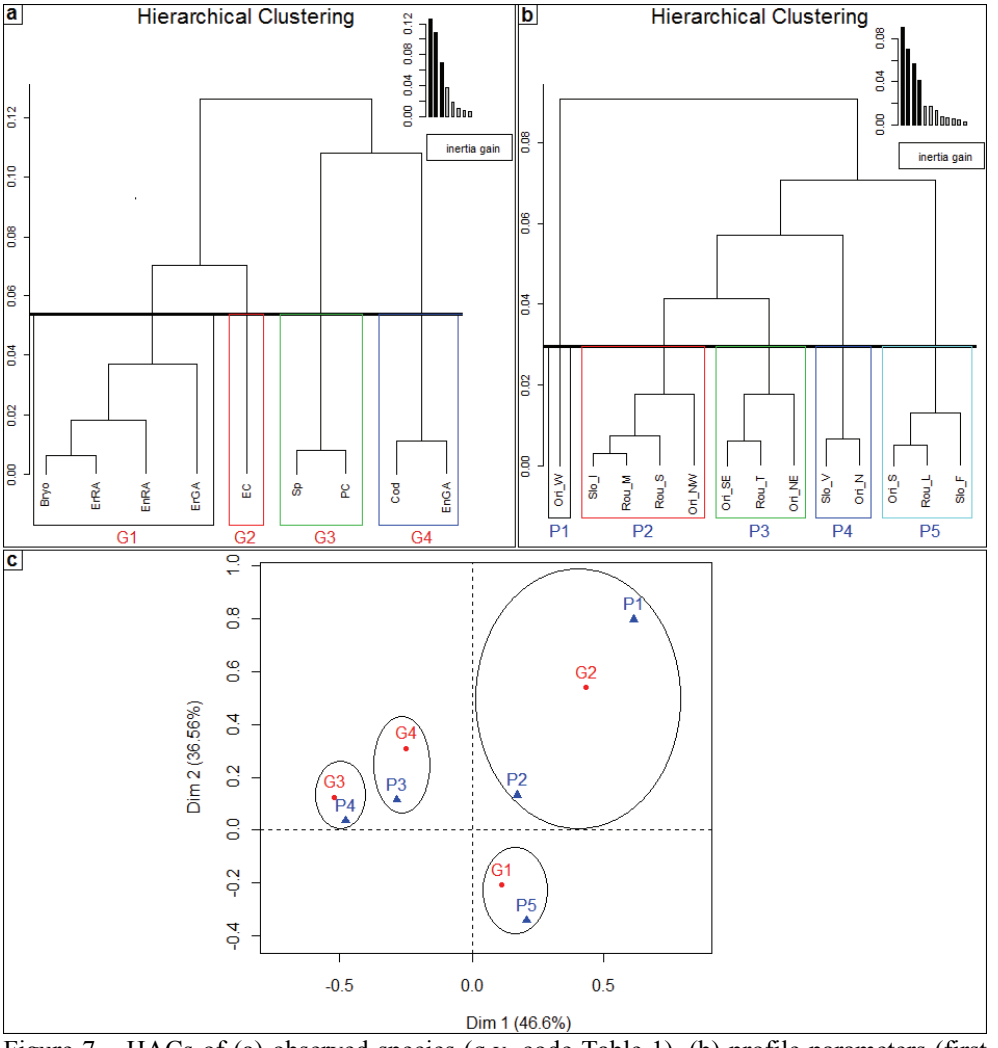


Figure 7 – HACs of (a) observed species (q.v. code Table 1), (b) profile parameters (first letter of the variable with : "Ori" for orientation, "Rou" for Roughness and "Slo" for Slope) and (c) FCA on frequencies of different groups of species (G1 to G4) and groups of profile parameters (P1 to P5) obtained from the HAC. n = 117.

Table 2 – Table of frequencies for each species group and profile parameter (%).

Catégories	P1	P2	P3	P4	P5
G1	0,59	4,77	3,20	2,53	7,65
G2	0,84	1,75	0,63	0,45	0,89
G3	0,19	1,33	1,89	2,46	1,11
G4	0,05	0,64	0,99	0,13	0,20

The aim of the AFC is to associate, at best, a single group of species with a single group of profile parameters. Here, the AFC is a representation of the grouping based on proximity most common observations. Groups can be combined as follow: the group G1 is associated with P5. As P1 contains only one profile parameter, the group G2 is associated with P1 and P2. G3 and G4 groups are respectively associated with P4 and P3 groups.

Using Factorial Correspondence Analysis and Ascending Hierarchical Classifications, preferential profiles of coralligenous species can be determined: bryozoans, encrusting and foliose red algae and foliose green algae occur preferentially on horizontal walls South-oriented with large roughness. *Eunicella cavolinii* is mainly present on inclined walls facing West/North-West with low or medium roughness. Porifera and *Paramuricea clavata* are preferentially present on vertical walls facing north. *Codium* genus and encrusting green algae are more present on walls facing South-East and North-East with tiny roughness. A "profile" is thus a combination of orientation parameters combined with inclination and roughness. Further detailed study will be conducted on associations of profile features to determine the preferential profiles of different coralligenous communities. Moreover, a pre-established species list will be detailed in order to obtain more accurate results. The identification of various difficulties encountered during the contextualization work will allow other participants of the CIGESMED program to use this method in Greek and Turkish coralligenous habitats.

Data and network organisation

The aim of the data systems in CIGESMED is to make data reusable and scalable with other observatory networks. The primary principle of data organisation in CIGESMED is to not centralize data, but to adopt all accessible formats, and propose (i) a typology to contextualize taxonomic observations and ecological structure studies and (ii) tools to share these data at a large scale (Mediterraean Sea) and over the long term (at least 10 years). Achievement of these objectives requires to use (i) open access, open data, and open source software, (ii) more exchanges between country scientific communities about coralligenous studies, and (iii) make sure that surveys and protocols are reusable (cost effective, security in dive and analytics methods, knowledge of managers…). The first tool that must permit it is a plugin able of self-automatic installation on the web site of the partners, in order to install the database with formatted fields and some options to share data. The sharing flux is provided on different formats (XML, RDF), permitting to build graphs out of the information system and to request a selected part of the dataset. These data of one partner can be aggregated with other partner datasets, respecting the same query. Indexations servers and programs are responsible to build these graphs. Moreover specific ontologies will permit to densify the links between the different objects presented above (assigned points, photos, segments...). At this scale, the query is using metrics of

contextualization to improve the sensibility of each analysis and to understand what is comparable between Mediterranean regions and what is not. The objective is to build new representations of data, used to find correlation between discrete and non-ordinate values, and not only systems of metrics.

Discussion

Assessment of the environmental status

The assessment consists of the analysis of the coralligenous megabenthic assemblages by means of direct observation, photographic/video surveys (which are directly influenced by competencies and experience of the operators). This is done in order to investigate basic demographic characteristics of key associated species populations and disturbances and threats and to correlate biotic (*i.e.* alien species) and abiotic factors affecting the coralligenous habitats. This requires understanding the effect of each factor on data variability.

The quality of the photo-quadrats dataset suffers from the difficulty to properly isolate each variable studied. This is mainly due to the complexity of undersea operations, but it can be improved by experience gain.

The study of both sampling methods (linear or random transect) shows that observations made at a phylum level of identification are comparable if they are made according to the one or the other method. But to compare *Porifera* observations from one site to another, it would be recommended to implement the same method in both sites, as it's shown that there might be significant difference of headcounts according to the applied method. As the random patch method is easier to perform, it should be recommended.

Concerning the effect of the quality of the camera on the observations made, it has been proven that the medium quality camera is sufficient to identify as much species as the high quality camera. Difference between the two is observed for species difficult to differentiate at the genus level. Indeed the high quality camera enables to give a taxonomic level at some individuals that were indeterminate with the medium quality camera. As the medium quality camera is more affordable for managers of marine protected area (5 time less expensive), it should be preferred. Finally the study of the operator's knowledge shows that the discussion between operators enables novices to quickly improve their capacity of identification. Discussion is very useful at the beginning, and then operators reach a step, and would need a proper training to progress, if more precise identification is needed.

It should be noted that the results presented in this document are preliminary, and that work is still running. The non-independence of the variables studied implies that the hypothesis should be stipulated very carefully.

Population studies

The objective of the cartography goes beyond than mapping habitats; it provides information about environmental profiles (slope, orientation, roughness and main stands) that will be used to understand species preferences. The population genetics approach which will complete the analyses is essential to investigate species diversity, population structure and connectivity.

Myriapora truncata is an erect and robust bryozoan that is a coralligenous bio-constructor found in most of the coralligenous stands. Mitochondrial and nuclear genetic markers are now tested on samples collected throughout littoral of the three Mediterranean countries implied in CIGESMED. Sequencing data of different genes are used to study (i) the phylogeny of *Myriapora truncata* and look for cryptic species, (ii) the phylogeography in order to understand its distribution area, (iii) the connectivity between populations. *Lithophyllum cabiochiae* is a red calcareous alga taking a major part in the coralligenous structure. Due to its photosynthetic and calcifying ability, it has an important ecological role in the carbon flux, benthic productivity, and the habitat complexity of this typically sciaphilic Mediterranean ecosystem.

Lithophyllum cabiochiae is considered to be a Mediterranean species and *L. stictaeforme* its Atlantic sister species. Despite their ecological importance, the identification in these species is quite hard and their taxonomic status, presently based on morphology, is constantly changing. Our work aims to answer the taxonomic questions about these species complexes by means of molecular markers and to provide information on their population structure and phylo-geography within the Mediterranean basin. The connectivity amongst the study sites could be evidenced, helping the decisions taken about management and conservation issues.

The first results about genetic differentiation of the populations of distinct localities for each taxon illustrate the fact that gene flow (migration) is limited even at the small scale of the Marseilles region for those important coralligenous builders. Genetic barrier were previously evidenced for other species analysed in population genetics as different as the mysid *Hemimysis margalefi* [10] (Lejeusne and Chevaldonné 2006) or the irregular sea urchin *Echinocardium cordatum* [5] (Egea 2011). It may also be linked to ecological conditions; the study of the distribution of divergent groups of each species depending on currents and ecologic profiles has to be carried out.

For this study, the mapping of coralligenous habitats can be improved. Species presents in some groups are very different. So, for futures mapping, these species will be separated. For example, the group Erect Green Algae will become *Halimeda tuna* and *Flabellia petiolata*; Red Algae will become *Mesophylum*, *Lithophyllum* and Peyssonneliaceae; Sponges will become erect sponges and encrusting sponges. Moreover, the aim of the mapping is to associate the coralligenous communities with their preferential profile (orientation, slope, roughness) and not just with parameters. So, when the dataset will be enough large, these analyses will be conducted on all of profiles possibilities.

Data and network organisation

New representations of data, built to find correlation between discrete and non ordonate values, and not only systems of metrics depends on the quality of ontologies. The network and partners are testing now the robustness of each typology for characterization, in order to be sure that all measurements have the same meaning.

Metrics that happen to be discriminant on these representations will be tested. The aim is to apply a new index of conservation state along French coasts, to test this index on coralligenous bottoms in the Eastern Mediterranean basin and compare it with other methods used to evaluate the conservation status of the benthic communities of coralligenous bottoms.

Conclusion

A difficulty generally not taken in account is the human factor. The aim of this pilot study of operational ecology is to improve knowledge on the dynamic of coralligenous habitats and also to improve their management by creating protocols and tools. This preliminary work yet enables to better understand the importance of the level of the skill and training of operators, the type of sampling methodology implemented. CIGESMED integrated approach of complexity of coralligenous habitats must permit to mutualize and visualize large data collections, and manage knowledge to study ecosystems. Indicators, from communities to infra-specific level, will be co-constructed and their variability will be tested by scientists, marine natural parks and reserves managers, and through the implementation of a "citizen science" network. The use of new representations of data and controlled qualifications as links in a graph as tools to sort, organize and illustrate very large heterogeneous data sets is an original approach. The outcome will be an integrative assessment of the GES within the MSFD.

To continue to build the network, the community is developing a metadata catalogue and some shared typologies; we are working now on i) harmonization of data collection methods and normalization of data access (European norms) ii) initiation/animation of thematic network about coralligenous habitats in Mediterranean Sea gathering all competent actors. This network is meant to be perennial, open, and fully decentralized (to allow for continuous update) at local, regional, national and international scales

This organisation will permit data diffusion and upper accessibility through local, regional, national and international reports and Quality Management System at each geographic level (local, regional, national levels) to ensure continuous improvement.

Acknowledgements

We are gratefull to the organizing committee who invite one of us (RD). We thank A. Haguenauer, S. Chenesseau, F. Zuberer and all the underwater diving team for their participation to the sampling in Marseilles and T. Dailianis for samples from Crete. We thank M. Verlaque who helps us for the identification of Corallinales and A. Le Gall for helpful advices. Thanks are also due to S. Chenesseau and A. Ereskovsky for the SEM. A special thanks to the SIP (Computing Service of the OSU Pytheas) who provides all the hardware part and assistance for developments.

This project was funded by the SeasEra program CIGESMED (CNRS- ANR convention n° 12-SEAS-0001-01).

References

[1] Ballesteros E. (2006) - *Mediterranean coralligenous assemblages: a syntheses of present knowledge*, Oceanogr. Mar. Biol., Annu. Rev. 44, 123 – 195.

[2] Bandelt H. J., Forster P., Röhl A. (1999) - *Median-joining networks for inferring intraspecific phylogenies*, Mol. Biol. Evol. 16, 37 – 48.

[3] Broom J. E. S., Hart D. R., Farr T. J., Nelson W. A., Neill K. F., Harvey A. S., Woelkerling W. J. (2008) - *Utility of psbA and nSSU for phylogenetic reconstruction in the Corallinales based on New Zealand taxa*, Mol. Phylogenet . Evol. 46, 958 – 973.

[4] Chenuil A., Hoareau T. B., Egea E., Penant G., Rocher C., Aurelle D., Mokhtar-Jamai K., Bishop J. D. D., Boissin E., Diaz A., Krakau M., Luttikhuizen P. C., Patti F. P., Blavet N., Mousset S. (2010) - *An efficient method to find potentially universal population genetic markers, applied to metazoans*, BMC Evol. Biol. 10, 276.

[5] Egea E (2011) - *Histoire évolutive, structures génétique, morphologique et écologique comparées dans un complexe d'espèces jumelles: Echinocardium cordatum (Echinoidea, Irregularia)*. Aix-Marseille Université, Marseille.

[6] Gérard K., Guilloton E., Arnaud-Haond S., Aurelle D., Bastrop R., Chevaldonné P., Derycke S., Hanel R., Lapègue S., Lejeusne C., Mousset S., Ramšak A., Remerie T., Viard F., Féral J. -P., Chenuil A. (2013) - *PCR survey of 50 introns in animals: Cross-amplification of homologous EPIC loci in eight non-bilaterian, protostome and deuterostome phyla*, Mar. Genomics. 12, 1 – 8.

[7] Hong, J. -S. (1980) - *Etude faunistique d'un fond de concrétionnement de type coralligène soumis à un gradient de pollution en Méditerranée nord-occidentale (Golfe de Fos)*, Thèse de doctorat. Aix-Marseille II.

[8] Laborel J. (1961) - *Le concretionnement algal "coralligène" et son importance géomorphologique en Méditerranée*. Recueil Travaux Station Marine d'Endoume. 23, 37-60.

[9] Laubier L. (1966) - *Le coralligène des Albères. Monographie Biocoenotique.* Ann. Inst. Océan., Paris.

[10] Lejeusne C, Chevaldonné P (2006) - *Brooding crustaceans in a highly fragmented habitat: the genetic structure of Mediterranean marine cave-dwelling mysid populations*. Mol Ecol 15, 4123–40

[11] Marion A. F. (1983) - *Esquisse d'une topographie zoologique du Golfe de Marseille*. Ann. Mus. Hist. Nat., Marseille, Marseille.

[12] Martin S., Charnoz A., Gattuso J. -P. (2013) - *Photosynthesis, respiration and calcification of the Mediterranean crustose coralline alga Lithophyllum cabiochae (Corallinales, Rhodophyta)*. Eur. J. Phycol. 48, 163 – 172.

[13] Noisette F. (2013) - *Impacts de l'acidification des océans sur les organismes benthiques calcifiants des milieux côtiers temprérés*. Thèse de doctorat. Université Pierre et Marie Curie.

[14] Pedel L., Fabri M. C., Menot L., Van Den Beld I (2013) - *Mesure de l'état écologique des habitats benthiques du domaine bathyal à partir de l'imagerie optique. (Sélection de métriques et proposition d'une stratégie de surveillance).* Convention 13/1210491/NYF Convention MEDDE-Ifremer pour le Bon Etat Ecologique des habitats benthiques profonds.

[15] Porter J. S., Ryland J. S., Carvalho G. R. (2002) - *Micro- and macrogeographic genetic structure in bryozoans with different larval strategies.* J. Exp. Mar. Biol. Ecol. 272, 119-130.

[16] RAC/SPA UNEP – MAP (2006) - *Classification des biocénoses benthiques marines de la région Méditerranéenne*. CAR/ASP (Tunis), 13 p.

[17] RAC/SPA UNEP – MAP (2009) - *Proceedings of the 1st Mediterranean symposium on the conservation of the coralligenous and other calcareous bio-concretions,* RAC/SPA (Tabarka), January. 1–278.

[18] Sartoretto S. (1996) – *Vitesse de croissance et bioérosion des concrétionnements 'coralligènes' de Méditerranée nord-occidentale. Rapport avec les variations Holocènes du niveau marin.* Thèse doctorat d'écologie. Université d'Aix-Marseille II 194 p.

[19] Schwaninger H. R. (1999) - *Population structure of the widely dispersing marine bryozoan* Membranipora membranacea *(Cheilostomata): implications for population history, biogeography, and taxonomy*. Mar. Biol. 135, 411 – 423.
[20] Trygonis V., Sini M. (2012) - *PhotoQuad: A dedicated seabed image processing software, and a comparative error analysis of four photoquadrat methods*. J. Exp. Mar. Biol. Ecol. 424, 99-108.
[21] Walsh P. S., Metzger D. A., Higuchi R. (1991) - *Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material*, Biotechniques 10, 506-513.

SITIS 2015, 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 23-27 nov. 2015, Bangkok, Thailand, pp. 232-239, DOI:10.1109/SITIS.2015.119

A first prototype for indexing, visualizing and mining heterogeneous data in Mediterranean ecology

within the IndexMed consortium interdisciplinary framework

R. DAVID, J.-P. FERAL, S. GACHET, A. DIAS
Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE)
CNRS, Aix Marseille Université
IRD, Université d'Avignon
Marseille, FRANCE
romain.david@imbe.fr, jean-pierre.feral@imbe.fr, sophie.gachet@imbe.fr, alrick.dias@imbe.fr

C. BLANPAIN, J. LECUBIN
Service informatique (SIP)
OSU Pythéas, CNRS
Aix Marseille Université
Marseille, FRANCE
cyrille.blanpain@osupytheas.fr, julien.lecubin@osupytheas.fr

C. DIACONU
Centre de Physique des Particules, CNRS
Aix Marseille Université
Marseille, FRANCE
diaconu@cppm.in2p3.fr

C. SURACE
Laboratoire d'Astrophysique de Marseille (LAM)
CNRS, Aix Marseille Université
Marseille, FRANCE
christian.surace@lam.fr

K. GIBERT
Department of Statistics and Operations Research
Universitat Politecnica de Catalunya
Barcelona, SPAIN
karina.gibert@upc.edu

***Abstract*—: Although biodiversity has been extensively studied over the last centuries, recent evidences suggest that links between collected data are still be missing. In order to fill this knowledge gap and at the initiative of the CNRS Institute of Ecology and Environment (INEE), IndexMed <www.indexmed.eu>, a unique and multidisciplinary consortium consisting of ecologists, sociologists, economists, mathematicians, IT specialists and astronomers was created. Through exploratory projects, IndexMed develops new methods for analyzing data on Mediterranean biodiversity and implements solutions based on interoperability technologies already deployed in other disciplines. In particular, IndexMed aims to build a prototype of such data graphs and "data cubes". This paper will first explore the ability of tools and methods by means of graphs to connect biodiversity objects with non-centralized data. It will then introduce the use of algorithms and graphs to analyze environmental and societal responses and presents a prototype under development.**

Indexing data, data qualification, data traceability, decentralized information systems, data-mining, ecology

I. INTRODUCTION

Although considered by most scientific disciplines and industries producing and using information as the most promising opportunity for progress and discoveries, the use of Big Data in Ecology is still lagging behind [34]. Information systems linking objects with qualifications (links) are omnipresent, the first object being the consumer. All these links can be used for data mining process. Data mining is the computational process allowing discovering patterns in large data sets («big data») and involves methods at the crossroads of artificial intelligence, machine learning, statistics, and databasing systems. (<www.kdd.org/>). Nowadays, many businesses have understood its huge analytical potential (insurance to manage risk, games companies to find cheaters, banks for investments, traders to increase their margins, advertisers and networks to increase their social/commercial impacts, etc.) Today, the environmental emergency requires a response through a shared system connected to local and global issues and beyond scientific questions such as: "Is this degradation related to this particular pressure?" instead addressing the following question: "How can we improve/preserve the ecological condition of an environment in the most efficient way?". Such queries help identify the limits not to be exceeded, for a set of conditions which may have opposite or complementary effects.

Data in ecology are extremely heterogeneous (Fig. 1). To get access to the different type of data and data formats, the system must be distributed, *i.e.* data remain where they were produced, although normalized flows are deployed and can be accessed by all members of the network, both at local and international levels and with an index to list and describe each record based on shared typologies. There are already some initiatives using Semantic Web technologies for retrieving Biodiversity data [1], Improve knowledge and developing methods for linking of biodiversity and environmental data, but they often concern only an "inventory" aspect of biodiversity (collection, observations, repositories and distribution) and far less functional aspects (Catalog of life, Data-ONE [Data Observation Network for Earth], EMODnet [European Marine Observation and Data Network], GEO and EU-BON [Group On Earth Observations Biodiversity Observation Network] and [European Biodiversity

Observation Network], GBIF [Global Biodiversity Information Facilities], LifeWatch, OBIS [Ocean Biogeographic Information System], TDWG [Biodiversity Information Standards] - with Darwin Core and ABCD - are well-known examples of efforts for interoperability and standardizing data collection regarding biodiversity).

The newly created IndexMed consortium has set a goal to build such a distributed system, with the help of institutes from different disciplines, in order to overcome skills lacking among ecologists.

IndexMed was created by the axis "Management of biodiversity and natural spaces" of the IMBE (Mediterranean Institute of Biodiversity and marine and terrestrial Ecology) with the aim to develop the knowledge of databases and their effective use in the ecological research community.

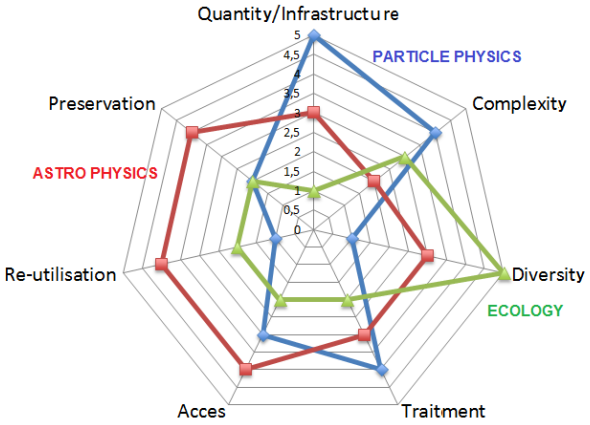


Figure 1. First summary of drawn conclusions in the MASTODONS (very large scientific datasets) meeting. Big Data in ecology are compared to astrophysics and particle physics data and demonstrate that diversity is the most important aspect to consider when dealing with such data.

In particular, this consortium responds to project calls and uses databases and address ecological issues in the Mediterranean Basin, therefore promoting multidisciplinary and collaboration across CNRS [National Centre for Scientific Research] institutes, other research entities and universities. The projects developed by members of the IndexMed consortium must be based on various national and international initiatives and promote international collaborations, therefore connecting existing networks to initiatives at national and international levels.

IndexMed's short-term goal is to establish a platform indexing Mediterranean biodiversity data and environmental parameters which are of interest for many researchers. This index will employ the tools and methods recommended at both national (SINP [National Information System on Biodiversity], RBDD [Network of Research Databases]) and international levels (MedOBIS [Mediterranean Ocean Biogeographic Information System], OBIS, GBIF, Life-Watch, GEO-BON, etc.) along with other research entities (i.e. IRD [Institute of Research for the Development] or MNHN [National Museum of Natural History]).

The architecture of the proposed prototype of information systems for projects being developed is decentralized, and is a first step towards indexing, classifying, mapping and interfacing data from coastal and marine Mediterranean environments, essential in ecology research and natural spaces management tools.

Building on the efficiency of this prototype, the project will develop an "object resolution service" (*i.e.* a web service that finds links and dependencies among indexed objects, based on unique objects identification). This object resolution service should allow an inventory of biodiversity descriptors, estimate the capacity of data to describe socio-ecological systems at temporal and geographic scales, and must permit links between economic and social approaches. It is based on large panels of participants, data and skills being developed. The project should also develop new trans-disciplinary methods of data analysis, focusing on open data, open source and free methods and development tools.

II. METHOD

A. Decentralized information systems

Integrating databases into one large centralized database has been attempted by many programs, but always failed. Then, we must ensure our ecological, economics and social sciences data are interoperable, connectable and comparable, and to manage to analyze them without placing one above the other.

The principle of information systems (IS) decentralization is unavoidable once one looks at the real-time data analysis produced by different actors in various fields. Whether it is used for biodiversity studies or for the knowledge of socio-ecological systems, the production of data is expensive and rarely automated. The long time series and/or large spatial extent studies are difficult to conduct, and when dealing to "interpretive data" the use of too many observers affect the reliability of the observation.

Reproducibility today is frequently questioned or even refuted. In the context of multi-source data production, it is essential to help each producer as well as external users to install and maintain a suited IS for their needs (maintenance, development of software, developments on the data scheme, scope and standards of data). This decentralization requires working on data models, their evolution, but must remain consistent with the data collection protocols and their evolution.

A "modular" organization (for administering a type of object or data independently by the most competent actor) is preferred to centralized systems: in this interdisciplinary framework, based on systems observing at large scale, each participant cannot consolidate data from all disciplines. The data serve as a model and concern marine habitat and common terrestrial habitat for all disciplines. This type of methodology may be declined on many environmental models, including but not limited to terrestrial and marine habitats, organisms, communities and species assemblages.

Standardization makes possible such a work, as well as a special task on interoperability qualities and accessibility of non-centralized data. It use aggregation for public display, multi-interface, multi-use and multi-format, and must permit (i) the connection between many databases, and (ii) the preparation of inter-calibration works.

B. Indexing data

Indexing is an alternative to centralization, which allows global approaches in ecology. Global ecology considers ecological systems and their complexity in terms of composition, structure and interactions. It identifies the parameters able to lay the foundations for sustainable management of resources and services they provide, to better understand and anticipate the risks and their consequences, and to participate in the improvement of the quality of life of societies.

The IS data is the keystone of the linkage of different databases formats. It provides unique identifiers for data objects which can return metadata about themselves and which can be brought together into a distributed collaborative information system as recommended by GBIF [6]. It identifies each record, recordings of each state (version), and solving for each of these states all the data that have been used and their condition, in order to reach a new transformed state. It allows describing all the previous states, thus ensuring provenance, traceability and intellectual property of this data if it exists, to identify the adjectives that this new data can or cannot inherit, and thus it complements criteria that may serve as an additional descriptor for data mining. The resolution of these indexed web service relies on the unique identifiers of databases, creates some where none already exist, and also creates relationships between them where there is more than one.

This Information System allows making an informed choice of data aggregation as nodes, because it does not contain data considered as sensitive by a data producer. These indexing nodes will be "clonable" on a discretionary basis with enrichment rules and sharing licenses corresponding to "creative common" type "sharing the same conditions", allowing others to copy, distribute and modify the index (Fig 2), provided they publish express any adaptation of the index under the same conditions (open-source, open data). These rules will encourage the emergence of standards to improve the interoperability of data and promote the participation of new contributor laboratories taking into account their contribution to the technical possibilities as and when the project develops.

The prerequisites of these tasks are accessibility of data, normalized and qualified flow with open data accessible by means of fluxes. Tools to be improved are resolution service, unique identifiers, equivalence between identifiers, and reproducible indexing nodes. The main objective is availability and traceability of data; it should permit a heritability for data qualifiers in the future, stories of data / data life studies in ecology and associated disciplinary, mainly economics, sociology and law.

C. Quality and qualification of data

In a production framework of multi-source data, the equivalence of observation system's problematics and inter-calibration of observers then become crucial. Increasingly, the need for integrative multi- or trans-disciplinary approaches becomes necessary, in the study of systems where data output in each discipline is discontinuous, imprecise and badly distributed.

Yet all the variables of these systems interact in time and at each spatial scale (biotic, abiotic variables, anthropogenic and natural pressures, perceived and provided services, societal perception, etc.) [4] [9] [13]. To prepare a true integrated management biodiversity [10], the data collection protocols must allow a fine and standardized description with a shared typology about (i) conditions of measurement and ecological status, (ii) pressures on these areas, and (iii) economic and sociological context. All this knowledge is necessary and should be standardized on a large scale to enable decision support.

Scientific challenges about data quality are complicated, given (i) their volume and the dynamics of their update, the update repositories and the standards necessary for administering the data, (ii) their intrinsic heterogeneity and complexity, especially related to cross biodiversity data and contextual variables, (iii) the heterogeneity of users, networks of producers actors and their motivations to maintain and supply their information systems.

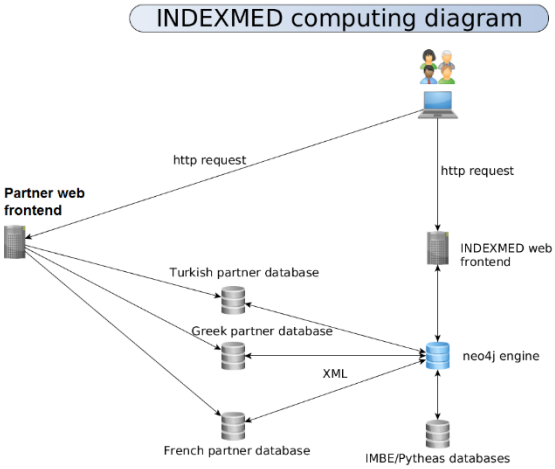


Figure 2. Global hardware architecture: it provides graph results from local or remote databases. Each partner will store data locally and the system will be able to get results using standard protocols. Graph results can be obtained by end users using a specific web frontend. 3 countries are testing the architecture under European programs and inventories of marine biodiversity (CIGESMED, DEVOTES, ZNIEFF)
In addition, queries on remote databases are being tested (*e.g.* GBIF (species occurrences), Tela Botanica (French Flora), European Pollen databases. Other thematic databases will be integrated in the development of this project.

Work on data quality and their equivalence is required. It firstly involves the analysis and description of the common elements of each piece of information, and what differentiates them (name fields, formats, update rate, precision, observers or sensors, etc.) These descriptions are added to the data and form a body of criteria used for data mining. Secondly, it is intended to give the equivalence of data, based on data dictionaries and thesaurus. Some database conjunctions allow to deduce other, using firstly their own ontology for each domain and multidisciplinary. Out of all these logical relationships, we can deduce new qualifiers that are either new data quality or a way to find common qualifier to heterogeneous data that can serve as an additional descriptor as part of the data mining.

This work is possible by accessible and decentralized IS, commonly related by other systems and normalized, accessible and configurable stream of qualifications. These standards/data dictionary/thesaurus ontology are

commonly constructed and will be improved over the next years in an open database. This work will allow jointed analysis of different data corpus, and inter-calibrations of data productions.

III. First Results

A. Data visualization and graphs

The visualization prototype is initiated in the VIGI-GEEK project (VIsualisation of Graph In transdisciplinary Global Ecology, Economy and Sociology data-Kernel). It currently involves representations of graphs, but in the coming years it will explore other types of heterogeneous data representations.

A major goal is to produce a "multidisciplinary" tool for construction and graph visualization through IndexMed. These graphs are constructed from aggregated information through the nodes indexing and qualifying data concerning Mediterranean marine and coastal environment, in various disciplines (socio-ecology, econometrics, ecology [structural and operational], town planning, management, etc.) at the Mediterranean scale. The development of this prototype is to make customizable graphs to search and visualize multidisciplinary data putting on the same level socio-ecological, economical, ecological, molecular and functional data types (trophic relationships, functional traits, etc.) Each object or concept describing biodiversity can be a node, and the terms of the attributes that describe them are as many links which can give specific properties (attraction or repulsion of other objects). Particular graphs can mix heterogeneous objects, if these different objects share identical terms for any attributes (Fig. 3).

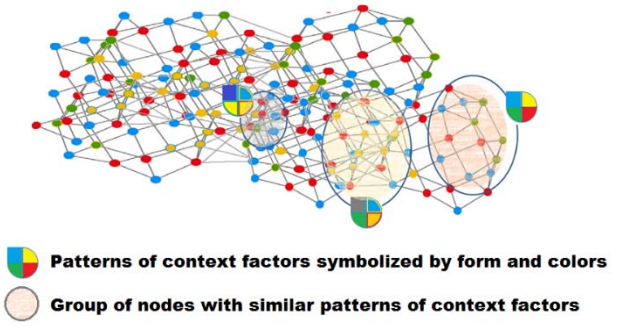


Figure 3. IndexMed project graphic sample using IMBE dataset: different objects (3 in this case: species, quadrat photo samplings and localities) can be linked if they share the same attribute modality. This interface allows a generic integration of different objects when links are possible with common attribute values. This enables the design of graphs which are then analyzed thanks to the stream generated in JSON or XML. This flow will be operated by the algorithms selected by a decision tree, depending on the type of objects generated graph (not yet developed). The computing infrastructure will be used to browse these graphs.

The interface use Neo4j <neo4j.com/>, a graph database implemented in java and released in 2010. The community edition of the database is licensed under the free GNU General Public License (GPL) v3. The database and its additional modules (online backup or high availability) are available under a commercial license.

Neo4j is used to represent data as objects connected by a set of relations, each object having its own properties. When the database is requested, a graph appears and it is possible to interact with it, using the web browser.

In Neo4j, everything is stored in form of either edge, node or attribute. Each node and edge can have any number of attributes. Both the nodes and the edges can be labelled and colored.

IndexMed technical staff is developing a specific web frontend using Ajax/Jquery language. It may be possible to request a database asking for specific objects and specific relations between them, without using a technical query language such as SQL or Cypher. The prototype is developed for a generic and is enough to integrate any type of data in the form of "object, attribute, and attribute value" (Fig. 4).

This prototype will be available as an open-source format to develop, on the medium term, usage of these graphs for decision support in environmental management and as part of a research project to be submitted to European calls for projects (BiodivERsA, ERDF, SeasEra, H2020…)

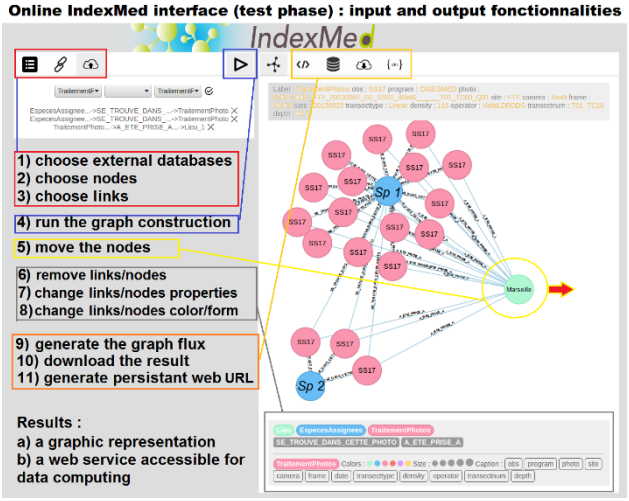


Figure 4. Data visualization coming from our specific web frontend: different types of objects can be mixed, and relation between them is due to the value of terms which describe them. Links and nodes can be selected or colored by values of term. Fluxes content and generated request are display in the front of this interface. At this state, databases on marine benthic environment, terrestrial flora and pollen databases are tested. Others data content (dendrology, archeology, economic and social data) will be included after works on common thesaurus and attributes value (which permit to construct links).

B. Testing other data devices: the astronomy example ("CHARLIEE" project with IndexMed consortium)

Astronomy is a good example of interoperability of data and associated service. Since the first works of Messier and the NGC catalogues (first typology of astronomic objects in 1888), all the extragalactic astronomical objects have been identified and have been associated with a unique identifier. Each new object is defined with a specific ID, and with a combination of its position in the sky at a given time and epoch of the observation.

In 2002 the International Virtual Observatory Alliance (IVOA <www.ivoa.net>) has been created to gather efforts on data standardization and dissemination. Since then, the Virtual Observatory (VO) allowed to spread validated data

all over the world and to use data from everywhere from earth. Infrastructure, standards and tools have been developed to easily search for data and use and export all structured objects. From Solar system and stellar objects to galactic objects, theoretical data and tabular data are covering several quantities such as astrometric, photometric, spectroscopic data. Most of the characteristics of the objects have been characterized and formatted. Format exchanges have been described as data models and serialized as XML formatted data. By the same time, access protocols have been setup to access images, spectra, tabular and temporal data. For several years, software is being developed to facilitate the cross-use and discovery of astronomical data. Among the most used software, one can cite:
ALADIN (http://aladin.u-strasbg.fr/)
TOPCAT (http://www.star.bris.ac.uk/~mbt/topcat/)
VOSPEC (http://www.sciops.esa.int/index.php?project=SAT&page=vospec)
Such tools hide the complexity of the VO infrastructure, to facilitate the use of data.

The goal is to transcript the ecological data into VO formatted data in order to use the existing astronomical tools in the study. Data will be translated using Unified Content Descriptors (UCDs). This will allow the use of astronomical tools in a comparative way. For example we can use the density maps (Fig 5).

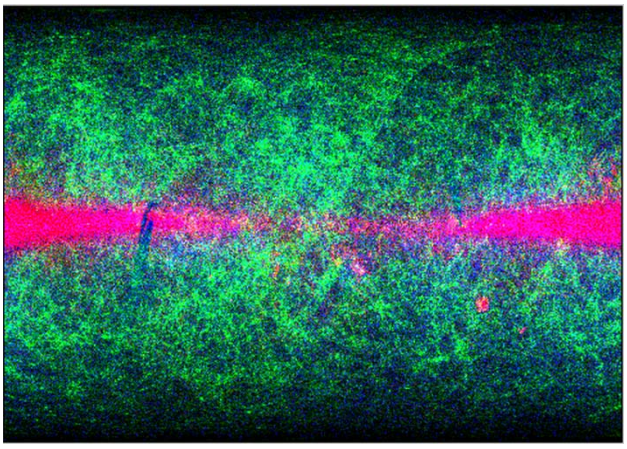


Figure 5. TOPCAT density map. The intensity of color at each pixel reflects the number of points that fall within its bounds. TOPCAT is an interactive graphical viewer and editor for tabular data. It presently provides most of the facilities that astronomers need for analysis and manipulation of source catalogues and other tables. This is very useful for analysis and visualization of extremely crowded plots. In the IndexMed framework, TOPCAT will be used for ecological, economical and/or sociological data visual representation. (http://www.starlink.ac.uk/topcat/)

IV. DISCUSSION AND PERSPECTIVES

In such an *environmental* framework, the development of a network that integrates researchers in Humanities and Social Sciences across the Mediterranean basin is essential. Indeed, the question of the compatibility of models and data from different disciplines is a key issue for building models and measuring relevant data within each discipline which follows logics and different rules. Some workshops will build-up protocols compatibility between models and ecological, social and economic data, as well as to identify priority actions for data access to be inserted in the graphs. Major tasks to be carried out after indexing and building the first graphs are to (i) build ontology and their implementations in existing graphs, (ii) create data mining tools with heterogeneous data in ecology and (iii) build a bridge between artificial intelligence tools used to analyze graphs and decision support. The decisions support objectives of the project also need to consider ways of preserving the short and long term data and thus the generated qualifiers.

A. Ontologies for complex graphs approaches

Ontology is a specification of a conceptualization for a domain of knowledge [18] and therefore it results of a choice to formally describe this area, based on a controlled vocabulary, creating dependencies and inheritance between these concepts.

Biodiversity is a multiple field of knowledge where concepts and data abound [28]. For example, it has recently been described in GEO-BON as 22 essential variables [33]. Translation into ontologies is essential for a better qualification of these data [25], to show new links between different objects that make it up (housing, key-stone species, predator species, etc.) These links give additional properties and dimensions to graphs constructed under IndexMed and increase the perimeters of the investigation areas. The use of ontologies within IndexMed project will build on existing eco-informatics and ongoing work related to thesaurus [20] [22] [42].

B. Data mining tools with heterogeneous data in ecology

Data mining emerged in the late 90s [8] as a discipline to extract relevant, novel and understandable knowledge from the analysis of datasets, easier to record, but that refer to increasingly complex phenomenon, among which we find ecology.

A main issue in this field is the need to analyze the relationships between heterogeneous data for a proper understanding of reality. The data mining on heterogeneous data is complex and convoluted by a longer work on quality data. However, it must allow the extraction of relevant decisional knowledge or acquaintance from large amounts of data, using supervised or fully automatic algorithms.

Acquiring a better global understanding of the balance of Socio-Ecological Systems (SES) and their impact on biodiversity is one of the main current scientific challenge. Advances in this understanding process must consider the construction and testing of methods for joint interpretation of these heterogeneous data.

Some research systems started to present logical inter-dependencies in SES for facilitating building of biodiversity and ecosystems services [23]. New opportunities are created by open data formats in ecology [38] and qualification standards usable in data mining are developed with the Taxonomic Databases Working Group consortium <www.tdwg.org> by the Darwin Core Task Group [43]. Some works focus on the integration of declarative knowledge with numerical and qualitative data [14] or on the post-process of results required to provide understandable knowledge to the end-user [5] [15]. Data mining methods must be able to bring new perspectives to the disciplinary research on these complex systems, studying ultimately interrelated objects (environmental chemistry, genomics, transcriptomic, proteomics, metabolomics, stands ecology, socio-ecological systems, or landscape ecology are some examples) as well as

dealing with the intrinsic space-time of the ecological phenomenon. It will use indexed data, data qualification, and data traceability for discovering patterns in the data values conjunctions with scientific significance. In the IndexMed Project, supervised clustering, graph algorithms, statistical ecology [17] and collaborative clustering methods [12] are planned to be used. Another issue is to use "unsupervised" mode, raising the possibility to compare the results of different algorithms to achieve consensus, which acts / results in the most likely scenario. The data mining helps finding managerial values such as scenarios, and provides standardized descriptors essential for approaches such as machine learning. The ambition of IndexMed consortium is to achieve operational objectives in terms of decision support, based on the exploitation of these complex multidisciplinary graphs. Scientific challenges concerning the quality of data are complicated by the heterogeneity of sometimes contradictory norms and standards between disciplines. However, adopting too many standards can prevent the pooling of heterogeneous objects. A process of simplification and approximation will probably be necessary.

The ongoing more advanced initiatives concerning open linked data construct analyses with data mostly linked by geographical approaches (see [40]). One challenge raised by the consortium IndexMed consists to achieve common ontologies between several disciplines, and to offer new dimensions of analysis beyond the usual geographical and time fields (see [1]), usable by intelligent decision support machine.

C. From artificial intelligence to intelligent decision support

The area of Decision Support Systems (DSS) focuses on development of interactive software that can analyze data and provides answers to relevant decisional questions from the users, thus enhancing a person or group to make better decisions. Early DSS [24] used simple monitoring; later, model-based simulation introduced what-if analyses. Intelligent DSS (IDSS) [26] included specific domain knowledge and automatic reasoning capabilities. Till now, important efforts to develop dedicated (I)DSS are required for every particular application [41][36] and some successful experiences appear in several fields, like self-care management [27], water management [32], forest ecosystems [30] or air pollution [31]. However, upgrading of these platforms to incorporate new risk factors control, new sensors connections or to take into account new predictive models becomes costly and time consuming.

New generation IDSS provides sufficient integration for achieving a really holistic approach (taking into account not only monitoring of isolated parameters, but also information from the different data sources available, activities developed in the community and all types of available information (images, measures, qualitative data, knowledge, documents, tweet, etc.) and to get a sufficiently flexible DSS system architecture to make easy adaptation of the system to advances in the ecological state of the art, and new architectures must be designed to permit flexible upgrades or domain-changes of these kind of platforms in an easy way [3] [29]. Today, it seems clear [35] [37] [16] that IDSS must combine data-driven, analytical and knowledge-based models (including prior declarative expert knowledge), as well as some standardized reasoning [21] [19] [39] to provide a relevant support to managers, even if there are not yet much real experiences on-going under this approach.

Ecology belongs to a set of critical domains where wrong decisions may have tragic consequences. Decision-making performed by IEDSSs should be collaborative, not adversarial, not only finding optimal or suboptimal solutions, but making the entire process more open and transparent. The system will have to deal with inherent uncertainty of decisions and decisions must inform and involve those who must live with the (consequences of the) decisions.

D. Preservation of heterogeneous data

The scientific data is collected with large material and human efforts and tend to be unique due to the ever increasing experimental complexity and because of the time-stamped nature of the data itself, as it is very often in ecology studies. The data preservation is therefore of crucial importance and may open new avenues for research at low cost, and for instance when older data sets are combined with newer data in order to enhance the precision or to detect time-dependent variations. Various disciplines have organized the preservation of larger or smaller samples of experiments in a different manner, most of the time adjusted to the community needs [6]. However, there is a large potential for improvement and unless an immediate and massive action is started, the danger to simply loosing unique data sets remains significant. Investigations in a multi-disciplinary context have revealed many similarities across heterogeneous data sets within some scientific disciplines; for instance, several experiments in high-energy physics may choose to standardize their preservation approaches [2] and astrophysics data is massively mutualized in the IVOA. Moreover, there are similarities and constructive complementarities in the scientific needs and methods for data preservation across different scientific disciplines. These arguments suggest that a rigorous approach of the heterogeneous data treatment, from collection, treatment and access to mining, visualization and storage can also be beneficial for the long term preservation. In addition, the preservation of data sets already collected can be reorganized through novel algorithms in order to enhance the robustness of the data preservation systems, thereby extracting more science and enhancing the original investment in experiments and data collection. Due to these considerations, this is one of IndexMed Consortium goals.

ACKNOWLEDGMENT

The construction of the first prototype for IndexMed consortium is funded by the *CNRS défi* "VIGI-GEEK (VIsualisation of Graph In transdisciplinary Global Ecology, Economy and Sociology data-Kernel)" and CNRS INEE with the "CHARLIEE" project. Data used for this article was obtained through the CIGESMED project [11] <www.cigesmed.eu>.

We acknowledge all the field helpers and students who have participated in data collection in the field and in the lab, as well as in data management. Thanks are also due to Dr D. Vauzour for improving the English text.

REFERENCES

[1] F.K. Amanqui, K.J. Serique, S.D. Cardoso, J.L. dos Santos, A. Albuquerque and D.A. Moreira, "Improving Biodiversity Data Retrieval through Semantic Search and Ontologies," in Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Conference on Web Intelligence, 11-14 August 2014, Warsaw, Poland, vol.1 (WI), pp.274-281, doi: 10.1109/WI-IAT.2014.44

[2] Z. Akopov, S. Amerio, D. Asner, E. Avetisyan, O. Barring, J. Beacham, M. Bellis, G. Bernardi, S. Bethke, A. Boehnlein, T. Brooks, T. Browder, R. Brun, C. Cartaro, M. Cattaneo, G. Chen, D. Corney, K. Cranmer, R. Culbertson, S. Dallmeier-Tiessen, D. Denisov, C. Diaconu, V. Dodonov, T. Doyle, G. Dubois-Felsmann, M. Ernst, M. Gasthuber, A. Geiser, F. Gianotti, P. Giubellino, A. Golutvin, J. Gordon, V. Guelzow, T. Hara, H. Hayashii, A. Heiss, F. Hemmer, F. Hernandez, G. Heyes, A. Holzner, P. Igo-Kemenes, T. Iijima, J. Incandela, R. Jones, Y. Kemp, K. Kleese van Dam, J. Knobloch, D. Kreincik, K. Lassila-Perini, F. Le Diberder, S. Levonian, A. Levy, Q. Li, B. Lobodzinski, M. Maggi, J. Malka, S. Mele, R. Mount, H. Neal, J. Olsson, D. Ozerov, L. Piilonen, G. Punzi, K. Regimbal, D. Riley, M. Roney, R. Roser, T. Ruf, Y. Sakai, T. Sasaki, G. Schnell, M. Schroeder, Y. Schutz, J. Shiers, T. Smith, R. Snider, D.M. South, R. St. Denis, M. Steder, J. Van Wezel, E. Varnes, M. Votava, Y. Wang, D. Weygand, V. White, K. Wichmann, S. Wolbers, M. Yamauchi, I. Yavin, H. von der Schmitt [DPHEP Study Group]. "Status Report of the: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics", DPHEP Study Group Collaboration, May 2012, 93 pp., DPHEP-2012-001, FERMILAB-PUB-12-878-PPD, e-Print: arXiv: 1205.4667 [hep-ex].

[3] O.J. Bott, M. Marschollek, K.H. Wolf and R. Haux, "Towards new scopes: sensorenhanced regional health information systems - part 1: architectural challenges". Methods of Informations in Medecine vol. 46(4), 2007, pp. 476–483.

[4] N. Conruyt, D. Sébastien, S. Cosadia, R. Vignes-Lebbe and T. Touraivane, "Moving from biodiversity information systems to biodiversity information services". In: Information and Communication Technologies for Biodiversity, Conservation and Agriculture, L. Maurer and K. Tochtermann (Eds.). Shaker Verlag: Aachen, August 2010.

[5] P. Cortez and M.J. Embrechts "Using sensitivity analysis and visualization techniques to open black box data mining models". Information Sciences, vol. 225, March 2013, pp. 1-17, doi:10.1016/j.ins.2012.10.039

[6] P. Cryer, R. Hyam, C. Miller, N. Nicolson, É.Ó Tuama, R. Page, J. Rees, G. Riccardi, K. Richards, and R. White. "Adoption of persistent identifiers for biodiversity informatics: Recommendations of the GBIF LSID GUID task group", 6 November 2009. Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark (version 1.1, last updated 21 Jan 2010) 62.

[7] C. Diaconu [Ed.]. PREDON group, "Scientific Data Preservation 2014", February 2014, 61pp., http://predon.org

[8] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery: An overview". In Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Fall 1996, pp.37-54.

[9] J.-P. Féral [Ed.], "Concepts and methods for studying marine biodiversity, from gene to ecosystem". Océanis, documents océanographiques, vol. 24(4) [1998], Institut Océanographique, Paris, March 2001, 420 pp., ISSN: 0182-0745.

[10] J.-P. Féral and R. David, "L'environnement, un système global dynamique. - 22. Zone côtière et développement durable, une équation à résoudre". In: Le développement durable à découvert, A. Euzen, L. Eymard, F. Gaill (Eds), CNRS éditions: Paris, September 2013, pp. 96-97, ISBN : 978-2-271-07896-4

[11] J.-P. Féral, C. Arvanitidis, A. Chenuil, M.E. Çinar, R. David, A. Frémaux, D. Koutsoubas and S. Sartoretto, "CIGESMED, Coralligenous based Indicators to evaluate and monitor the « Good Environmental Status » of the MEDiterranean coastal waters, a SeasEra project (www.cigesmed.eu)". Proceedings RAC/SPA 2nd Mediterranean Symposium on the Conservation of coralligenous and other calcareous bio-concretions, Portorož, Slovenia, October 2014, pp. 15-21

[12] G. Forestier, C. Wemmert and P. Gançarski, "Multi-source Images Analysis Using Collaborative Clustering". EURASIP Journal on Advances in Signal Processing, Special issue on Machine Learning in Image Processing, 2008, Article ID 374095, 11 pp., doi:10.1155/2008/374095

[13] S. Gachet, E. Véla and T. Tatoni, "BASECO: a floristic and ecological database of Mediterranean French flora". Biodiversity and Conservation, vol. 14, April 2005, pp.1023-1034, doi: 10.1007/s10531-004-8411-5

[14] K. Gibert, A. Valls and M. Batet, "Introducing semantic variables in mixed distance measures. Impact on hierarchical clustering", Knowledge and Information Systems, vol. 40(3), September 2014, pp. 559-593, doi: 10.1007/s10115-013-0663-5

[15] K. Gibert, D. Conti and D. Vrecko, "Assisting the end-user in the interpretation of profiles for decision support. An application to wastewater treatment plants", Environmental Engineering and Management Journal , vol. 11(5), May 2012, pp. 931-944

[16] K. Gibert, M. Sánchez-Marrè and I. Rodríguez-Roda, "GESCON-DA: An intelligent data analysis system for knowledge discovery and management in environemntal databases". Environmental modelling and software, vol. 21(1), January 2006, pp. 115-120, doi:10.1016/j.envsoft. 2005.01.004

[17] O. Gimenez, S.T. Buckland, B.J.T. Morgan, N. Bez, S. Bertrand, R. Choquet, S. Dray, M.P. Etienne, R. Fewster, F. Gosselin, B. Mérigot, P. Monestiez, J. Morales, F. Mortier, F. Munoz, O. Ovaskainen, S. Pavoine, R. Pradel, F.M. Schurr, L. Thomas, W., Thuiller, V. Trenkel, P. de Valpine and E. Rexstad, "Statistical ecology comes of age". Biology Letters, vol. 10, December 2014, 4 pp., doi: 10.1098/ rsbl.2014.0698.

[18] T.R. Gruber, "A translation approach to portable ontology specifications". Knowledge Acquisition, vol 5(2), June 1993, pp. 199-220, http://dx.doi.org/10.1006/knac.1993.1008.

[19] A. Helmer, B. Song, W. Ludwig, M. Schulze, M. Eichelberg, A. Hein, U. Tegtbur, R. Kayser, R. Haux and M. Marschollek, "A sensor-enhanced health information system to support automatically controlled exercise training of COPD patients". In: 4th International Conference on Pervasive Computing Technologies for Healthcare. Munich: IEEE, 22-25 March 2010, pp. 1-6, doi: 10.4108/ CSTPERVASIVEHEALTH2010.8827

[20] J. Kattge, S. Díaz, S. Lavorel, I.C. Prentice, P. Leadley, G. Bönisch, E. Garnier, M. Westoby, P.B. Reich, I.J. Wright, J.H.C. Cornelissen, C. Violle, S.P. Harrison, P.M. van Bodegom, M. Reichstein, N.A. Soudzilovskaia, D.D. Ackerly, M. Anand, O. Atkin, M. Bahn, T.R. Baker, D. Baldocchi, R. Bekker, C. Blanco, B. Blonder, W. Bond, R. Bradstock, D.E. Bunker, F. Casanoves, J. Cavender-Bares, J. Chambers, F.S.I. Chapin, J. Chave, D. Coomes, W.K. Cornwell, J.M. Craine, B.H. Dobrin, W. Durka, J. Elser, B.J. Enquist, G. Esser, M. Estiarte, W.F. Fagan, J. Fang, F. Fernández, A. Fidelis, B. Finegan, O. Flores, H. Ford, D. Frank, G.T. Freschet, N.M. Fyllas, R. Gallagher, W. Green, A.G. Gutierrez, T. Hickler, S. Higgins, J.G. Hodgson, A. Jalili, S. Jansen, A.J. Kerkhoff, D. Kirkup, K. Kitajima, M. Kleyer, S. Klotz, J.M.H. Knops, K. Kramer, I. Kühn, H. Kurokawa, D. Laughlin, T.D. Lee, M. Leishman, F. Lens, T. Lenz, S.L. Lewis, J. Lloyd, J. Llusià, F. Louault, S. Ma, M.D. Mahecha, P. Manning, T. Massad, B. Medlyn, J. Messier, A. Moles, S. Müller, K. Nadrowski, S. Naeem, Ü. Niinemets, S. Nöllert, A. Nüske, R. Ogaya, J. Oleksyn, V.G. Onipchenko, Y. Onoda, J. Ordoñez, G. Overbeck, W. Ozinga, S. Patiño, S. Paula, J.G. Pausas, J. Peñuelas, O.L.

Phillips, V. Pillar, H. Poorter, L. Poorter, P. Poschlod, R. Proulx, A. Rammig, S. Reinsch, B. Reu, L. Sack, B. Salgado, J. Sardans, S. Shiodera, B. Shipley, E. Sosinski, J.-F. Soussana, E. Swaine, N. Swenson, K. Thompson, P. Thornton, M. Waldram, E. Weiher, M. White, S.J. Wright, S. Zaehle, A.E. Zanne and C. Wirth, "TRY – a global database of plant traits". Global Change Biology, vol. 17, June 2011, pp. 2905-2935, doi: 10.1111/j.1365-2486.2011.02451.x

[21] S. Koch and M. Hagglund, "Health informatics and the delivery of care to older people". Maturitas, vol. 63(3), July 2009, pp.195-199.

[22] M.A. Laporte, I. Mougenot and E. Garnier, "ThesauForm – Traits: a web based collaborative tool to develop a thesaurus for plant functional diversity research". Ecological Informatics, vol. 11, September 2012, pp. 34-44, doi:10.1016/j.ecoinf.2012.04.004

[23] M.A. Laporte, I. Mougenot and E. Garnier, U. Stahl, L. Maicher and J. Kattge, "A semantic web faceted search system for facilitating building of biodiversity and ecosystems services". In: H. Galhardas & E. Rahm [Eds], Data Integration in the Life Sciences, DILS 2014, pp. 50-57. Springer, Switzerland, doi: 10.1007/978-3-319-08590-6_5

[24] J.D. Little, "Models and Managers: The Concept of a Decision Calculus". Management Science, vol. 16(8), April 1970, B-466-B485.

[25] J.S. Madin, S. Bowers, M.P. Schildhauer and M.B. Jones, "Advancing ecological research with ontologies". Trends in Ecology and Evolution, vol. 23(3), March 2008, pp. 159-168.

[26] G. M. Marakas, Decision support systems in the twenty-first century, 1999, Prentice Hall, Inc. Upper Saddle River, N.J., ISBN:0-13-744186-X

[27] M. Marschollek, "Decision support at home (DS@ HOME)–system architectures and requirements." BMC medical informatics and decision making, May 2012, 12/43, 8 pp., doi: 10.1186/1472-6947-12-43.

[28] W.K. Michener and M.B. Jones, "Ecoinformatics: supporting ecology as a data-intensive science". Trends in Ecology & Evolution, vol. 27(2), February 2012, pp. 85-93, doi: 10.1016/ j.tree.2011.11.016

[29] J. Misyak, C. Giupponi and P. Rosato, "Towards the development of a decision support system for water resource management". Environmental Modelling and Software, vol. 20(2), February 2005, pp. 203-214, doi: 10.1016/j.envsoft.2003.12.019

[30] D. Nute, W.D. Potter, F. Maier, J. Wang, M. Twery, H.M. Rauscher, P. Knopp, S. Thomasma, M. Dass, H. Uchiyama and A. Glende, "NED-2: an agent-based decision support system for forest ecosystem management". Environmental Modelling & Software, vol. 19(9), September 2004, pp. 831-843, doi: 10.1016/ j.envsoft.2003.03.002.

[31] M. Oprea, M. Sanchez-Marré and F. Wotawa, "A case study of knowledge modelling in an air pollution control decision support system". AI Communications, Binding Environmental Sciences and AI, vol. 18(4), December 2005, pp. 293-303, ISSN:0921-7126

[32] S. Pallottino, G.M. Sechi and P. Zuddas, "A DSS for water resources management under uncertainty by scenario analysis", Environmental Modelling & Software, vol. 20(8), August 2005, pp. 1031-1042, doi: 10.1016/j.envsoft.2004.09.012.

[33] H.M. Pereira, S. Ferrier, M. Walters, G.N. Geller, R.H. Jongman, R.J. Scholes, M.W. Bruford, N. Brummitt, S.H. Butchart, A.C. Cardoso, N.C. Coops, E. Dulloo, D.P. Faith, J. Freyhof, R.D. Gregory, C. Heip, R. Höft, G. Hurtt, W. Jetz, D.S. Karp, M.A. McGeoch, D. Obura, Y. Onoda, N. Pettorelli, B. Reyers, R. Sayre, J.P. Scharlemann, S.N. Stuart, E. Turak, M. Walpole, M. Wegmann, "Essential biodiversity variables". Science, vol. 339(6117), January 2013, pp. 277-278. doi: 10.1126/science.1229931.

[34] D.P.C. Peters, K.M. Havstad, J. Cushing, C. Tweedie, O. Fuentes, and N. Villanueva-Rosales, "Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology". Ecosphere, vol. 5(6), Art. 67, June 2014, 15 pp., http://dx.doi.org/10.1890/ ES13-00359.1

[35] M. Poch, J. Comas, I. R-Roda, M. Sànchez-Marrè and U. Cortés, "Designing and building real environmental decision support", Systems Environmental Modelling & Software, vol. 19(9), September 2004, pp. 857-873, doi: 10.1016/j.envsoft.2003.03.007

[36] D.J. Power, "A Brief History of Decision Support Systems", DSSResources.COM (Editor), World Wide Web, version 4.0", March 2007, http://dssresources.com/history/ dsshistory.html.

[37] V. Rajasekaram and K.D.W. Nandalal "Decision Support system for Reservoir Water management conflict resolution". Journal of water resources planning and management, vol. 131(6), November 2005, pp. 1-10, doi: 10.1061/(ASCE)0733-9496(2005)131:6(410).

[38] O.J. Reichman, M.B. Jones and M.P. Schildhauer, "Challenges and opportunities of open data in ecology". Science, vol. 331(6018), February 2011, pp. 703-705, doi: 10.1126/science.1197962

[39] M. Sànchez-Marrè and K. Gibert, "Improving ontological knowledge with reinforcement in recommending the data mining method for real problems". In Proceedings of Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA), Albacete, 9-12 November 2015, (*in press*).

[40] S.D. Cardoso,· F.K. Amanqui, K.J.A. Serique, J.L.C. dos Santos, D.A. Moreira, "SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data". Future Generation Computer Systems, vol. 54, January 2016, pp. 389-398, doi: 10.1016/j.future.2015. 05.006

[41] U. Varanon, C.W. Chan and P. Tontiwachwuthikul, "Artificial Intelligence for monitoring and supervisory control of process systems". Engineering applications of artificial intelligence, vol. 20(2), March 2007, pp. 115-131, doi: 10.1016/j.engappai.2006. 07.002

[42] R.L. Walls, J. Deck, R. Guralnick, S. Baskauf, R. Beaman, S. Blum, S. Bowers, P.L. Buttigieg, N. Davies, D. Endresen, M.A. Gandolfo, R. Hanner, A. Janning, L. Krishtalka, A. Matsunaga, P. Midford, N. Morrison, É.Ó. Tuama, M. Schildhauer, B. Smith, B.J. Stucky, A. Thomer, J. Wieczorek, J. Whitacre and J. Wooley, "Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies". PLoS ONE, vol. 9(3), e89606, March 2014, doi: 10.1371/journal.pone.0089606

[43] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson and D.Vieglais, "Darwin Core: An evolving community-developed biodiversity data standard". PLoS ONE, vol. 7(1), e29715, January 2012, doi:10.1371/journal.pone. 0029715.

International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

-

IndexMed projects: new tools using the CIGESMED DataBase on Coralligenous for indexing, visualizing and data mining based on graphs

Romain David[1], Jean-Pierre Féral[1], Anne-Sophie Archambeau[2], Nicolas Bailly[3], Cyrille Blanpain[4], Vincent Breton[5], Aurélien De Jode[1], Aurélie Delavaud[6], Alrick Dias[1], Sophie Gachet[1], Dorian Guillemain[1], Julien Lecubin[4], Geneviève Romier[5], Christian Surace[7], Laure Thierry de Ville d'Avray[1], Christos Arvanitidis[3], Anne Chenuil[1], Melih Ertan Çinar[8], Drosos Koutsoubas[3,9], Stéphane Sartoretto[10], Thierry Tatoni[1]

(1) Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE), CNRS, Aix Marseille Université, IRD, and Université d'Avignon, Station Marine d'Endoume, Chemin de la Batterie des Lions, 13007 Marseille, France. romain.david@imbe.fr, jean-pierre.feral@imbe.fr, anne.chenuil@imbe.fr, aurelien.dejode@imbe.fr, alrick.dias@imbe.fr, sophie.gachet@imbe.fr, dorian.guillemain@imbe.fr,laure.thierry@imbe.fr, thierry.tatoni@imbe.fr
(2) GBIF-France, MNHN, CP 48, 43 rue Buffon, 75005 Paris, France. archambeau@gbif.fr, gbif@gbif.fr
(3) HCMR/IMBBC Hellenic Centre for Marine Research, Institute of Marine Biology, Biotechnology & Aquaculture, LifeWatchGreece, Gouves, 71500 Heraklion, Crete, Greece. nbailly@hcmr.gr; arvanitidis@hcmr.gr
(4) Service informatique (SIP), OSU Pythéas, CNRS, Aix Marseille Université, 13007 Marseille, France. cyrille.blanpain@osupytheas.fr, julien.lecubin@osupytheas.fr
(5) Institut des Grilles et du cloud (IDG) France Grilles % LPC Clermont-Ferrand 4 avenue Blaise Pascal 63178 Aubière Cedex breton@idgrilles.fr , genevieve.romier@idgrilles.fr
(6) FRB ECOSCOPE - Pôle pour l'observation et la diffusion des données de recherche sur la biodiversité, Fondation pour la Recherche sur la Biodiversité, 195 rue Saint-Jacques - 75005 Paris, France. aurelie.delavaud@fondationbiodiversite.fr
(7) Laboratoire d'Astrophysique de Marseille (LAM), CNRS, Aix Marseille Université, rue Frédéric Joliot-Curie, 13013 Marseille, France. christian.surace@lam.fr
(8) Department of Hydrobiology, Faculty of Fisheries, Ege University, Bornova, Izmir, Turkey. melih.cinar@ege.edu.tr
(9) National Marine Park of Zakynthos, 29100, Zakynthos, and Dept. Marine Sciences, University of the Aegean, 81100 Mytilini, Greece drosos@aegean.gr
(10) IFREMER,Quartier Brégaillon, 83500 La Seyne-sur-Mer, France. Stephane.Sartoretto@ifremer.fr

Abstract: Data produced by the CIGESMED project (Coralligenous based Indicators to evaluate and monitor the "Good Environmental Status" of the MEDiterranean coastal waters) have a high potential for use by several stakeholders involved in environmental management. A new consortium called IndexMed whose task is to index Mediterranean biodiversity data, makes it possible to build graphs in order to analyse the CIGESMED data and develop new ways for data mining of coralligenous data. This paper presents the prototypes under development that test the ability of graphs approach to connect biodiversity objects with non-centralized data. This project explores the ability of two scientific communities to work together. The uses of data from coralligenous habitat demonstrate the prototype functionalities and introduce new perspectives to analyse environmental and societal responses.

Keywords*: data qualification, graph, thesaurus, distributed information system, Coralligenous habitats*

1 INTRODUCTION

1.1 Context : Big data and interoperability in ecology

Data mining emerged in the late 90s [Fayyad et al., 1996] as a discipline to extract relevant novel and understandable knowledge from the analysis of preexistent datasets and evolved to an increasingly complex approach which includes ecology, among other disciplines. Although currently it is considered

by most information producers and users in scientific disciplines and industry as the most promising way for making progress and leading to discovery, the use of Big Data in Ecology is still lagging behind from other disciplines [Peters et al., 2014]. In marine biodiversity and its connection with the coastal socio-ecological systems (SES), data production is still very expensive and with a low level of automation. Studies on long term data series and/or large spatial areas are difficult to conduct, and when it is necessary to involve several observers, it must be noted that the robustness and reproducibility of the observation is very often more difficult to obtain.
In a production framework of multi-source data in ecology, the equivalence of observation systems and inter-calibration become crucial. Increasingly, integrative transdisciplinary approaches become necessary in the study of systems where information in each discipline is patchy, imprecise and poorly distributed. Yet all variables (biotic, abiotic, anthropogenic and natural pressures, perceived and rendered ecosystem services, societal perception, etc.) of these systems interact in a wide range of spatio-temporal scales [Féral et al., 2001] [Gachet et al., 2005], [Conruyt et al., 2010]. Some research systems tried to bring out logical interdependencies in socio-ecological systems to facilitate the building of biodiversity and ecosystems services [Laporte et al., 2014]. Several authors and international initiatives also tried to specify, through a hierarchical approach of biodiversity [Noss, 1990], a common minimum set of variables to be measured, complementary to one another and covering the interlinked biodiversity organization levels. They should allow to capture, with current means and tools, the maximum possible information on biodiversity state and trends with the least effort [Pereira et al., 2013], [Kissling et al., 2015]. Similar initiatives are ongoing for climate, weather and ocean [Connecting GEO] to foster the discovery and the analysis of complementary data across spatial and temporal scales.
New opportunities are created by open data formats in ecology [Reichman etal., 2011] and qualification standards usable in data management are developed with the Biodiversity Informatics Standards (formerly Taxonomic Database Working Group) consortium <www.tdwg.org> (Darwin Core Task Group) [Wieczorek et al., 2012]. Other studies focus on the integration of declarative knowledge with numerical and qualitative data [Gibert et al., 2014] or on the post-process of results required to provide understandable knowledge to the end-user [Cortez et al., 2012] [Gibert et al., 2012].

Finally, methods for linking biodiversity and environmental data exist, but they are often limited to an "inventory" aspect of biodiversity (collection, observations, repositories and distribution) and neglect functional aspects. Initiatives like CoL [Catalogue of Life], Data-ONE [Data Observation Network for Earth], EMODnet [European Marine Observation and Data Network], GEO-BON [Group On Earth Observations Biodiversity Observation Network], EU-BON [European Biodiversity Observation Network], GBIF [Global Biodiversity Information Facilities], LifeWatch, OBIS [Ocean Biogeographic Information System], and TDWG [Biodiversity Information Standards] along with Darwin Core and ABCD [Access to Biological Collections Data], are well-known examples for achieving interoperability and standardizing data collection. However, integrative approaches in the coastal management zone need more interoperability at each scale [Féral and David, 2014].

1.2 Coralligenous habitat's case

The "coralligenous habitat", an endemic bioherm of the Mediterranean Sea, offers such a particularly complex case. Coralligenous habitats are difficult to study because they are patchy, not easily accessible (between 20 m and 120 m deep) and highly variable in local contexts [Ballesteros, 2006]. Due to these difficulties and the intrinsic complexity of this habitat type, comprehensive studies were rare until the 2000s [Laborel, 1961], [Laubier, 1966], [Hong, 1982], [Sartoretto, 1994]. Most of the proposed monitoring protocols / indicators for its ecological health are developed locally or regionally [Deter et al., 2012],[Sartoretto et al., 2016], on a single type of this habitat [Pergent-Martini et al., 2014], [Sini et al. 2015] and use rapid assessment techniques [Bianchi et al., 2007], [Kipson, 2011], [Deter et al., 2012], [Gatti et al., 2015], depending on prevailing environmental conditions.

Coralligenous habitats have been systematically studied at a larger scale within the CIGESMED ERANET'S program (*Coralligenous based Indicators to evaluate and monitor the "Good Environmental Status" of the MEDiterranean coastal waters*). The main CIGESMED's goal was to understand the connections between pressures (natural or anthropogenic) and the ecosystem functioning in order to define and maintain the Good Environmental Status (GES) of the Mediterranean Sea, by studying the typical, complex and poorly known habitats built by calcareous encrusting algae: the coralligenous habitats. This program is in support of the implementation of the Directive 2008/56/EC of the European Parliament and of the Council of the 17th June 2008. It

participates establishing a framework for stakeholders community action in the field of marine environmental policy (Marine Strategy Framework Directive - MSFD), and highlighting descriptors 1 (biological diversity), 2 (non-indigenous species) and 6 (seafloor integrity). The Marine Strategy Framework Directive (MSFD) is directing European Member States towards an implementation of the assessment of marine environmental status. Due to their very high specific richness, including commercial species, and the number of aesthetically important seascapes they hold, coralligenous habitats are one of the most popular marine environments [Ballesteros, 2006]. The community of CIGESMED redefined it as : "*reefs in dim-light conditions mainly bio-constructed on hard substratum by calcifying coralline algae widespread throughout the Mediterranean sea, including patchwork of habitats complicated by the action of bio-eroders. These complex biogenic formations provide a number of different conditions of light, food and shelter. They are often considered as biodiversity hotspots gathering numerous sessile and sedentary species such as sponges, bryozoans, corals and gorgonians depending on the region and on the depth, to which hundreds of sciaphilic species are associated. These complex environments are a reservoir of natural resources (fisheries, red coral) and form highly valued landscapes sought by divers*". The data provided by CIGESMED are now used by the IndexMed consortium as a model, for developing data mining and decision support.

1.3 IndexMed, an open consortium

IndexMed is a new consortium in charge of indexing Mediterranean biodiversity data, building and analyzing graphs from heterogeneous databases. It aims to develop new ways for data mining in ecology. This consortium aims to identify and overcome the scientific barriers encountered when working on the quality and heterogeneity of data. The use of emerging data mining methods like graph-based models and analyses allows us to address these issues for improving decision-making support. These methods enable us to detect new patterns of contexts factors, invisible when using multidimensional analyses, that have an accurate capacity to indicate particular situations [Klemes, 2015].

2 METHODS

2.1 The challenge of quality data management to enhance results of data mining

Besides theoretical scientific issues (such as the intrinsic heterogeneity and complexity of biodiversity data, from genes to ecosystems, and their links to environmental parameters), the improvement of data quality is hindered by data management issues, such as: i) the dynamic update of voluminous datasets, ii) the update of reference repositories and standards supporting data management, iii) the heterogeneity of data producers and their motivation to maintain and supply their information systems, and iv) the diversity of the targeted end-users and their skills.

An integrated approach of the complexity of coralligenous was implemented to mutualize dataset production methods and visualize large data collections, and extract knowledge to study ecosystems. Health quality Indicators, targeting different levels of biodiversity (from communities to genomes), were co-constructed and tested by scientists, stakeholders, and by a citizen science network. Within the CIGESMED program, an upgrade of the design of each protocol and the inter-calibration exercises between various observers, materials, methods and organizations allowed to obtain: i) an assessment of the data variability due to natural or anthropogenic conditions and ii) a comparison of the different methods or observers and their efficiency. It showed that, under the above consensual definition of coralligenous, coralligenous habitats are made of a large panel of different species assemblages. For instance, coralligenous habitats from the Eastern and Western Mediterranean basins may share only 2 or 3 conspicuous species. Comparisons between regions are complicated by this lack of common species and the environmental conditions that can deeply change. Other links between typical environments and species (e.g. traits, contexts, structures, etc.) should be used to build indicators and to compare the environmental status between regions at a Mediterranean scale. Only multi-criteria contextualization by common factor value level allows constructing and adapting the indicators at a local scale and highlighting the significance of this indicator. Finally, it was decided to use competencies and tools developed by the IndexMed consortium to analyse all heterogeneous data, and integrate multidisciplinary data related to coralligenous habitats within the same multi-criterial approach but considering them at a comparable level of importance.

Coralligenous habitats are under anthropogenic uses and threats [Ballesteros, 2006]. Elaborating and testing "co-interpretation methods" of analyses (i.e. in the same time, at a same level of integration) is thought to be a keystone to mix in the same studies these heterogeneous data from different socio-economics disciplines and biodiversity disciplines, from the genome to the seascapes level. The ultimate goal is to propose scenarios to reach the sustainable management of biodiversity balancing exploitation and conservation. Data mining methods will be able to bring new perspectives to the disciplinary researches that finally examine interrelated objects (e.g. environmental chemistry, genomics, transcriptomics, metabolomics, population ecology / landscape, socio-ecological systems).

2.2 Workflow and e-services

To be able to use different and distributed datasets for data mining, a prototype of "object resolution service" (i.e. a web service that finds links and dependencies among indexed objects, based on unique objects identification (Figure 1)) that can be replicated by stakeholders is shared on a nodal point.
The aims of this prototype is i) listing available data and data stream, ii) analyzing content of datasets and data streams with standards referential, iii) qualifying streams, datasets with unique identifiers if there is no identifiers, iv) suggesting matches between fields to users /matches between equivalent data rows. The role of this object resolution service is to establish links between data row with different "unique identifiers" (e.g. different versions of data raw, interdependencies between raw data and transformed data, etc.).

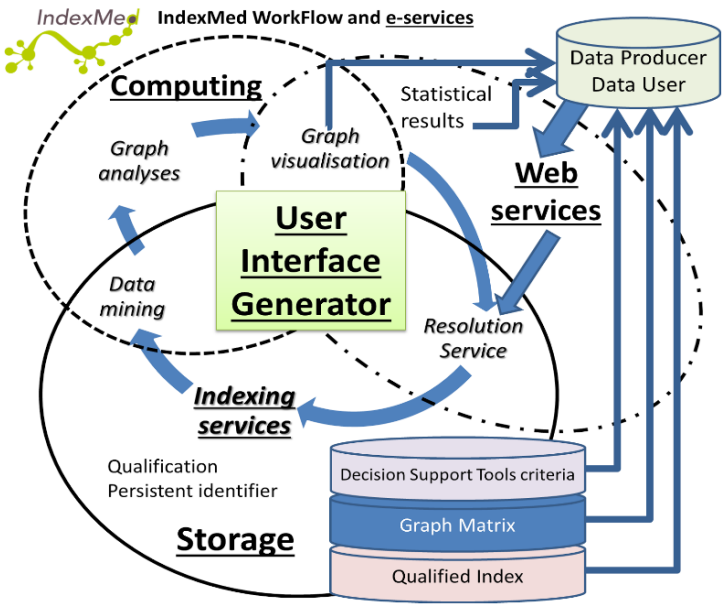


Figure 1 - IndexMed WorkFlow and e-services: the resolution service is able to compare the index with storage data in e-infrastructures and other distant XML, JSON Flux from different databases. When necessary, it creates a persistent identifier or link datasets or data records with existing identifiers if they are enough robust. A qualification process is allowed by a scientist interface, adapted to the level and needs of each user. The indexing service allows data for computing services like data mining and graph analyses, and statistical results and graph models are stored and proposed as a new persistent flux. This system is intended to be replicable as a free software and a free service from European grids (EGI and others).

When it is possible, data qualification uses tools, standards and recommendations at both national (SINP [National Information System on Biodiversity], RBDD [Network of Research Databases]) and international levels (MedOBIS [Mediterranean Ocean Biogeographic Information System], OBIS, GBIF [Cryer et al., 2009], Life-Watch, GEO-BON, etc.) along with other research entities (i.e. IRD [Institute of Research for the Development] or MNHN [National Museum of Natural History, Paris]).
Heterogeneity in datasets may be the result of a lack of standards to name and describe data [Kattge et al., 2011; [Madin et al., 2008]. Thus, attention must be paid to the characterisation of concepts by using controlled vocabulary and semantic links between these concepts, which implies building a thesaurus in the first place (a thesaurus appears more appropriate than an ontology because of its flexibility). Several eco-informatic initiatives attempted to build such thesaurus (see [Michener & Jones, 2012; Laporte, 2012]) and it is expected to take them in account.

New data qualifications generated by IndexMed prototypes aim at following the "guidelines on Data Management in Horizon 2020" (V2.1, 15 Feb. 2016) and cover as it is recommended the handling of research data during and after the project, what data are collected, processed or generated, what methodology and standards are applied, whether data will be shared /made open access and finally

how data will be curated and preserved (Horizon 2020 Annotated Grant Agreement for articles 29.2 and 29.3, IP/12/790 on open access in Horizon 2020. <http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm>)

3 RESULTS

The first tool developed commonly by the CIGESMED and the IndexMed community for scientists working on biodiversity is a prototype building dynamic maps of data and their possible links based on Graphs [David et al., 2015]. It can be used to establish links between objects of different disciplines and is able to connect data without centralizing them. The first aim is to teach the community how to use the graph approach, featuring a didactic and ergonomic interface (Figure 2) with the aims to improve by step user level. It allows evaluating the data quality level and identifying the best ways to improve their efficiency (e.g. density, sensibility, velocity, accuracy, etc.). A tool permits to keep the new models designed by users (i.e. link and node selections) and new items (more than 1 object in a node) and produces a single stream usable by data centers at different formats (NoSQL exports in RDF, Json, XML formats) with a persistent URL.

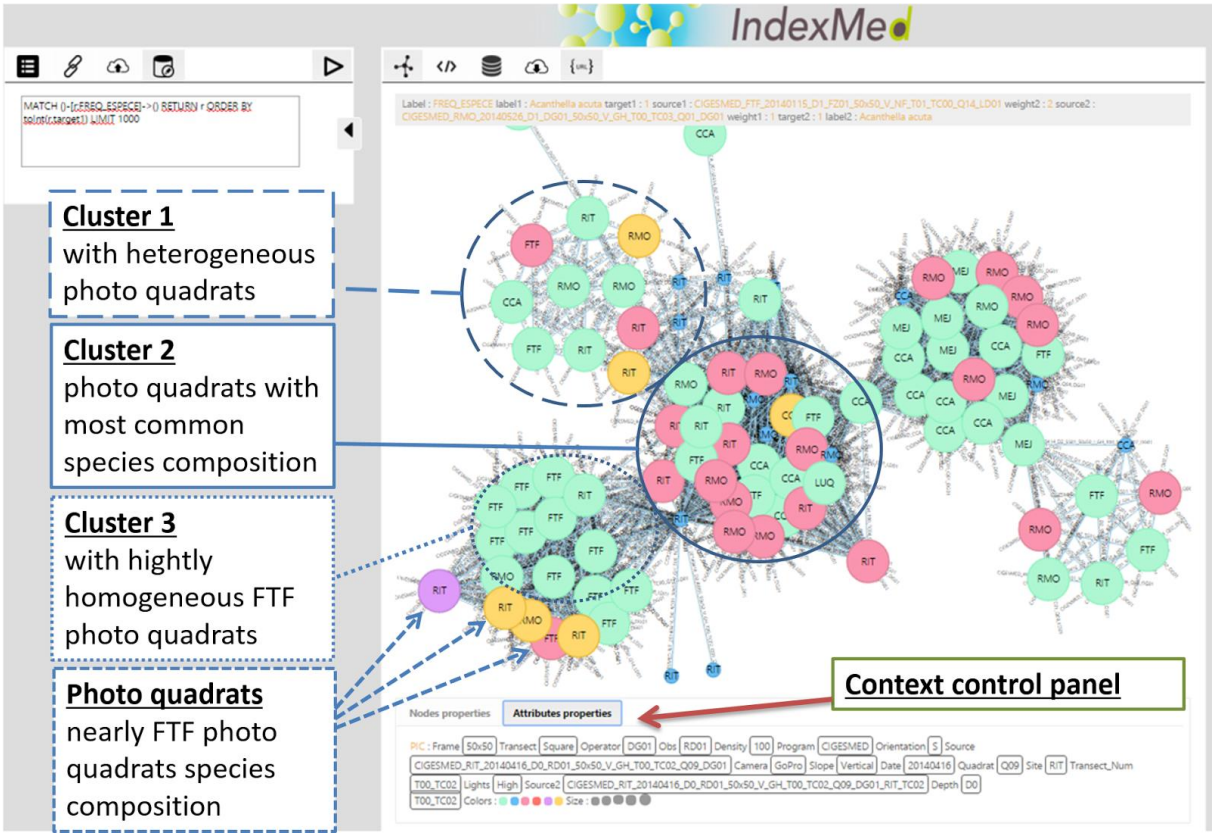


Figure 2 : In this use case of the prototype, photo quadrats selected by the interface are the nodes of the graph, species frequency selected by the user build 1000 links between photo quadrats, the colours of nodes highlight different elements of contexts (here, different observers). Node legends are the names of the observed sites. The more photo quadrats contain similar species assemblages, the more they are attracted. We can observe that some photo quadrats are equivalent of many different sites, and very ubiquitous (cluster 2 in the centre of the graph). In cluster 3 photo quadrats are homogeneous and typical for a site, and near this cluster some photo quadrats of other sites constitute a particular group. Cluster 1 show another type of photo quadrats, less present in each site but represented everywhere.

In the example of figure 2, datasets come from 3 different protocols and data production systems, including one based on photo quadrats analyses with the software photoQuad [Trigonis and Sini, 2012], a cartography of ecological/physical contexts and genetic data. Data objects represented on the graph are photos coming from different sites. Objects can be selected using the context control

panel. Clusters visible in figure 2 can be modified by selection of context or species choice for the links. The relative importance of each species can be modified in the links (e.g. depending of the status or a specific trait) and some context elements can be selected to participate to links between nodes. Observers, highlighted by colours of nodes are not evenly distributed. Experience of observers is reflected in the size of nodes.

This interface uses indexed data, data qualification, and data traceability for discovering patterns in the conjunctions of data values with scientific significance.
The graph design can be manipulated on a web browser interface, the request and manipulations steps can be stored on a personal account and the result allows installing a flux at XML or JSON format available on a web service for data mining or another indexing service..

4 DISCUSSION

4.1 organizing data means to organize access and to improve quality

The description of data quality is an objective of the IndexMed consortium <http://www.indexmed.eu> that can be useful for data about coralligenous, based on an analysis of both similarities and differences between databases. Descriptions as metadata form a set of criteria used for data mining. The graph-based model is an abstraction tool that enables the comparison of various databases despite their differences and that improves decision support using emerging data mining methods. Practically, it is intended to give the equivalence of data, based on data dictionaries, thesauri and ontologies. From the established logical relationships, new qualifiers can be deduced including across data heterogeneity.

This work on CIGESMED data quality and their equivalence with other observatory systems involves first the analysis and description of the common elements of each piece of information, and of what differentiates them (fields name, formats, update rate, precision, observers or sensors, etc.). These descriptions are added to the data, and form a supplementary set of criteria used for data mining.
Standard formats and protocols are used to interconnect CIGESMED data with other databases. Standardization makes possible such a work, as well as a special task on interoperability qualities and accessibility of non-centralized data. It uses aggregation and new visualizations for public display, multi-interface, multi-use and multi-format, and must allow (i) the connection between many databases, and (ii) the preparation of inter-calibration works.

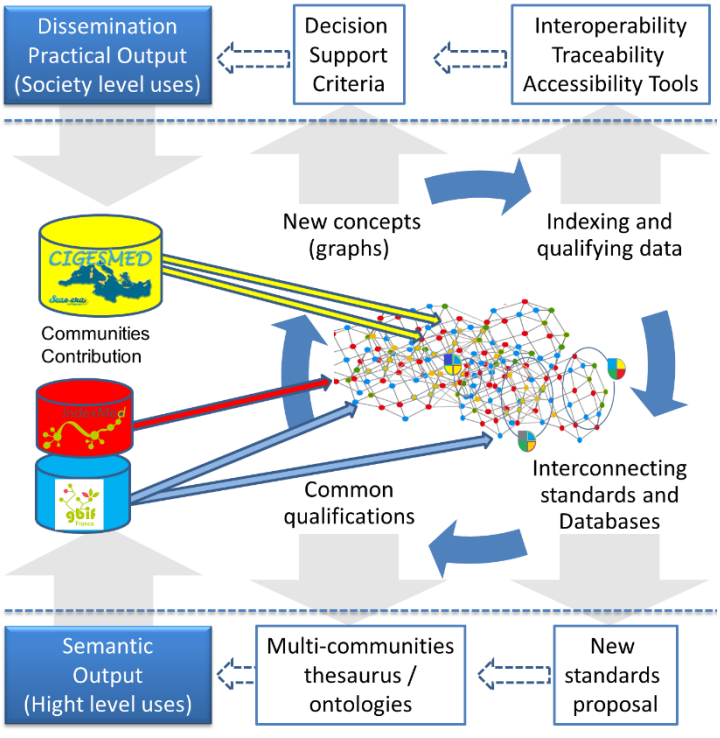


Figure 3: Iterative quality approach and IndexMed Output : It is intended to give the equivalence of data, based on data dictionaries and thesaurus. Some database equivalences allow deducing others, using first specific standards if it exists in each domain and secondary multidisciplinary approach. Equivalencies are used to link heterogeneous objects and construct graphs, where objects are nodes and attribute modalities are links. The main output is a new model of dataset, stored in a graph database (graph matrix) and accessible with web-services for visualization and integrated flux. A second output is the improving of multi-community thesaurus necessary to build new common ecological concepts. The next step of this project is the recognition of patterns of context in the graph matrix that will constitute decision support criteria.

Semantic approaches greatly increase their interoperability and some initiatives using Semantic Web technologies for retrieving Biodiversity data [Amanqui et al., 2014] and developing methods for linking of biodiversity and environmental data already exist, but they often concern only an "inventory" aspect. However, it remains that specific scientific objectives, organizational logic of projects and collection of information are leading to a decentralized data distribution which may hamper environmental research development. In such a heterogeneous system footed on different organizations and data formats, not everything can be homogenized. IndexMed workflow permits to implement an iterative quality demarche (Figure 3) increasing step by step the capacity of each data to be connected to others (i.e. contextual data like biotic, abiotic, anthropogenic and natural pressures, etc.)

The resulting cluster and their correlations to context patterns can be compared with other kinds of analyses: supervised clustering results, statistical ecology approach [Gimenez et al., 2014] and collaborative clustering methods [Forestier et al., 2008] are planned to be used at each part (e.g. a group of nearby nodes with similar patterns of context) of the graph, using job middleware (DIRAC3, [Tsaregorodtsev, 2009]). Another issue is to use "unsupervised" mode, raising the possibility to compare the results of different algorithms to achieve consensus, which acts / results in the most likely scenario. The data mining helps finding managerial values of qualifiers to propose scenarios, and provides new standardized descriptors essential for approaches such as machine learning.

4.2 Efficiency for data mining approach and links with decision support

The chain between data and decision making can be superimposed to the DIKW(U) hierarchy: Data, Information, Knowledge, Wisdom, Understanding [Zeleny, 1987; Ackoff, 1989], replacing Wisdom by Management. In a simplistic view, scientists produce knowledge by analysing data into information and by elaborating theories from information. Data constitute the primary material from which hypotheses are 1) elaborated, and 2) tested. However, even if biodiversity data have been produced in the common framework of the theory of evolution, it has often been done independently in different domains from genes to ecosystems; moreover, biodiversity data are historical in essence, they have a time component: that a species has been observed at a given location at a given time is not reproducible like for physico-chemical experiments. Consequently, every piece of data in each domain may be of importance, and for older ones, they may need to be re-expressed to fit under their current conceptual and standard forms, in particular to use them all in a common approach like here. We are thus dealing with millions of pages of scientific literature and the increasing number of data repositories since Aristotle [Voultsiadou, 2007]. The Biodiversity Heritage Library <http://www.biodiversitylibrary.org/> is already making available almost 50 million pages (and still increasing), mainly up to 1930s because of copyright issues: since the scientific production progresses exponentially, we may talk here about billions of pages. Even narratives of travels and expeditions can be used to extract biodiversity semi-quantitative data [Al-Abdulrazzak et al., 2012].

The development of new data mining tools becomes crucial to explore automatically all sources of biodiversity data, or currently more reasonably semi-automatically, in order to produce the most complete knowledge that constitutes the decision support material (but not the decision-making tools by themselves!). This knowledge is the basis for developing alternative future scenarios about the biodiversity management among which decision and policy-makers will make a political choice. The graph approach may allow going a step further by integrating socio-economic knowledge in these scientifically supported scenarios. Currently, this integration is made at the decision making level, where biodiversity and socio-economics scenarios are on the contrary put in competition, most often to the benefit of the socio-economics scenarios, with too many examples from the domain of fisheries [Froese, 2011].

5 CONCLUSIONS AND RECOMMENDATIONS

Compared with dimensional and multidimensional analyses, which are often used for ecological and environmental purposes, such innovative approaches make possible the investigation of complex research questions and the emergence of new scientific hypotheses. Regarding the first results, environmental scientists and environmental managers from CIGESMED and from the IndexMed consortium have to face different challenges about links between data and well understanding of the meaning of new objects and their variation. For better analysing heterogeneously distributed data spread in different databases and for identifying statistical relationships between observed data and

the emergence of contextual patterns, it will be necessary to create matches and incorporate some approximations.
The area of Decision Support Systems (DSS) focuses on development of interactive software that can use data mining export of IndexMed prototypes. This prototype aims to provide qualifiers that can be interpreted in patterns as answers to relevant decisional questions from the users, thus enhancing a person or a group to make better decisions. Till now, important efforts to develop links between Indexmed and dedicated DSS are required and possible for every particular application [Varanon et al., 2007, Power, 2007]. Specific DSS linked to IndexMed must be experimented where some successful experiences appear in several fields, like self-care management [Marschollek, 2012], water management [Pallottino, 2005], forest ecosystems [Nute, 2004] or air pollution [Oprea, 2005]. IndexMed community is open to contribution. IndexMed software are open source and privileged case studies are open data, and the involved teams plan to set up a forge and a contributory platform for expanding testing graph approaches.

Multidisciplinary approaches are a key as well as the most difficult way to improve data mining and DSS. At a “human” level, it is seriously necessary to encourage the data openness and data sharing, as the only way to give value to data after their primary use [McNutt et al., 2016]. A good start might be to organize more events dedicated to the sharing of experience and expertise, the acquisition of practical methods to construct graphs and value data through “metadata and data papers".

ACKNOWLEDGMENTS

The construction of the first prototype for IndexMed consortium is funded by the CNRS défi “VIGI-GEEK (VIsualisation of Graph In transdisciplinary Global Ecology, Economy and Sociology data-Kernel)” and CNRS INEE through the “CHARLIEE” project. Data used for this article were obtained through the CIGESMED project <www.cigesmed.eu> (ANR conventions n° 12-SEAS-0001-01, 02 and 03 for France; GSRT - 12SEAS-12-C2 for Greece; TUBITAK Project No: 112Y393 for Turkey). The authors acknowledge the support of France Grilles for providing computing resources on the French National Grid Infrastructure. Supplementary acknowledgement to Gergely SIPOS, Jan BOT and Roberta PISCITELLI for the helpful support provided at “design your e-infrastructure” EGI <www.egi.eu> workshop. We acknowledge all the field helpers and students who have participated in data collection in the field and in the lab, as well as in data management.

REFERENCES

Ackoff, R. L., 1989. From data to wisdom. Journal of applied systems analysis, 16(1), 3-9.

Al-Abdulrazzak, D., Naidoo, R., Palomares, M. L. D., Pauly, D., 2012. Gaining perspective on what we've lost: the reliability of encoded anecdotes in historical ecology. PloS one, 7(8), e43386. doi:10.1371/journal.pone.0043386.

Amanqui, F.K., Serique, K.J., Cardoso, S.D., dos Santos, J.L., Albuquerque, A. and Moreira, D.A., 2014. Improving biodiversity data retrieval through semantic search and ontologies. Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Conference on Web Intelligence, 11-14 August 2014, Warsaw, Poland, vol.1 (WI), pp.274-281, doi: 10.1109/WI-IAT.2014.44.

Ballesteros, E., 2006. Mediterranean coralligenous assemblages: a synthesis of present knowledge. Oceanography and Marine Biology: An Annual Review, 44, 123-195.

Bianchi, C. N., Cattaneo-Vietti, R., Morri, C., Navone, A., Panzalis, P., Orru, P., 2007. Coralligenous formations in the Marine Protected Area of Tavolara Punta Coda Cavallo(N-E Sardinia, Italy). Biologia marina mediterranea, 14(2), 148-149.

Conruyt, N., Sébastien, D., Vignes-Lebbe, R., Cosadia, S., 2010. Moving from biodiversity information systems to biodiversity information services. Information and Communication Technologies for Biodiversity Conservation and Agriculture, 107-128.

Cryer, P., Hyam R., Miller C., Nicolson, N., Tuama, É.Ó, Page, R., Rees, J., Riccardi, G., Richards, K., Whitev, R., 2009. Adoption of persistent identifiers for biodiversity informatics. Report published by GBIF Secretariat.

Cortez, P., & Embrechts, M. J., 2013. Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225, 1-17.

David, R., Féral, J. P., Blanpain, C., Diaconu, C., Dias, A., Gachet,S., Gibert, K., Lecubin, J., Surace, C., 2015. A First Prototype for Indexing, Visualizing and Mining Heterogeneous Data in

Mediterranean Ecology: Within the IndexMed Consortium Interdisciplinary Framework. In 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS).232-239.

Deter, J., Descamp, P., Ballesta, L., Boissery, P., & Holon, F., 2012. A preliminary study toward an index based on coralligenous assemblages for the ecological status assessment of Mediterranean French coastal waters. Ecological indicators, 20, 345-352.

Deter, J., Descamp, P., Boissery, P., Ballesta, L., & Holon, F., 2012. A rapid photographic method detects depth gradient in coralligenous assemblages. Journal of Experimental Marine Biology and Ecology, 418, 75-82.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P., 1996. From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Féral, J.-P., David, R., 2013. L'environnement, un système global dynamique. Zone côtière et développement durable, une équation à résoudre. In: Euzen, A., Eymard, L., Gaill, F. (Eds), 2013. Le développement durable à découvert, CNRS éditions: Paris, September,96-97, ISBN : 978-2-271-07896-4.

Féral, J.-P., Arvanitidis, C., Chenuil, A., Çinar, M.E., David, R., Frémaux, A., Koutsoubas, D., Sartoretto, S., 2014. CIGESMED, Coralligenous based Indicators to evaluate and monitor the « Good Environmental Status » of the MEDiterranean coastal waters, a SeasEra project (www.cigesmed.eu). Proceedings RAC/SPA 2nd Mediterranean Symposium on the Conservation of coralligenous and other calcareous bio-concretions, Portorož, Slovenia, October, 15-21.

Forestier, G., Wemmert, C., Gançarski, P., 2008. Multisource images analysis using collaborative clustering. EURASIP Journal on Advances in Signal Processing, 2008, 133.

Froese, R., 2011. Fishery reform slips through the net. Nature, 475(7354), 7-7.

Gachet, S., Véla, E., Tatoni, T., 2005. BASECO: a floristic and ecological database of Mediterranean French flora. Biodiversity & Conservation, 14(4), 1023-1034.

Gatti, G., Bianchi, C. N., Morri, C., Montefalcone, M., Sartoretto, S., 2015. Coralligenous reefs state along anthropized coasts: Application and validation of the COARSE index, based on a rapid visual assessment (RVA) approach. Ecological Indicators, 52, 567-576.

Gibert, K., Valls, A., Batet, M., 2014. Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering. Knowledge and Information Systems, 40(3), 559-593.

Gibert, K., Conti, D., Vrecko, D., 2012. Assisting the end-user in the interpretation of profiles for decision support. an application to wastewater treatment plants. Environmental Engineering and Management Journal, 11(5), 931-944.

Gimenez, O., Buckland, S.T., Morgan, B.J.T., Bez, N., Bertrand, S., Choquet, R., Dray, S., Etienne, M.P., Fewster, R., Gosselin, F., Mérigot, B., Monestiez, P., Morales, J., Mortier, F., Munoz, F., Ovaskainen, O., Pavoine, S., Pradel, R., Schurr, F.M., Thomas, L., Thuiller, W., Trenkel, V., de Valpine, P., Rexstad, E., 2014. Statistical ecology comes of age. Biology letters, 10(12), 20140698.

Hong,J. S., 1982. Contribution to the study of populations of the coralligenous concretionary bottom from the Marseille region on the northwestern Mediterranean coast. Bull. Korea Ocean Res. Dev. Inst., 4: 1–2.

Kattge, J., Ogle, K., Bönisch, G., Díaz, S., Lavorel, S., Madin, J., Nadrowski, K., Nöllert, S., Sartor, K., Wirth, C., 2011. A generic structure for plant trait databases. Methods in Ecology and Evolution, 2(2), 202-213.

Kipson, S., Fourt, M., Teixidó, N., Cebrian, E., Casas, E., Ballesteros, E., Zabala, M., Garrabou, J., 2011. Rapid biodiversity assessment and monitoring method for highly diverse benthic communities: a case study of Mediterranean coralligenous outcrops. PLoS One, 6(11), e27103.

Kissling, W. D., Hardisty, A., García, E. A., Santamaria, M., De Leo, F., Pesole, G., Freyhof, J., Wissel, S., Konijn, J., Los, W., 2015. Towards global interoperability for supporting biodiversity research on essential biodiversity variables (EBVs). Biodiversity, 16(2-3), 99-107.

Klemeš, J.J. (Ed.), 2015. Assessing and measuring environmental impact and sustainability. Butterworth-Heinemann. 608 pp. ISBN: 9780127999685

Laborel, J., 1961. Le concrétionnement algal "coralligène" et son importance géomorphologique en Méditerranée. Recueil des travaux de la Station marine d'Endoume, 23(37), 37-60.

Laporte, M. A., Garnier, E., 2012. ThesauForm—Traits: A web based collaborative tool to develop a thesaurus for plant functional diversity research. Ecological Informatics, 11, 34-44.

Laporte, M. A., Mougenot, I., Garnier, E., Stahl, U., Maicher, L., Kattge, J., 2014. A semantic web faceted search system for facilitating building of biodiversity and ecosystems services. In Data Integration in the Life Sciences (pp. 50-57). Springer International Publishing.

Laubier, L., 1966. Le Coralligène des Albères. Monographie biocénotique. Adaptations chez les Annélides Polychètes interstitielles. Thès. Fac. Sci. Univ. Paris (Sér. A, No. 4693 No d'ordre 5541), 139-316.

Madin, J. S., Bowers, S., Schildhauer, M. P., Jones, M. B., 2008. Advancing ecological research with ontologies. Trends in Ecology & Evolution, 23(3), 159-168.

Marschollek, M., 2012. Decision support at home (DS@ HOME)–system architectures and requirements. BMC medical informatics and decision making, 12(1), 43.

McNutt, M., Lehnert, K., Hanson, B., Nosek, B. A., Ellison, A. M., King, J. L., 2016. Liberating field science samples and data. Science, 351(6277), 1024-1026.

Michener, W. K., Jones, M. B., 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends in ecology & evolution, 27(2), 85-93.

Noss, R. F., 1990. Indicators for monitoring biodiversity: a hierarchical approach. Conservation biology, 4(4), 355-364.

Nute, D., Potter, W. D., Maier, F., Wang, J., Twery, M., Rauscher, H. M., Knopp, P., Thomasma, S., Dass, M., Uchiyama, H., Glende, A., 2004. NED-2: an agent-based decision support system for forest ecosystem management. Environmental Modelling & Software, 19(9), 831-843.

Oprea, M., 2005. A case study of knowledge modelling in an air pollution control decision support system. Ai Communications, 18(4), 293-303.

Pallottino,S., Sechi, G.M., Zuddas, P., 2005. A DSS for water resources management under uncertainty by scenario analysis. Environmental Modelling & Software, 20 (8): 1031-1042, doi: 10.1016/j.envsoft.2004.09.012.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M., 2013. Essential biodiversity variables. Science, 339 (6117): 277-278. doi: 10.1126/science.1229931

Pergent-Martini, C., Alami, S., Bonacorsi, M., Clabaut, P., Daniel, B., Ruitton, S., Sartoretto, S., Pergent, G., 2014. New data concerning the coralligenous atolls of cap corse: an attempt to shed light on their origin. RAC/SPA 2nd Mediterranean Symp. on the Conservation of coralligenous and other calcareous bio-concretions, Portorož (Slovenia), 29-30/10/2014, 129-134.

Peters, D.P., Havstad, K. M., Cushing, J., Tweedie, C., Fuentes, O., Villanueva-Rosales, N., 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere, 5(6), 1-15.

Power,D.J., 2007. A Brief History of Decision Support Systems, DSSResources.COM (Editor), World Wide Web, version 4.0", March 2007, http://dssresources.com/history/ dsshistory.html.

Reichman, O. J., Jones, M. B., Schildhauer, M. P., 2011. Challenges and opportunities of open data in ecology. Science, 331(6018).

Sartoretto, S., 1994. Structure et dynamique d'un nouveau type de bioconstruction à Mesophyllum lichenoides (Ellis) Lemoine (Corallinales, Rhodophyta). Comptes rendus de l'Académie des sciences. Série 3, Sciences de la vie, 317(2), 156-160.

Sartoretto,S., Schohn, T., Bianchi, C.N., Morri, M.C., Garrabou, J., Ballesteros, E., Ruitton, S., Verlaque, M., Daniel, B., Charbonnel, E., Blouet, S., David, R., Féral, J.-P., Gatti, G., 2016. An integrated approach to evaluate and monitor the conservation state of coralligenous habitats: the Index-Cor approach. submitted in Ecological Indicators.

Sini, M., Kipson, S., Linares, C., Koutsoubas, D., Garrabou, J., 2015. The Yellow Gorgonian Eunicella cavolini: demography and disturbance levels across the Mediterranean Sea. PloS one, 10(5), e0126253.

Tsaregorodtsev, A., 2009. DIRAC3 . The New Generation of the LHCb Grid Software. Journal of Physics: Conference Series, vol. 219 062029, n° 6.

Trygonis, V., Sini, M., 2012. photoQuad: A dedicated seabed image processing software, and a comparative error analysis of four photoquadrat methods. Journal of Experimental Marine Biology and Ecology, 424, 99-108.

Uraikul, V., Chan, C. W., Tontiwachwuthikul, P., 2007. Artificial intelligence for monitoring and supervisory control of process systems. Engineering Applications of Artificial Intelligence, 20(2), 115-131.

Voultsiadou, E., 2007. Sponges: An historical survey of their knowledge in Greek antiquity. Journal of the Marine Biological Association of the United Kingdom, 87(06), 1757-1763.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D., 2012. Darwin Core: An evolving community-developed biodiversity data standard. PloS one, 7(1), e29715.

Zeleny, M., 1987. Management support systems: towards integrated knowledge management. Human systems management, 7 (1): 59-70.

Visualisation de données sous forme de graphes en archéologie. Rencontre opérationnelle des archéologues d'ArkeoGIS et des écologues d'IndexMed

Data visualisation in archaeology based on graph approach. Operational meeting of ArkeoGIS archaeologists and IndexMed ecologists

Romain David[1], Loup Bernard[2], Cyrille Blanpain[3], Alrick Dias[1], Jean-Pierre Féral[1], Sophie Gachet[1], Julien Lecubin[3], Christian Surace[4], Thierry Tatoni[1]

[1] Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE), CNRS, Aix Marseille Université, IRD, et Université d'Avignon, Station Marine d'Endoume, romain.david@imbe.fr, alrick.dias@imbe.fr, jean-pierre.feral@imbe.fr, sophie.gachet@imbe.fr, thierry.tatoni@imbe.fr
[2] Université de Strasbourg, Université de Haute-Alsace, CNRS, Archimède UMR 7044, Strasbourg, loup.bernard@unistra.fr
[3] Service informatique (SIP), OSU Pythéas, CNRS, Aix Marseille Université, Marseille, cyrille.blanpain@osupytheas.fr, julien.lecubin@osupytheas.fr
[4] Laboratoire d'Astrophysique de Marseille (LAM), CNRS, Aix Marseille Université, Marseille, christian.surace@lam.fr

RÉSUMÉ. Un point commun des études en archéologie, en écologie ou sur les systèmes sociaux est que la production de données est à la fois coûteuse et peu automatisée. Les suivis de longues séries temporelles et/ou à larges emprises spatiales sont difficiles à mener, dès lors qu'il faut recourir sur une longue durée à plusieurs observateurs. La robustesse et la reproductibilité de l'observation sont aussi plus difficiles à obtenir, voire impossibles en archéologie, même si les méthodes de modélisation se développent.

Dans un cadre de production de données multi-sources, l'équivalence des systèmes d'observations et l'inter-calibration d'observateurs deviennent cruciales. Des approches intégratives, pluri- ou trans- disciplinaires, deviennent nécessaires à l'étude de systèmes où la production de données dans chaque discipline est discontinue, plus ou moins précise et mal répartie. Pourtant, toutes les variables (caractérisation des activités économiques, des installations humaines, études des productions, objets reconstitués ou découverts, données biotiques et abiotiques, cartographies des pressions anthropiques et naturelles, services rendus et ressentis, image sociétale...) de ces systèmes interagissent dans le temps et à chaque échelle spatiale.

Après quelques années d'existence, ArkeoGIS agrège aujourd'hui 67 bases de données représentant plus de 50 000 objets (sites, analyses...). Fort de cette normalisation de l'information archéologique et paléo-environnementale, il nous a semblé important de tester de nouvelles méthodes de fouille de données, afin de mettre en évidence de possibles données « connexes » et complexes possiblement reliables à ces jeux de données. Le lien entre les extraits des bases agrégées au sein d'ArkeoGIS nous a permis de tester ces approches grâce à un prototype "open source" développé par le consortium IndexMed. Ce prototype permet la mise en place de liens entre objets de bases de données différentes.

Le consortium IndexMed a pour objectif d'identifier puis de lever les verrous scientifiques liés à la qualité des données et à leur hétérogénéité. La représentation de l'information sous forme de graphe rend possible la prise en compte des données malgré leur disparité et sans les hiérarchiser, et permet d'améliorer la précision des outils d'aide à la décision utilisant des méthodes émergentes d'analyse de données (*clustering* collaboratif, classification collective, fouille de graphes, analyse de réseau, extraction de communautés). Adapter ces méthodes à l'archéologie nous permet d'aller au-delà de la « simple » agrégation de données : ArkeoGIS peut donc aussi servir à alimenter les outils de fouille utilisés au sein de nos données et métadonnées.

ABSTRACT. The one thing in common "archaeological", "biodiversity" or "social systems" studies share is that data production is both expensive and few automated. Long time series and / or large spatial surveys are difficult to conduct, since it is necessary to use several observers. The robustness and reproducibility of the observation are also harder to get and is obviously impossible in archaeological sciences, even if modeling methods are improved.

In a context of multi-source data production, the equivalence of observation systems and the inter-calibration of the observers become crucial. Multi-disciplinary integrative approaches become necessary to study systems where the output

of data, in each discipline, is discontinuous, imprecise and poorly distributed. Yet, all variables (characterization of economic activities and human installation, productions studies, characteristics of the discovered or reconstituted objects, biotic or abiotic data, maps of anthropogenic and natural pressures, rendered services and feelings, societal perception...) of these systems interact over time and at each spatial scale.
After a few years of existence, ArkeoGIS aggregates 67 databases representing over 50 000 objects (sites, analyzes...). With this standardization of archaeological and paleoenvironmental information, it seemed important to test new data mining methods, to see whether "related" and complex data can be linked to these archaeological data sets. The link between aggregated-bases extracts within ArkeoGIS allowed us to set up a cross-requesting and test possibilities in a prototype developed by the consortium IndexMed. This prototype, open source, allows the establishment of links between objects from different databases.
The consortium IndexMed aims to identify and to raise the scientific challenges related to data quality and heterogeneity. The use of graphs allows us to consider data despite their disparity and without prioritization, and improve decision support using emerging data mining methods (collaborative clustering, machine learning, graphs approaches, representation knowledge). Adapting these methods in archeology allows us to go beyond the "simple" data aggregation: ArkeoGIS can therefore also be used to power such tools allowing us to mine our data and metadata.
MOTS-CLÉS. visualisation, qualification de données, graphes, système d'information décentralisé, archéologie.
KEYWORDS. visualisation, data qualification, graph, distributed information system, archeology.

Introduction / contexte

Si les archéologues utilisent l'informatique depuis plusieurs décennies, la mise en commun des données produites et leur interrogation à l'aide de méthodes plus novatrices que les désormais traditionnelles CAH et AFC restent un défi. Ces données hétérogènes sont difficilement exploitables de manière intégrée par les techniques couramment utilisées par les chercheurs en archéologie. L'analyse de cette grande quantité de données diversifiées et produites par des sources hétérogènes constitue également un vrai défi pour la science des données. Désormais, la croissance exponentielle de la quantité de données nécessite l'utilisation des outils les plus récents.

La version 4 d'ArkeoGIS (arkeogis.org) permet d'agréger des bases de données disparates au sein d'un outil libre et en ligne à l'aide d'une ontologie "bottum up" construite par les acteurs de la discipline. Forte de plusieurs décennies d'expériences plus ou moins comparables (Archaeomedes puis ArchaeDYN, Fastionline), et agrégeant des projets aussi bien archéologiques (Digital Atlas of Roman and Medieval Civilizations - DARMC, sont à l'étude Chronocarto, NOMISMA, artefacts…) que des projets environnementaux (European Pollen Database, MedMAX, Banadora ou DCCD en cours), le projet ArkeoGIS a trouvé avec Indexmed une équipe développant un outil très adapté, bien qu'initialement développé dans le cadre d'études en écologie marine (David et al 2016).

Récemment, les 3èmes journées organisées par le consortium IndexMed (https://indexmed2016.sciencesconf.org/) ont mis en évidence non seulement le potentiel des approches basées sur les graphes, mais aussi les lacunes en termes de compétences et d'expérience de la communauté des écologues et des archéologues pour adapter et utiliser ces méthodes, afin d'analyser de manière totalement intégrée leurs jeux de données multisources. Une première discussion autour de la visualisation de données hétérogènes, à laquelle ont participé les animateurs d'ArkeoGIS a notamment permis de montrer que des techniques à base de graphes peuvent être adaptées pour modéliser plus efficacement les composantes d'interactions spatiales malgré l'hétérogénéité de ce type d'information. Les participants à ces journées (STIC -Sciences des Techniques de l'Information et de la communication-, écologues et archéologues) ont sollicité l'organisation de rencontres et le développement de collaborations entre la communauté des chercheurs en science de l'écologie et de la biodiversité et celle des chercheurs en science des données et STIC.

IndexMed est un consortium pluridisciplinaire créé par l'axe *Gestion de la biodiversité et des espaces naturels* de l'IMBE (Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale). Son objectif principal est de développer la culture des bases de données et leur utilisation efficace dans le milieu de la recherche en écologie et biodiversité. Ce consortium s'est étendu à plusieurs UMR de disciplines différentes (notamment, de l'environnement pour l'expertise qualitative

de la donnée, et de l'astronomie pour l'expertise en matière de gestion des grosses masses de données). Il doit permettre de répondre à des appels à projets dans le domaine des bases de données en écologie méditerranéenne en favorisant l'interdisciplinarité et les collaborations avec d'autres entités du CNRS. Les projets qui y seront développés doivent s'appuyer sur les différentes démarches nationales et internationales et promouvoir un travail partenarial international. IndexMed doit notamment servir de relais aux réseaux et démarches en place nationalement et internationalement, et proposer une réponse aux obligations européennes (Aarhus, INSPIRE,…) auxquelles les laboratoires de recherche travaillant dans les domaines de l'environnement et des sciences humaines sont et seront de plus en plus soumis. L'objectif à court terme d'IndexMed est de mettre en place une plateforme d'indexation des données sur la biodiversité méditerranéenne et des paramètres environnementaux ayant un intérêt pour la recherche (David et al 2015). Cette indexation utilisera les outils et méthodes préconisés nationalement (SINP - Système national d'Information sur la Nature et les Paysages, MNHN - Muséum National d'Histoire Naturelle, SPN - Service du Patrimoine Naturel, RBDD Réseau Bases De Données du CNRS) ou internationalement (OBIS, GBIF, LifeWatch, GEOBON, CoL, WoRMS…) et s'appuiera sur les catalogues développés à ce niveau (IDCNP - Inventaire des Dispositifs de Collecte sur la Nature et les Paysages du SINP, Réseaux d'acteurs de la FRB - Fondation pour la Recherche sur la Biodiversité).

Les données environnementales et écologiques accessibles sont d'un grand intérêt pour contextualiser les données issues des prospections archéologiques. Les méthodes de fouille de données basées sur la fouille de graphes peuvent être transposables en archéologie, et s'appuyer sur les résultats de requêtes réalisées à l'aide d'ArkeoGIS. Cet article présente les principes de l'utilisation des graphes envisagés en lien avec ArkeoGIS, ainsi que quelques grands types de questionnements scientifiques testés grâce aux premiers exports. Il met en lumière les verrous scientifiques et techniques identifiés lors de ces premières prospections. Il dresse enfin la liste de quelques pistes pour lever ces barrières et développer ce nouveau mode de prospection en archéologie, basé sur les données hétérogènes et multi-sources.

Théorie et méthode

Principes de la représentation de l'information sous forme de graphes

Un graphe est un ensemble de points que l'on appelle des nœuds (sommets en mathématique ou cellules en informatique) reliés par des traits (segments) ou flèches nommées arêtes (ou bien liens ou arcs). L'ensemble des arêtes (edges en anglais) entre nœuds (nodes en anglais) forme une figure similaire à un réseau (Aggarwal & Wang 2010).

Afin de construire les graphes qui seront présentés, un export de données a été réalisé à partir d'ArkeoGIS. La représentation de ces données sous forme de graphes permet de relier des objets (champs de la base de donnée ou valeurs de ces champs) ayant des formats différents (quantitatif, qualitatifs ordonnés ou non ordonnés); les valeurs d'attributs d'un second champ de la base de donnée permettent de créer les liens entre ces objets. Les liens sont matérialisés par des descripteurs (une variable ayant plus d'une valeur possible, qui est aussi la valeur ou une transposition de la valeur que prend un champs de la base de donnée pour un enregistrement). Les descripteurs quantitatifs sont en général transformés en classes de valeurs.

Plusieurs descripteurs peuvent être assemblés pour former - selon la combinaison de leurs valeurs respectives - un motif, qui pour ces valeurs données sont appelés patrons (patterns en anglais). Ces patrons peuvent décrire des objets et/ou des liens et/ou des contextes (contextes qui ne participent pas à la topologie du graphe).

Les objets ayant le plus de liens en commun sont les plus proches, ceux ayant les liens les plus ténus (c'est à dire le moins de chemins possibles pour les relier à entre eux et beaucoup de nœuds

intermédiaires) sont les plus éloignés dans la représentation. La représentation sous forme de graphe permet de représenter de nouveaux objets en combinant les valeurs de différents champs. On peut ainsi traiter les champs un à un ou bien en groupe de valeurs.

D'autres champs de la base de données, nommés "contextes", sont ensuite utilisés pour colorer ou changer la forme et la grosseur des nœuds. Ils ne participent pas à la topologie du graphe. Les motifs ainsi projetés dans le graphe peuvent être (i) dispersés, auquel cas les liens qui organisent le graphe ne sont pas liés aux éléments de contexte ; ou bien (ii) regroupés dans une ou plusieurs parties du graphe, auquel cas il existe un lien entre la façon dont les nœuds sont organisés et un ou plusieurs contextes.

L'analyse des fréquences relatives de ces "motifs" et des redondances entre "plus proches voisins" par rapport à leur fréquence dans tout ou partie du graphe donne une idée de l'importance de ces corrélations. La significativité de ces motifs peut ensuite être testée par des méthodes comme le *clustering* de graphes. Le clustering (classification non supervisée en français, mais c'est le terme anglais qui est le plus usité à la place de "classification", "groupe", ou "regroupement") consiste à regrouper des éléments. Cette agrégation est un élément-clé pour l'analyse de grands graphes. Une fois les groupes obtenus, on peut ré-appliquer l'opération pour obtenir un clustering hiérarchique (basé sur une autre variable par exemple). Cette décomposition hiérarchique (ou multi-échelle) permet de modifier la complexité des algorithmes de fouille, de faciliter l'exploration des données, et de proposer une visualisation paramétrable : on parle aussi de navigation multi-échelle (Lambert et al 2013).

Dans des graphes plus complexes ou le nombre de combinaison et de liens peut croître exponentiellement, l'étude de la corrélation entre fréquence de contextes et "clusters" de nœuds peut demander de paralléliser les calculs nécessaires à une investigation des parcours possibles. Selon la question scientifique sous-jacente aux objets représentés par le graphe, certains éléments dans les liens ou les nœuds peuvent être ignorés ou simplifiés. Cet aspect prospectif dans les graphes est en cours d'élaboration avec la communauté STIC .

Le potentiel actuel d'ArkeoGIS

ArkeoGIS fonctionne comme un agrégateur requêtable de bases de données. Cela signifie que toute information présente dans ArkeoGIS peut faire l'objet de requêtes sur son emplacement, l'état de la recherche, et les périodes concernées. Chaque site peut ensuite être interrogé à l'aide de cinq filtres selon que l'information concerne les structures, le mobilier, les productions, le paysage et les analyses, ou les textes et l'iconographie. Le résultat de la requête s'affiche sous forme de carte dans l'application ; les lignes d'informations correspondantes peuvent ensuite être exportées dans un format très simple (CSV), permettant une réutilisation dans tout type de logiciel. Ce format assure une compatibilité avec des tableurs, des bases de données, des systèmes d'information géographique, des logiciels d'analyse ou de modélisation (comme R ou Netlogo p.ex.), ou encore avec le prototype de visualisation sous forme de graphe de données distantes en cours de développement dans le cadre d'IndexMed.

Initialement développé afin de mutualiser les données archéologiques et paléoenvironnementales de la vallée du Rhin, ArkeoGIS est un système d'information géographique (SIG) libre, en ligne et multilingue (allemand-anglais-espagnol-français). Actuellement dans sa quatrième version, ArkeoGIS permet de mettre en commun les données scientifiques spatialisées concernant le passé. Les bases de données sont issues de travaux de chercheurs institutionnels, d'étudiants avancés, de sociétés privées, de services d'archéologie, mais aussi de travaux de paléo-environnementalistes, d'historiens et de géographes. Tous ces travaux et bases de données sont accessibles et requêtables en ligne. L'étendue chronologique de l'outil est désormais ouverte et permet d'agréger des informations depuis la Préhistoire jusqu'à nos jours. L'emprise spatiale d'ArkeoGIS permet d'afficher des informations sur toutes les régions du monde. A ce jour, les régions les mieux renseignées sont le Rhin Supérieur, l'Ouest méditerranéen et le Proche Orient. Plusieurs dizaines de milliers de sites, objets et analyses sont d'ores et déjà accessibles. ArkeoGIS fait aussi le lien vers différents outils numériques, permettant à ses utilisateurs d'avoir connaissance de différents projets numériques en ligne.

Chaque utilisateur peut interroger en ligne tout ou partie des bases disponibles, afficher ses résultats sur plusieurs fonds de carte et exporter les résultats de sa requête vers d'autres outils. ArkeoGIS peut servir à tout travail de recherche individuel ou collectif. Il permet entre autres de gérer le *Data Management Plan* (DMP) de contrats de recherche, et constitue un outil puissant pour la préparation de recherches théoriques ou de terrain (fouilles, synthèses, thèses etc..)

Chaque auteur mettant à disposition des informations géoréférencées au format ArkeoGIS reste maître de celles-ci et peut seul décider de les modifier, un identifiant unique pérenne (DOI-HANDLE, répondant aux normes INSPIRE) qualifie chacune des bases. Le contributeur peut ainsi très facilement accéder aux informations des autres contributeurs afin d'implémenter sa base tout en citant ses sources. Un annuaire permet de mettre en contact les chercheurs, afin de développer les échanges entre pays et entre institutions.

Aspect heuristique et représentation de la connaissance

Pour l'archéologue, la spatialisation des données a une forte valeur heuristique. Elle permet de saisir instantanément l'état de la recherche ou l'avancement de la mise en commun des données, et rend possible une approche interdisciplinaire et diachronique.

Cette représentation spatiale de la connaissance est immédiatement exploitable pour les données que le chercheur maîtrise ; en revanche, pour les données issues de disciplines connexes (dans le temps, l'espace ou en provenance de travaux environnementaux ou paléo-environnementaux), d'autres outils sont nécessaires. L'exploration des jeux de données peut avantageusement compléter la représentation géographique et permet de mettre en évidence des ressemblances entre sites dans le jeu de données, sans que ceux-ci soient géographiquement proches (ce qui fait l'objet de la présente collaboration).

Enfin, les métadonnées et la façon dont nos bases sont renseignées livrent, grâce à ces représentations de la connaissance, des informations parfois inattendues sur la densité de la recherche par thématique, et se révèlent un outil précieux pour les questions de "*state of art*".

Curation de données

La mise en commun de données issues de différents chercheurs permet une curation simple, sur l'emplacement des sites par exemple, mais aussi une amélioration itérative des données. C'est-à-dire que chaque chercheur peut compléter sa base avec les informations mises à disposition par les collègues, et au passage corriger d'éventuelles erreurs. Ce travail commun implique de facto la constitution d'un vocabulaire contrôlé commun à large échelle. Lorsqu'il est validé par une communauté homogène et organisé avec des liens d'équivalence et des liens hiérarchiques, on parle alors de micro-thésaurus. Lorsque deux communautés travaillent à l'interprétation interdisciplinaire de leurs données, comme c'est ici l'ambition (archéologie / palynologie / paramètres environnementaux), une confrontation entre ces micro-thésaurus est nécessaire. Celle-ci doit prioritairement établir des correspondances sur les descripteurs et la valeur des descripteurs qui, utilisés en commun, permettent une analyse conjointe des données de disciplines différentes (cf. aussi Bernard et al 2015 pour une approche plus "traditionnelle").

Export des données pour construire les graphes

Afin de pouvoir exploiter les données exportées depuis ArkeoGIS, un certain nombre de modifications ont dû être effectuées : les champs non relevants (présence de Geonames p.ex.) ont été supprimés, les vides remplacés par "NULL" et les champs alphanumériques (Bibliographie et Remarques) ont également été ignorés. La fonction d'export de la V4 d'ArkeoGIS (figure 1) permet maintenant une intégration quasiment directe dans le prototype de visualisation d'IndexMed.

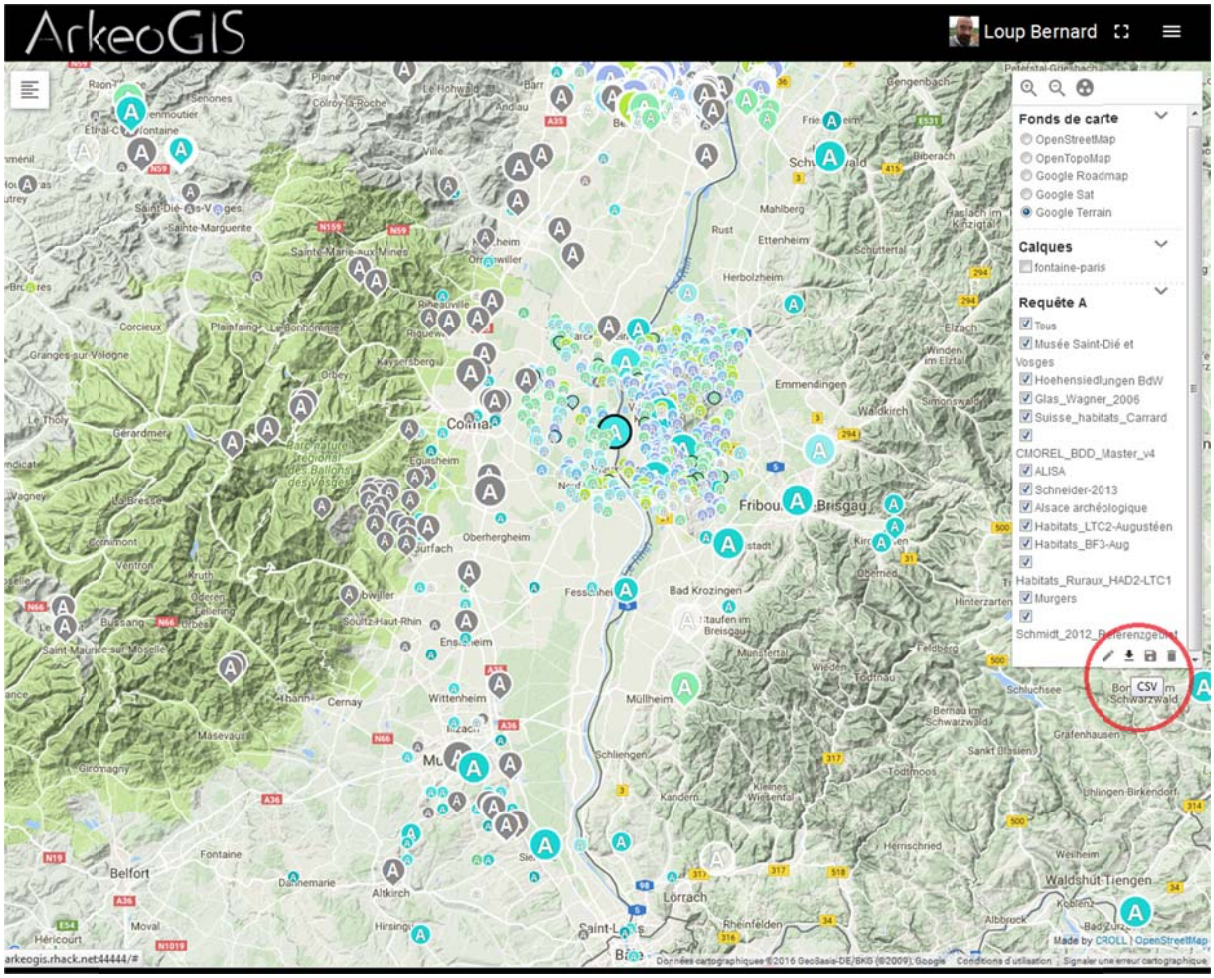


Figure 1. *Capture d'écran de la version 4 d'ArkeoGIS. Les bases ayant fourni des informations sur la région du Rhin Supérieur (extrait de requête sur le Rhin Supérieur de Bâle à Freiburg) sont listées à droite. La taille des points est indexée sur l'état de la recherche ; les sites qui n'ont pas de pointe vers le bas sont des centroïdes. Les codes couleurs correspondent à la chronologie choisie, ici "Europe continentale du Néolithique à nos jours". Les sites en gris sont de période indéterminée. Le curseur de la souris (cercle rouge) est sur l'icône qui permet d'enregistrer le résultat de cette requête au format CSV. C'est ce type de fichier qui a ensuite été utilisé avec des modifications mineures afin d'alimenter l'outil développé par le consortium IndexMed et de produire une représentation des données sous forme de graphes.*

Après cette présentation synthétique et assez formelle de l'intérêt de la mise en commun et de la représentation de données hétérogènes sous forme de graphes, voici parmi de multiples possibilités, deux premières représentations des données d'ArkeoGIS.

Prototype de visualisation et résultats préliminaires

Prototype de visualisation

L'interface du prototype utilise Neo4j <neo4j.com/>, une base de données à base de graphes mise en œuvre en java et publiée en 2010. L'édition communautaire de la base de données est sous licence GNU GPL v3. La base de données et ses modules supplémentaires (sauvegarde en ligne ou haute disponibilité) sont disponibles sous licence commerciale. Le prototype d'IndexMed permet à un opérateur néophyte d'importer des données (en CSV, XML ou JSON). Il permet d'interroger Neo4j pour produire le graphe et d'interagir avec lui à l'aide du navigateur Web. Le personnel technique d'IndexMed développe un frontend Web spécifique à l'aide du langage Ajax / Jquery. Il peut être possible de demander une base de données demandant des objets spécifiques et des relations spécifiques entre eux, sans utiliser un langage de requête technique tel que SQL ou Cypher.

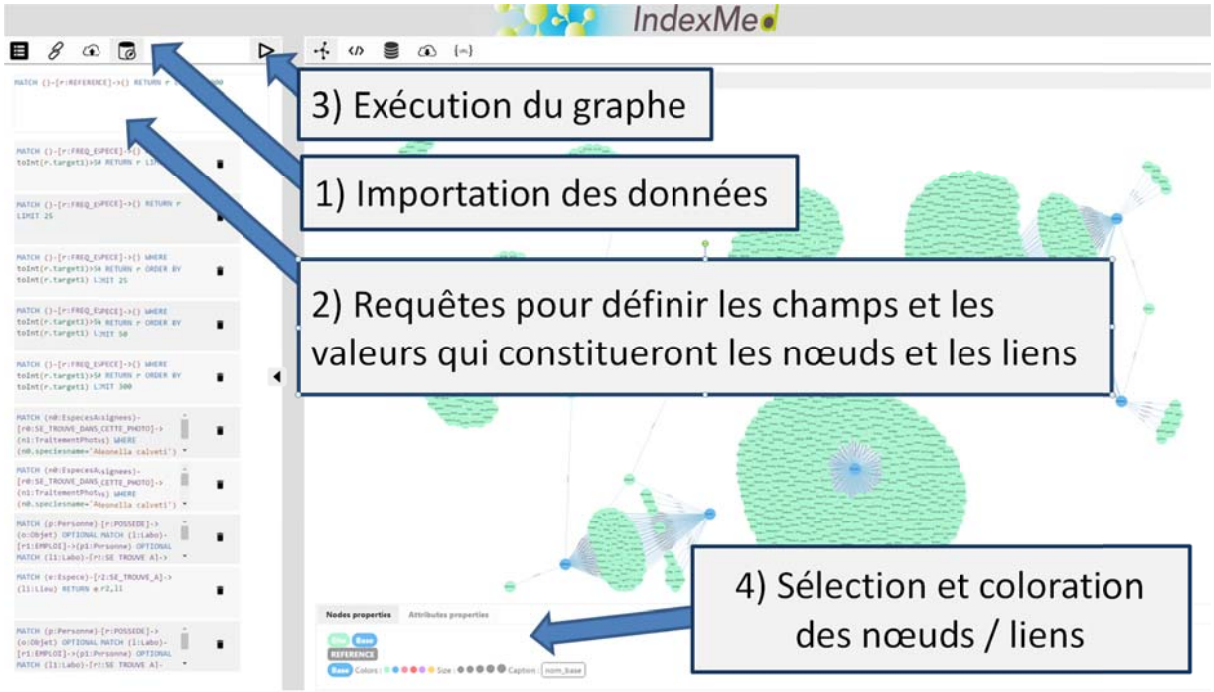


Figure 2. *Présentation générale du prototype d'IndexMed de visualisation des données représentant dans cet exemple 1492 sites d'archéologie en vert et 12 bases de données. Les informations proviennent de l'import d'ArkeoGIS mais pourraient directement être interrogées à distance (1) par le prototype sur les systèmes d'information des partenaires, (format JSON ou XML). La colonne de gauche permet d'importer, d'effectuer les requêtes (2) avec un formulaire ou le langage cypher et de les enregistrer. Un bouton (3) permet de lancer l'exécution de la nouvelle requête ou d'une requête pré-enregistrée. Le bandeau du bas (4) permet de configurer les couleurs des noeuds et des liens en fonction des valeurs de descripteurs.*

Le prototype est développé pour pouvoir être générique et permet d'intégrer n'importe quel type de données sous la forme de "valeur d'objet et d'attribut". Il suffit ensuite à l'opérateur de sélectionner la base à utiliser, les champs qui servent de noeuds, les champs qui servent de liens, et ceux qui servent à mettre en évidence des éléments de contextes. Il est aussi possible de faire ces opérations en sélectionnant certaines valeurs de champs. Ce prototype sera disponible sous forme de source ouverte pour développer, à moyen terme, l'utilisation de ces graphiques pour l'aide à la décision en matière de gestion environnementale et dans le cadre d'un projet de recherche à soumettre aux appels à projets européens (BiodivERsA, ERDF, SeasEra , H2020 ...)

Résultats préliminaires

Dans le premier graphe présenté, les bases sont des nœuds, les sites sont des nœuds reliés aux bases qui contiennent des données les concernant (figure 2), les descripteurs de base (auteurs, langues, types, descripteurs communs) permettent de colorer les nœuds ou de sélectionner une partie des bases ou des sites seulement.

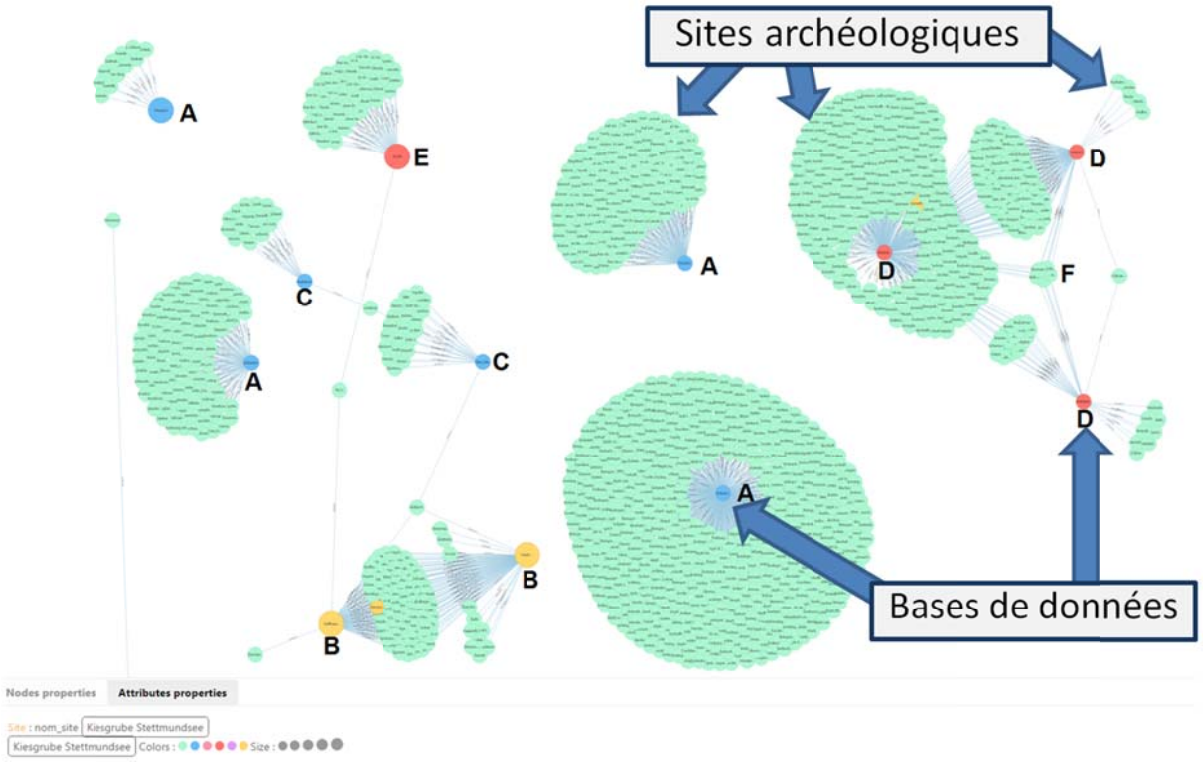


Figure 3. *Graphe avec extrait des données issus de la requête géographique présentée dans la figure 1 et reliant les bases avec les sites communs. Cette figure permet de mettre en évidence, concernant la requête réalisée, les bases qui référencent des informations sur des sites archéologiques qu'elles ne partagent avec aucune base (A), d'autres qui référencent l'essentiel des sites en commun avec une autre base (B). Dans le cas C, un seul site est commun aux autres bases. Les trois bases du cas D partagent une partie de leurs sites deux à deux, et Le cas F montre les sites communs aux trois bases à la fois. Le cas E est relié par un noeud aux cas B, par un "NULL", qui met en évidence une erreur dans le jeu de données (un enregistrement sans nom pour le site dans les deux bases de données reliées).*

Le fait de représenter les sites et les bases sur le même graphe (figure 3) permet d'étudier le système d'information en lui même. L'opérateur peut visualiser l'importance de chaque base dans la sélection géographique effectuée via Arkeogis. Il est possible aussi de sélectionner une période ou d'autres critères descripteurs des bases ou des sites eux même. Cette représentation permet aussi, comme le cas E dans la figure 3, de mettre rapidement en évidence des particularités ou des erreurs dans les jeux de données.

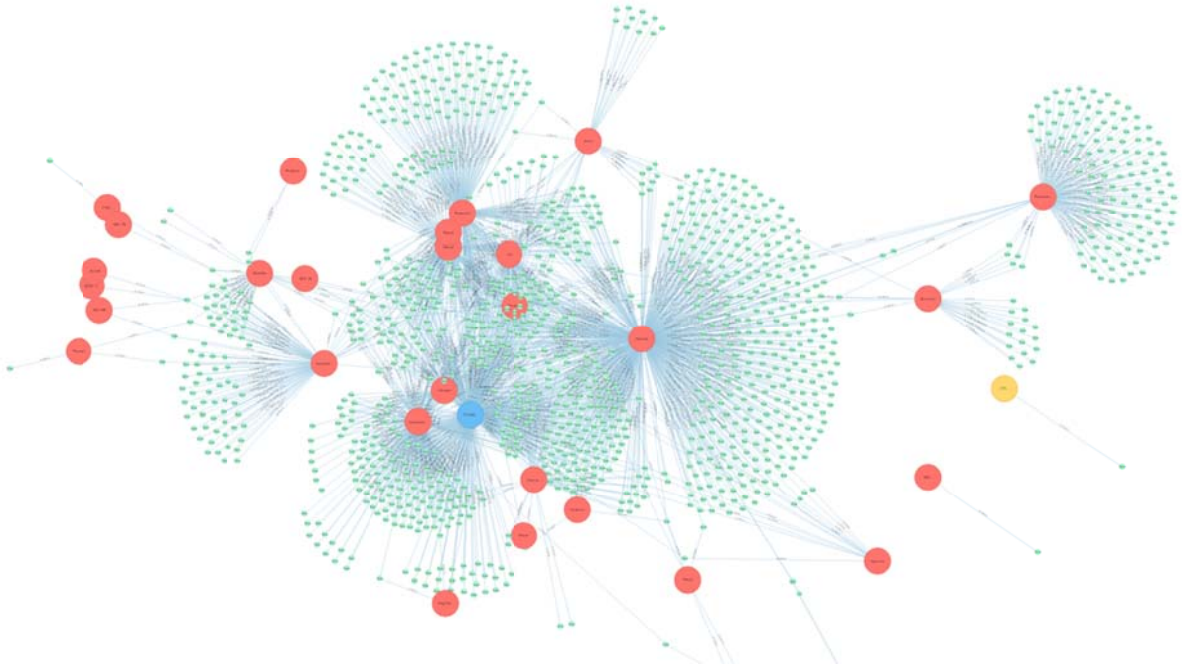


Figure 4. *Graphe issu de l'interface IndexMed, utilisant les données exportées à partir de la sélection géographique sur ArkeoGIS, représentant 1492 sites et 4950 liens matérialisant les types d'objets trouvés sur ces sites. Les sites sont représentés par les petits nœuds verts, les clusters rapprochant les sites ayant les mêmes topologies / patrons de descripteurs (ici sont utilisés uniquement les patrons d'objets de niveau 1 matérialisés par des nœuds rouges). Une sélection a été faite sur deux valeurs de descripteurs : "Bleu" pour les types d'objet de niveau 1 qualifiés de "céramique" (Le site contient donc au moins un objet de type céramique si il est lié à ce noeud), Jaune : le noeud correspond à un objet daté de la période exactement égale à "-900 -à -726".*

Le graphe suivant est un exemple de représentation du système observé (la sélection géographique de sites et les types d'objets de niveau 1). Les sites représentés au centre du graphe (figure 4) contiennent les patrons d'objets les plus communs, ceux à la périphérie des patrons plus particuliers. Certains sites déconnectés du graphe (exemple de la période sélectionnée égale à "-900 -à -726") contiennent des types d'objets (de niveau 1, les niveaux suivants étant moins systématiquement renseignés) que l'on ne trouve que sur ces sites. Des regroupements d'objets (groupes de nœuds en rouge à gauche du graphe) semblent typiques de quelques sites et sont donc rassemblés dans un cluster bien particulier, auquel correspond un contexte précis.

Lorsque des objets sont particuliers à un groupe de site (ici on a clairement un cluster d'objets en rouge, à gauche du graphe), on peut rechercher les contextes particuliers (groupes de valeurs de descripteurs significativement différents dans cette partie du graphe, par rapport aux autres clusters) grâce à des algorithmes adaptés. Le choix des algorithmes dépendent du type des objets représentés dans le graphe et de sa topologie.

L'étude de ces contextes associés à ces clusters est un champ d'investigation d'autant plus grand que le nombre de descripteurs de contextes consistant est important ; Sur 15 bases sélectionnées pour les graphes, en se limitant aux 16 champs communs de l'export utilisables en tant que descripteurs, on peut remarquer que même si 77% des champs sont renseignés de manière quasiment systématique (à plus de 90%), 20% des descripteurs sont renseignés en moyenne à moins de 60% (dont 12.1% des descripteurs sont renseignés en moyenne à moins de 10%) ce qui les rend inutilisables dans un des graphes tel que nous les proposons. Cela laisse apprécier la marge de progrès à faire si l'on souhaite élargir la liste de ces descripteurs.

Discussion et perspectives

Difficultés et propositions de résolutions

De manière générale, les verrous scientifiques à lever dans un projet de recherche interdisciplinaire futur devront être précisés par les experts du domaine des STIC, et concernent notamment i) l'augmentation des fréquences et de la densité d'acquisition des observations (développement des méthodes de reconnaissance automatique et déploiement d'outils d'acquisition moins onéreux), ii) la diversification des objets et des descripteurs d'objet intégrés dans les graphes, iii) la normalisation des descripteurs de la donnée et les méthodes permettant d'intégrer les différents niveaux de qualité des données.

Le premier verrou important, surtout lorsque l'on part de bases de données "empilées", est d'avoir un dénominateur commun suffisant à chacune de ces bases pour obtenir, pour chaque objet, au moins une valeur pour les descripteurs qui servent de liens. Cette recherche de consistance des données se traduit par l'élimination systématique des descripteurs qui ne sont pas majoritairement renseignés (idéalement, il faut au moins une valeur enregistrée pour chaque objet, en comprenant bien qu'un objet n'ayant pas de lien avec un autre n'a aucun intérêt à être représenté). Plus le nombre de descripteurs intégrés dans le modèle est grand, moins il doit y avoir d'objets sans valeur pour un descripteur.

La deuxième source de difficultées liée à cette approche est l'utilisation involontaire de descripteurs équivalents, ou au moins partiellement redondants : ceux-ci peuvent entraîner une "déformation" du graphe, et les rapports de distances entre les objets représentés être mal interprétés. Des techniques - à explorer - de comparaisons de topologies de graphes, faites avec les descripteurs dont on soupçonne la redondance, sont prévues dans le cadre du développement de cette recherche.

Pour intégrer ces formats différents de variables, il faut donc travailler sur le dénominateur commun pour chaque type de variable descriptive ou de contexte. Cela peut à minima être fait sous la forme de booléens (présence/absence), mais demande de faire accepter à tous les acteurs un processus permettant un consensus sur les descripteurs, leur définition, leurs nombres de valeurs possibles et les possibilités de dupliquer les descripteurs d'une même qualité avec des niveaux de précision différents et les équivalences entre ces niveaux.

Les perspectives interdisciplinaires de cette approche de fouille de données basée sur les graphes sont applicables à la majorité des sciences humaines et sociales. Elles requièrent la recherche et l'acceptation de termes/valeurs structurants dans les jeux de données, basé sur les standards dans chaque discipline (et donc une mise en cohérence des standards entre disciplines, ce qui est parfois un peu compliqué). L'internationalisation des standards, les nuances entre langues qui impactent la compréhension exacte des termes en anglais et l'évolution en propre des standards sont aussi des facteurs à prendre en compte.

Cette cohérence est une condition *sine qua* non de l'interdisciplinarité. Ce travail d'homogénéisation peut commencer par un relevé des problèmes de polysémie de descripteurs et de termes/valeurs structurants. Il doit éviter les jargons ou les usages trop locaux et/ou non entérinés par les communautés. L'objectif doit être d'élaborer une proposition d'amélioration des standards par recherche de consensus entre les communautés, et ceci de manière itérative.

La taxonomie utilisée pour désigner des descripteurs et leur donner des valeurs a une importance primordiale pour construire un graphe qui puisse répondre à un questionnement scientifique. Elle agit principalement sur sa topologie. Sans entrer dans les détails, par exemple, son pouvoir descripteur / discriminant dépend du nombre de valeur pour chaque descripteur et de la répartition des valeurs possibles du descripteur sur l'ensemble des enregistrements.

Les archéologues sont majoritairement des littéraires, leurs outils sont traditionnellement bibliographiques, et depuis quelques décennies maintenant statistiques. Il nous semble essentiel d'accompagner le transfert vers de nouveaux outils, par exemple en incluant de la bibliographie dans le code, ou en codant des articles ou des hypothèses.

Afin que les participants au projet ne se sentent pas tributaires d'une "*black box*" (i.e. le fait que les utilisateurs non-avertis ne comprennent pas l'intégralité des calculs effectués), le code sera fourni en open source (avec la bibliographie correspondante). Cela devrait permettre à l'opérateur d'identifier et / ou de modifier le code.

Concernant enfin la manipulation de variables dans les graphes, afin de lever les suspicions de boîte noire, une bonne méthode nous semble de tester des hypothèses simples et avérées. Pour une approche archéologique, cela pourrait correspondre à des faciès ou des groupes culturels connus, par exemple, afin de vérifier que ces entités apparaissent regroupées sur le graphe aussi clairement que sur le SIG. Les graphes pourraient permettre ici de tester des marqueurs annexes qui "ressortent" alors qu'ils n'avaient pas été pris en compte lors de l'hypothèse originale. Une démarche secondaire et complémentaire pourrait être de tester une hypothèse absurde afin de mettre en avant le fait que l'outil ne fonctionne pas s'il est mal utilisé.

Perspectives

Les premiers tests de ces méthodes sont néanmoins encourageants car ils permettent d'appréhender tout ou partie du système observé (un ensemble de sites) en intégrant des données de contextes de format différents. Ils permettent aussi d'étudier le système d'observation ainsi que les efforts de prospection, ou d'avoir une approche visuelle de la répartition des compétences en archéologie. Les aspects chronologiques et les requêtes spatiales n'ont pas été mises en oeuvre de manière plus poussée, car un travail d'homogénéisation de ces descripteurs est encore nécessaire (nombre de catégories et valeurs partagées par toutes les bases de données utilisées). Dans un second temps, ils feront l'objet d'analyses plus complexes basés sur des graphes issus de requêtes plus spécifiques. Chacun de ces trois aspects pourra être développé dans le cadre de futures recherches interdisciplinaires, dans lesquelles les questionnements scientifiques en archéologie côtoient les questions scientifiques en sciences et techniques de l'information et de la communication.

La mise en place d'une dynamique d'échange entre des experts en écologie / biodiversité / archéologie et des experts du domaine des STIC est la prochaine étape prévue par le consortium IndexMed. Proposée à différents financeurs sous la forme d'une action, celle-ci regroupe des représentants des deux champs disciplinaires et permettra de formaliser des besoins en terme d'analyses de données hétérogènes de la part de la communauté écologie / biodiversité / archéologie et de stimuler la recherche en STIC afin de proposer des solutions plus adéquates pour l'analyse et la gestion des données écologiques dans le contexte du Big Data (prise en compte de la dimension temporelle et spatiale et des données multi-échelles et hétérogènes).

En ce qui concerne les techniques et approches STIC étudiées et développées, la recherche sera dirigée vers des techniques de gestion et d'analyse de graphes qui puissent prendre en compte la complexité des données hétérogènes et notamment passer à l'échelle sur des jeux des données volumineux sans détériorer la qualité des résultats obtenus. Sous l'égide d'IndexMed, il est prévu de réaliser une première carte des compétences de laboratoires en informatique qui pourront apporter des outils méthodologiques ou des techniques algorithmiques adéquates pour l'analyse des données issue de l'écologie et des Sciences humaines et sociales.

Conclusion

Les approches proposées dans le cadre de la collaboration entre ArkeoGIS et IndexMed seront testées sur d'autres correspondances de patrons utilisant des objets hétérogènes dans les domaines des sciences environnementales liées à l'archéologie. L'ambition est de développer des modèles d'études libres et ouverts sur le long terme aux experts de l'analyse de données (STIC), ainsi que des processus de test et de choix de nouveaux algorithmes à l'échelle globale pour les utilisateurs potentiels (environnementalistes comme archéologues dans le cas décrit ici).

Cet élargissement de la collaboration sera l'occasion de préciser les besoins, intérêts et attentes de chaque communauté, en termes de recherche autant que de formation. Une plateforme ouverte aux collaborateurs désirant investiguer ces nouvelles méthodes, pourrait prendre par exemple la forme d'une « forge » permettant aux deux communautés de faire évoluer leurs recherches sur le long terme, mettant ainsi en place un vrai *Linked Open Data* utilisable par les chercheurs travaillant sur des sujets interdisciplinaires à l'aide de nos outils. Les acteurs de la recherche intéressés peuvent prendre contact avec les deux communautés via arkeogis.org et indexmed.eu

Remerciements

La construction du premier prototype du consortium IndexMed a été est financé par le défi CNRS « VIGI- GEEK1 » et le PEPS Blanc CNRS INEE avec le projet "Charliee2".

Nous remercions tous les membres actifs du consortium IndexMed pour leurs contributions et les GDR MaDICS et EcoStat pour leurs labellisations et soutiens. Les auteurs tiennent évidemment à remercier leurs communautés respectives, concernant ArkeoGIS plus particulièrement les auteurs des bases utilisés : G. Hoffmann, M. McCormick, C. Morrissey, C. Morel, M. Trautmann, N. Schneider, H. Wagner, C. Jeunesse, M. Roth-Zehner, D. Schwartz et C. Schmid-Merkl, et pour la relecture effectuée par Dino Ienco concernant les termes propres aux STIC.

Références

AGGARWAL, C & WANG, H. 2010. C. Aggarwal, H. Wang, *Managing and Mining Graph Data*, Springer, 1st Edition., 2010, XXII, 600 p.

BERNARD et al. 2015. Bernard L., Ertlen D., Schwartz D., "ArkeoGIS, Merging Geographical and Archaeological Datas Online",*in* Giligny F., Djindjian, F., Costa, L., MoscatiI, P., Robert, S. (éds.) *Concepts, methods and tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology* Paris, Archaeopress 2015 : 401-406.

DAVID et al 2015. David R., J.-P. Féral, C. Blanpain, C. Diaconu, A Dias, S. Gachet, K. Gibert, J. Lecubin, C Surace, "A first prototype for indexing, visualizing and mining heterogeneous data in Mediterranean ecology within the IndexMed consortium interdisciplinary framework". *In*: SITIS 2015, *11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Bangkok, Thailand, pp. 232-239, nov. 2015 doi: 10.1109/SITIS.2015.119.

DAVID et al 2016. David R., J.-P. Féral, A-S. Archambeau, N. Bailly, C. Blanpain, V. Breton, A. De Jode, A. Delavaud, A. Dias, S. Gachet, D. Guillemain, J. Lecubin, G. Romier, C. Surace, L. Thierry de Ville d'Avray, C. Arvanitidis, A. Chenuil, M.E. Çinar, D. Koutsoubas, S. Sartoretto, T. Tatoni ; "IndexMed projects : new tools using the CIGESMED DataBase on Coralligenous for indexing, visualizing and data mining based on graphs". In : Sauvage S, Sánchez-Pérez J-M., Rizzoli, A.E. (Eds.), *Proceedings of the 8th International Congress on Environmental Modelling and Software*, *Environmental modelling and software for supporting a sustainable future*, Vol. 3, pp.656-665, Toulouse, France. July 2016

LAMBERT et al 2013. Lambert A., R. Bourqui, D. Auber, "Graph Visualization for Geography. Methods for Multilevel Analysis and Visualisation of Geographical Networks", *Springer*, : 81-102, 2013. <hal-00841188>

Analysis on the Graph Techniques for Data-mining and Visualization of Heterogeneous Biodiversity Data Sets

Víctor Méndez Muñoz[1], Anna Cohen-Nabeiro[2], Romain David[3], Vicente José Ivars Camáñez[1], Alfons Nonell-Canals[5], Miquel Angel Senar[1], Denis Couvet[4], Jean-pierre Feral[3], Aurélie Delavaud [2] and Thierry Tatoni[3]

[1]*Department of Computer Architecture & Operating Systems (CAOS), Universitat Autònoma de Barcelona (UAB), Bellaterra (Barcelona), Spain*

[2]*Fondation pour la Recherche sur la Biodiversité (FRB), Paris, France*

[3]*Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE), CNRS, Aix Marseille Université, IRD, and Université d'Avignon, France*

[4]*Museum National d'Histoire Naturelle, Paris, France*

[5]*Mind the Byte, Barcelona, Spain*

Keywords: Biodiversity Data Mining, Ontology Engineering, Biodiversity Metadata Visualization, Graph.

Abstract: Extisting biodiversity databases contain an abundance of information. To turn such information into knowledge, it is necessary to address several information-model issues. Biodiversity data are collected for various scientific objectives, often even without clear preliminary objectives, may follow different taxonomy standards and organization logic, and be held in multiple file formats and utilising a variety of database technologies. This paper presents a graph catalogue model for the metadata management of biodiversity databases. It explores the possible operation of data mining and visualization to guide the analysis of heterogeneous biodiversity data. In particular, we would propose contributions to the problems of (1) the analysis of heterogeneous distributed data found across different databases, (2) the identification of matches and approximations between data sets, and (3) the identificaton of relationships between various databases. This paper describes a proof of concept of an infrastructure testbed and its basic operations, presenting an evaluation of the resulting system in comparison with the ideal expectations of the ecologist.

1 INTRODUCTION

Accurate and publicly available information on biodiversity observations can contribute to scientific knowledge, foster multidisciplinary studies, and provide new perspectives to environmental and societal responses including decision-making (Lausch et al., 2015). To this end, several biodiversity metadata projects have been established which describe and characterize the information hosted in a range of distributed databases (David et al., 2016; Dodge et al., 2013).

As these metadata projects grow horizontally, with more databases and types of data sets, as well as vertically, with more documents, the ecologist will need information management tools to enable the following common tasks:

1. To discover existing but heterogeneous, dispersed data sets of different origins and scales of observation;
2. To discover relationships between documents of potential interest to the scientist;
3. To interpret the semantic meaning of relations without the need to know the meta-model;
4. To enable the scientist to understand the data context and collection methods of multiple fields and topics;
5. To determine the quality associated with the data, including the data sets inter-calibration; and
6. To be aware of the conditions of access and use.

This proof of concept presents a case study of the ECOSCOPE metadata catalogue (Taffoureau et al., 2016; Eco,) which provides data mining and visualization capabilities. ECOSCOPE is a metadata collection service for databases of different fields of ecology in *lato sensu*.

We follow a behaviour-driven development (BDD)(Solis and Wang, 2011) of a minimal operational set and the assessment of the ecologist at each operation. The resulting evaluation is used to

Muñoz, V., Cohen-Nabeiro, A., David, R., Camáñez, V., Nonell-Canals, A., Senar, M., Couvet, D., Feral, J-p., Delavaud, A. and Tatoni, T.
Analysis on the Graph Techniques for Data-mining and Visualization of Heterogeneous Biodiversity Data Sets.
DOI: 10.5220/0006379701440151
In *Proceedings of the 2nd International Conference on Complexity, Future Information Systems and Risk (COMPLEXIS 2017)*, pages 144-151
ISBN: 978-989-758-244-8

propose a new graph catalogue service architecture. The new graph catalogue can be used for metadata discovery and visualization, integrated with the existing and future data management service. The current ECOSCOPE web catalogue is used to collect metadata in a standardized way, using an authorization service which provides the ecologist accredited access to various storage systems.

2 RELATED WORK AND MOTIVATION

The consortium IndexMed (renamed recently "IndexMeed - Indexing for Mining Ecological and Environmental Data" to build international projects) was created by the axis " Management of biodiversity and natural spaces " of the IMBE (Mediterranean Institute of freshwater and marine Biodiversity and Ecology) (David et al., 2015). Its main goal is to develop awareness of databases and their effective use in the ecological research community. This consortium is particularly useful as a bridge between existing networks and initiatives at national and international levels. The aim of the consortium is to index biodiversity data (and to provide an index of qualified existing open datasets) and to make it possible to build graphs to assist in the analysis and the development of new ways to mine data. Standards and specific protocols can be applied to interconnect databases. Semantic approaches greatly increase data interoperability. The project should develop new transdisciplinary methods of data analysis, focusing on open data, open source and free methods and development tools.

ECOSCOPE is an infrastructure funded and managed by French research organisations through the Foundation for Research on Biodiversity (FRB) which ensures its coordination. The scientific aim is to document the state and trends of biodiversity and ecosystem services, enabling scenarios for the future to be built. In this framework, ECOSCOPE promotes the complementarity of observations and links between research observation systems that vary across spatial and temporal scales, variables, studied ecosystems and kingdoms, levels of organization and data sources. In cooperation with existing initiatives, ECOSCOPE provides an entry point for the discovery of observations and datasets for research on biodiversity across the entire data life cycle, facilitating links between data producers and users.

Note - for deletion on final version: Scholes does not refer to ECOSCOPE. hence it cannot be used in this manner as a reference. You could say "These aims are consistent with Scholes et al.(2012)."

The ECOSCOPE metadata catalogue delivers freely available online information about who, where, what, when, why and how the research observation data were collected. It is build on the EBV concept, developed by GEO BON (Group on Earth Observation Biodiversity Observation Network), which is designed to serve as the foundation for interoperable sub-national, national, regional and global monitoring initiatives.

Precise and public information on biodiversity observation datasets contribute to data openess and reuse, in full conformity with data producers and owners. Metadata formats the description, characterisation and specification of data hosted in datasets, allowing the discovery of data, whether heterogeneous or dispersed and across locational and observational scales. Metadata permits the understanding of the context of the dataset, collection methods and data quality. It gives information on access and use of data and other resource conditions and the contact persons (Michener 2006).

Michener W.M. (2006) Meta-information concepts for ecological data management. Ecological Informatics 1:3-7

The ECOSCOPE metadata catalogue answers to this need thanks to providers and exchanges with other information systems. As it is based on standards in use, the metadata profile can be exported into other information systems, and metadata files (such as Ecological Metadata Language: EML) can be imported into the ECOSCOPE metadata portal. It contributes to global efforts to make research on biodiversity data more available for scientific projects, synthesis and indicators.

In this context of a prototype for collecting ecological metadata of various fields and topics, our motivation is to explore the possibilities of graph techniques for visualization and data mining supported in graph databases. The graph databases have been a proven feasible backend to provide semantic services (Riesen and Bunke, 2008; Angles and Gutierrez, 2008). Furthermore, the indexing capabilities of graph databases ensures the scalability of the system response (Williams et al., 2007), which is a critical factor in our needs to increase various databases, data sets, and document integration.

In a graph catalogue the database mapping model is isomorphic with the represented structure. The resulting model enables the evolution of applications with linear complexity in the data mining operation, which is critical for scale-up in data volume and variety.

The overall architecture of our vision is shown in Figure 1. There are increasing number of database

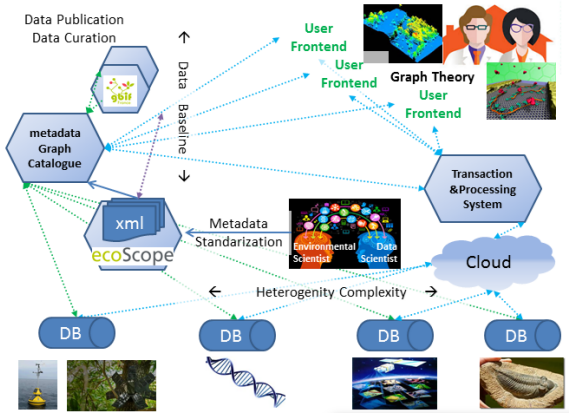


Figure 1: The proposed metadata graph catalogue in the IndexMed overall architecture.

managers adopting a metadata standarization process, using the ECOSCOPE portal to deliver a meaningful metadata description into the ECOSCOPE database. Other external sources are used to complement related information about curation and publication, like GBIF (Flemons et al., 2007). GBIF attempts to bring together all biodiversity and collections data to make them available to researchers and the general public. To do this, the GBIF provides a search engine for databases connected to GBIF in a standardized way. Data owners can connect all or part of their resources to GBIF to make them visible and interoperable, but they keep the control of their data, which they continue to host and use in their work.

In the current prototype architecture of GBIF, there is no vertical solution to secure access into the storage systems, neither high level facilities for semantic data. In this paper we are exploring and analyzing the possibilities of the high level semantic data operations.

3 A SEMANTIC METADATA SERVICE WITH GRAPH CATALOGUE

To prove the concept, we have dumped the current ECOSCOPE document database into a graph database and we test the ecologist operations. Scientists produce knowledge by analysing data into information and the goal is to elaborate theories from information. Data constitute the primary material from which hypotheses are first formulated, then refined and validated. Metadata permit the data openness and data sharing, as the way to give value to data after their primary use (McNutt et al., 2016).

3.1 Basic Visualization Operations

This section presents the general visualization of the graph with all the metadata nodes, and two other general visualizations of all data sets, but without all the metadata nodes.

3.1.1 Operation: Show All Graph

Given ecologist could not be aware of the meta models of multidisciplinary data sets.
When ecologist likes to analyze the possibilities of multidisciplinary studies because it is the only way to better understanding systemic interactions between factors.
Then it is needed and overall meta model view with browse capabilities.
Test:

```
MATCH (n)
OPTIONAL MATCH (n)-[r]-()
RETURN n,r
```

Assets: Displays the hold graph of the metadata catalogue without any previous knowledge of the meta-model. It can be a good starting point to get the number of nodes (300) and relations (1137). The nodes are: Address(5), Attribute(48), Dataset(30), Description(17), GPolygonOuterRing(14), GeographicCoverage(19), Keyword(79), Person(8), TaxonomicClassification(45), TaxonomicCoverage(31), TemporalCoverage(4)

Weakness: There is limited interactive usability of the graph method in large meta models, because it is difficult to visually manage too many objects—300 in our proof of concept—ideally less than 100 nodes are recommended.

3.1.2 Operation: Show Graph for Spatial and Temporal Relations

Given the complete graph nodes and their direct relations can be categorized as follows:

- Data set core information: Dataset→Description; Person→Address
- Temporal and spatial information: GeographicCoverage→GPolygonOuterRing; TemporalCoverage
- Information of data set classification: Attribute; TaxonomicCoverage→TemporalCoverage; Keywords

When core information is the spine in the structure and the two other categories are more specific,
Then it can be of interest the visualization and

browse of a graph focused in the core information with temporal and spatial information.
Test:

```
MATCH (n)
WHERE NOT n:Attribute
AND NOT n:Keyword
AND NOT n:TaxonomicClassification
AND NOT n:TaxonomicCoverage
RETURN n
```

Assets: Here a clear segmentation of the graph in 5 categories is obtained (Figure 2). This general spatial temporal graph shows two types of node segments. On one hand, some segments are irrelevant because a single data set (in blue) is related to its core information: the three segments in the bottom right. In the other hand, node segments with several data sets are related with temporal spatial information. For example, the segment on the top shows all the data sets (nodes in blue) are related to a single geographical area (node in yellow), even when they have been tagged with different geographical names in the database (nodes in pink).

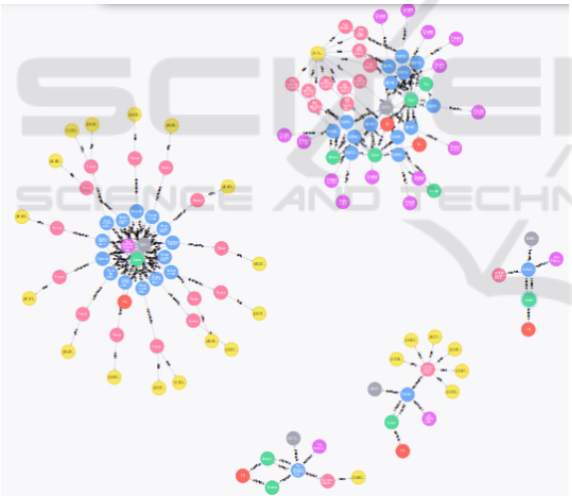


Figure 2: A general view of the spatial temporal graph.

Weakness: The resulting segmentation does not show relations between data sets of different databases. Eventually, a more precise matching in geographical area is needed, for example by area proximity or overlap. Another approach could be to draw such areas in the map to give to the ecologist a visual map of the data sets.

3.1.3 Operation: Show Graph for Taxonomy and Organizational Relations

Given the node classification above.
When it is needed for analysis of data set categories,
Then it can be of interest to the visualization and browse of a graph focused in core information with the organizational logic metadata.
Test:

```
MATCH (n)
WHERE NOT n:Attribute
AND NOT n:TemporalCoverage
AND NOT n:GeographicCoverage
AND NOT n:GPolygonOuterRing
RETURN n
```

Assets: Figure 3 shows a clear segmentation of data sets (nodes in blue) by the organization logic metadata of Keyword (in red), TaxonomicClassification (in green) and TaxonomicCoverage (in pink), but with all the nodes connected, which is of high interest to enable multidisciplinary relationship discovery. Our results show some metadata fields which are relation-hubs between data sets of different databases, particularly a few generalist TaxonomicClassification values and Keywords.

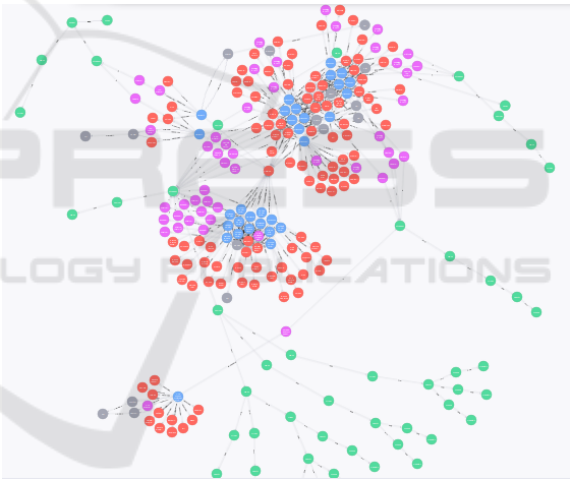


Figure 3: A general view of the organizational logic graph.

Weakness: Even when the visualization tool is able to do a zoom of Figure 3, this is not enough to ensure a systematic discovery.

3.2 Common Data Mining Operations

This sub-section presents operations in the metadata graph database to provide a subset of the graph according to the behaviours requited by the ecologist, which have been described in the enumeration of the introduction section.

3.2.1 Operation: Common 1

Given the general graph visualizations above,
When ecologist want to discover existing but hetero-

geneous, dispersed data sets of different origins and scales of observation;
Then restrict the graph of visualization operation 3.1.2, which contains geographical origins and temporal scales, to match a single hub node and related data sets. The hub node is taken from previous general view of operation 3.1.3
Test:

```
START keyword=node(*)
MATCH (n)<-[]-(d)-[r]->(keyword)
WHERE keyword.word = "AGROVOC"
AND NOT n:Attribute
AND NOT n:Keyword
AND NOT n:TaxonomicClassification
AND NOT n:TaxonomicCoverage
RETURN n,d,r,keyword
```

Assets: In Figure 4 a zoom view of all the data sets related to the hub metadata. In red the hub node *(Keyword=AGROVOC)*. The cluster on the top is a segment of DIMPIE data sets, of the same database, with a temporal coverage in green (1975) and to the left there is a blue data set *(Collection moisissures IFV)*, whith temporal coverage of 2008 in green. So both databases and datasets are related by the hub node.

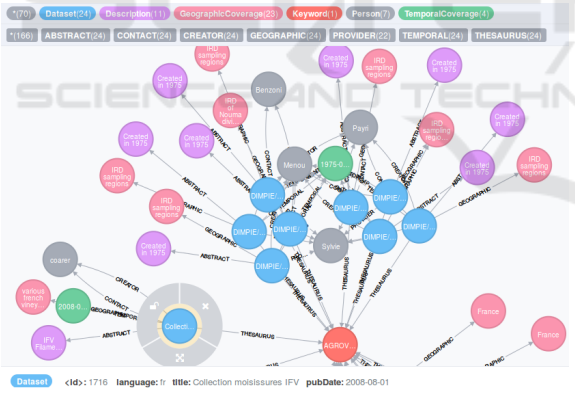


Figure 4: A zoom of different origins and scale matching with a hub node.

Weakness: There is no systematic way of filtering origins and scales.

3.2.2 Operation: Common 2

Given the metadata catalogue,
When the ecologist wants to discover relationships in a document of potential interest;
Then starting from operations to match data sets, filter the desired documents.
Weakness: In our case study the metadata source has not the details of each document. It is necessary to collect the metadata information of the document in the meta catalogue.

3.2.3 Operation: Common 3

Given a data set,
When the ecologist wants to interpret the semantic meaning of relationships without the need to know the meta-model;
Then to dig in the relationships without an explicit relationship label

Test:

```
START d1=node(*)
MATCH (d1)-[*1..5]->(n)
WHERE d1.title =~ "Donkey.*"
AND NOT n:Attribute
AND NOT n:Keyword
AND NOT n:TaxonomicClassification
AND NOT n:TaxonomicCoverage
RETURN d1,n
```

Assets: Figure 5 illustrates the great possibilities of the graph approach to describe meta-model semantics, without explicit knowledge of the model. The clause *MATCH (d1)-[*1..5]→(n)* gets a maximum depth of 5 levels, and the result shows a maximum of only two levels of relations from the given data set (node in blue). The rest of the MATCH clause is restricting the results to the basic information and the spatial temporal information of the visualization operation in 3.1.2. Another interesting filter would be to show the taxonomy and organizational relations of the visualization operation in 3.1.3.

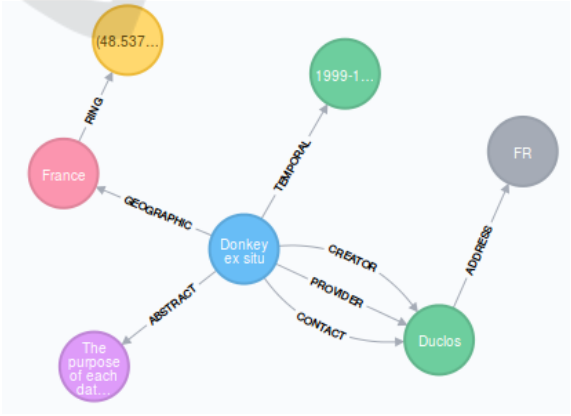


Figure 5: Semantic visualization of relations of a given data set.

Weakness: The test command shows only outbound relations from the given data set. Eventually, a more generalist operation shall ask the ecologist

whether it is requested outbound, inbound or both relations to or from a given data set.

3.2.4 Operation: Common 4

Given a data set,
When To guide the ecologist to understand the data context and collection methods;
Then to dig in the collection and context information of the data set.

Test:

```
START d1=node(*)
MATCH (d1)-[*1..8]->(n)
WHERE d1.title =~ "Donkey.*"
AND NOT n:Address
AND NOT n:Person
AND NOT n:TemporalCoverage
AND NOT n:GeographicCoverage
AND NOT n:GPolygonOuterRing
RETURN d1,n
```

Assets: Given a data set in blue, Figure 6 shows on one hand the information on collection methods about taxonomic coverage (in yellow) and the corresponding sub-graph of the taxonomic classification in green. The names displayed are the category of the classification, while by clicking in a particular green node will give the corresponding value for the data set. On the other hand, the context information is shown in the attributes (in grey) of the documents in the data set, as well as the keywords (in red) of the data set. Spatial and temporal information would be other interesting data context information.

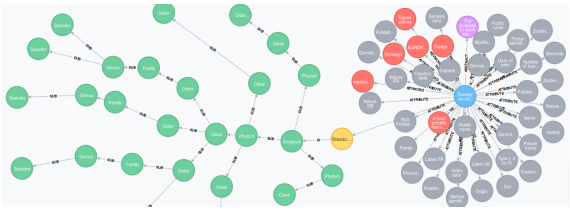


Figure 6: Data context and collection methods.

Weakness: Even when we have the attribute list of the documents, we don't have a document catalogue, so this information is of little value. Eventually we should include the document catalogue in the graph catalogue.

3.2.5 Operation: Common 5

Given a data set,
When the ecologist would like to estimate the quality associated with the data, including the data sets inter-calibration;
Then Show detailed information about the data quality of the corresponding nodes and inter-calibration.

Assets: Figure 7 shows the content of the Description node of a data set, which gives information about the associated data quality and a few calibration details.

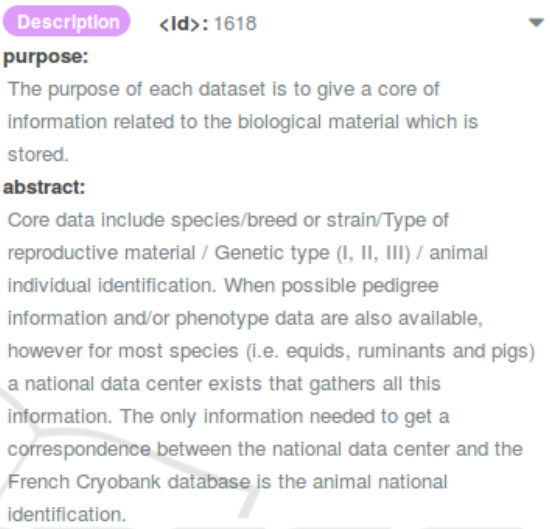


Figure 7: Data quality and callibration details.

Weakness: There is more quality and calibration information in some of the Attribute nodes. However, the name of the Attribute with valuable information is dependent on the particular data set. Therefore, it will be necessary to include a label in those Attribute nodes which are related to calibration and data quality to display such nodes for a given data set so the ecologist could browse the details.

3.2.6 Operation: Common 6

Given the metadata catalogue,
When the ecologist wants to be aware of the conditions of access and use of data sets and more documents;
Then provide access policy to sets and objects
Weakness: In our case study the metadata source only provides a secondary way to obtain the data, by giving the contact person and web information for a data set. So the ecologist can manually manage their access to the data, and there is no automation in this behaviour. This is a critical point to overcome the current collection scope by tools and methods to enable the access policy to the existing database objects.

4 SERVICE ARCHITECTURE

These tests have demonstrated the feasibility of graph techniques to provide semantic features in visualization and data mining of ecological metadata. However, to facilitate the ecologist's discovery and visualization, it also is necessary to provide high level applications alongside the existing graph database. The weakness analysis on the common expected operations, points to the need for more generic operations and systematic approaches, adapted to the characteristic multidisciplinary database of the ecologist.

The required behaviour on several of the common operations needs the inclusion of the metadata of the documents, not only as generic information of the data set.

- *Common 2* and *Common 4* need the integration metadata of the documents in the catalogue.
- *Common 6* is a critical operation to provide automated access policies to the documents.

For these reasons the present paper proposes a model-view-controller (MVC) service architecture(Deacon, 2009) as show in Figure 8

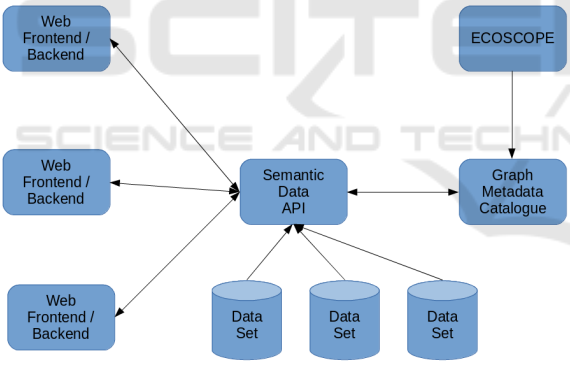


Figure 8: Proposed MVC Service Architecture.

- **View-Controller**
  - **Web Frontend/Backend** a common web framework for all the specific webs of the multidisciplinary studies, including basic frontend forms, backend handlers and driver connectivity to the common API. It supports the identity management and it is the user entry point to the data.
- **Model**
  - **Semantic Data API** Including all the high level methods for visualization, data mining and the gateway for data access policies to various storages.
  - **Graph Metadata Catalogue** With the existing meta model for data sets, but also including the metadata of the documents, as well as the access policies between identities and documents.
  - **Data Set** with secured access to the documents
  - **ECOSCOPE** the metadata collection and standardization portal.

The developing, releasing and deployment of the service components can be enabled by a container compose (Mulfari et al., 2015) or virtual machine infrastructure manager (Caballer et al., 2015).

5 CONCLUSIONS

This paper presents preliminary studies on metadata semantics. Indeed, it demonstrates improved metadata qualification needs using tools, standards and recommendations at both national (SINP [National Information System on Biodiversity], RBDD [Network of Research Databases]) and international levels (MedOBIS [Mediterranean Ocean Biogeographic Information System], OBIS, GBIF (Cryer et al., 2009), Life-Watch, GEO-BON, etc.) or shared by other research entities (i.e. IRD [Institute of Research for the Development] or MNHN [National Museum of Natural History, Paris])

The proof of concept demonstrates the pontential of graph databases to enable metadata visualization and common operations in a scalable way through the graph database capabilities in the horizontal relations. Furthermore, it has identified the commonalities for a high level semantic data API, into a service architecture for several specific web front-ends, contributing to economies of scale in the development and exploitation of the information system.

The promising results encourage future work following the proposed service architecture, to facilitate ecological studies in heterogeneous fields and topics with their increasingly complex requirements.

ACKNOWLEDGEMENTS

The authors would like to thanks Alison Specht, director of CESAB (FRB) and to Robin Goffaux from FRB for they advisory and support to this paper. This work is co-funded by the EGI-Engage project (Horizon 2020) under Grant number 654142 and by the Spanish MICINN project number TIN2014-53234-C2-1-R. IndexMed consortium is funded by the CNRS défi "VIGI-GEEK (VIsualisation of Graph

In transdisciplinary Global Ecology, Economy and Sociology data-Kernel)", CNRS INEE through the "CHARLIEE" project in 2015 and CNRS "Mission pour l'Interdisciplinarité" in 2016. Data used for this article were obtained through ECOSCOPE metadata tools. The authors acknowledge the support of France Grilles for providing computing resources on the French National Grid Infrastructure. Supplementary acknowledgement to organisers of the EGI workshop "design your e-infrastructure" which started this work.

REFERENCES

Ecoscope metadata portal. http://ecoscope.fondationbiodiversite.fr/fr/portail-de-metadonnees. Accessed: 2017-01-30.

Angles, R. and Gutierrez, C. (2008). Survey of graph database models. *ACM Computing Surveys (CSUR)*, 40(1):1.

Caballer, M., Blanquer, I., Moltó, G., and de Alfonso, C. (2015). Dynamic management of virtual infrastructures. *Journal of Grid Computing*, 13(1):53–70.

Cryer, P., R., H., C., M., Nicolson, N., Tuama, ., Page, R., Rees, J., Riccardi, G., Richards, K., and Whitev, R. (2009). Adoption of persistent identifiers for biodiversity informatics.

David, R., Feral, J.-P., Archambeau, A.-S., Bailly, N., Blanpain, C., Breton, V., De Jode, A., Delavaud, A., Dias, A., Gachet, S., et al. (2016). Indexmed projects: new tools using the cigesmed database on coralligenous for indexing, visualizing and data mining based on graphs.

David, R., Feral, J.-P., Gachet, S., Dias, A., Blanpain, C., Lecubin, J., Diaconu, C., Surace, C., and Gibert, K. (2015). A first prototype for indexing, visualizing and mining heterogeneous data in mediterranean ecology: Within the indexmed consortium interdisciplinary framework. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2015 11th International Conference on*, pages 232–239. IEEE.

Deacon, J. (2009). Model-view-controller (mvc) architecture. *Online][Citado em: 10 de março de 2006.] http://www. jdl. co. uk/briefings/MVC. pdf*.

Dodge, S., Bohrer, G., Weinzierl, R., Davidson, S. C., Kays, R., Douglas, D., Cruz, S., Han, J., Brandes, D., and Wikelski, M. (2013). The environmental-data automated track annotation (env-data) system: linking animal tracks with environmental data. *Movement Ecology*, 1(1):3.

Flemons, P., Guralnick, R., Krieger, J., Ranipeta, A., and Neufeld, D. (2007). A web-based gis tool for exploring the world's biodiversity: The global biodiversity information facility mapping and analysis portal application (gbif-mapa). *Ecological Informatics*, 2(1):49–60.

Lausch, A., Schmidt, A., and Tischendorf, L. (2015). Data mining and linked open data–new perspectives for data analysis in environmental research. *Ecological Modelling*, 295:5–17.

McNutt, M., Lehnert, K., Hanson, B., and Nosek, B. A.and Ellison, A. M. K. J. L. (2016). Liberating field science samples and data. *Science*, 6277:1024–1026.

Mulfari, D., Fazio, M., Celesti, A., Villari, M., and Puliafito, A. (2015). Design of an iot cloud system for container virtualization on smart objects. In *European Conference on Service-Oriented and Cloud Computing*, pages 33–47. Springer.

Riesen, K. and Bunke, H. (2008). Iam graph database repository for graph based pattern recognition and machine learning. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 287–297. Springer.

Solis, C. and Wang, X. (2011). A study of the characteristics of behaviour driven development. In *Software Engineering and Advanced Applications (SEAA), 2011 37th EUROMICRO Conference on*, pages 383–387. IEEE.

Taffoureau, E., Cohen-Nabeiro, A., and Touroult, J. (2016). Metadata on biodiversity: definition and implementation. In *DCMI International Conference on Dublin Core and Metadata Applications: DC 2016 Conference*.

Williams, D. W., Huan, J., and Wang, W. (2007). Graph database indexing using structured graph decomposition. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 976–985. IEEE.

Bilan des journées du GRAAL 2016 et avenir de l'utilisation des graphes en écologie

GRAIL DAYS 2016 results and future of the use of graphs in ecology

Romain David*[1], Anna Cohen Nabeiro[2], Jean-Pierre Féral[1], Aurélie Delavaud[2], Anne-Sophie Archambeau[3], Fanny Arnaud[4], David Auber[5], Nicolas Bailly[6], Loup Bernard[7], Cyrille Blanpain[8], Vincent Breton[9], Denis Couvet[10], Alrick Dias[1], Sophie Gachet[1], Robin Goffaux[2], Karina Gibert[11], Manuel Herrera[12], Dino Ienco[13], Romain Julliard[10], Julien Lecubin[8], Yannick Legre[14], Michelle Leydet[1], Grégoire Loïs[10], Victor Méndez Muñoz[15], Jean-Charles Meunier[16], Isabelle Mougenot[17], Sophie Pamerlon[3], Jean-Claude Raynal[18], Geneviève Romier[9], Dad Roux-Michollet[19], Alison Specht[2], Christian Surace[16], Thierry Tatoni[1]

Contact Author : Romain DAVID (IMBE, CNRS, Aix-Marseille Université), romain.david@imbe.fr (1)
Anna COHEN NABEIRO (ECOSCOPE, FRB) anna.cohen-nabeiro@fondationbiodiversite.fr (2)
Jean-Pierre FÉRAL (IMBE, CNRS, Aix-Marseille Université), jean-pierre.feral@imbe.fr (1)
Aurélie DELAVAUD (ECOSCOPE, FRB) aurelie.delavaud@fondationbiodiversite.fr (2)
Anne-Sophie ARCHAMBEAU (GBIF France) archambeau@gbif.fr (3)
Fanny ARNAUD (ENS Lyon, OHM Vallée du Rhône) fanny.arnaud@ens-lyon.fr (4)
David AUBER (LABRI, Bordeaux) david.auber@labri.fr (5)
Nicolas BAILLY (HCMR/IMBBC and MedOBIS, Heraklion, Greece) nbailly@hcmr.gr (6)
Loup BERNARD (ARCHIMEDE-UMR 7044, Université de Strasbourg/CNRS) loup.bernard@unistra.fr (7)
Cyrille BLANPAIN (SIP OSU Pytheas, CNRS) blanpain@osupytheas.fr (8)
Vincent BRETON (IdGC – LPC, CNRS, France Grilles), breton@idgrilles.fr (9)
Denis COUVET (Muséum national d'Histoire naturelle), couvet@mnhn.fr (10)
Alrick DIAS (IMBE, CNRS, Aix-Marseille Université), alrick.dias@imbe.fr (1)
Sophie GACHET (IMBE, CNRS, Aix-Marseille Université) sophie.gachet@imbe.fr (1)
Robin GOFFAUX (ECOSCOPE, FRB) robin.goffaux@fondationbiodiversite.fr (2)
Karina GIBERT (Department of Statistics and Operations Research, Universitat Politecnica de Catalunya Barcelona) karina.gibert@upc.edu (11)
Manuel HERRERA FERNANDEZ (EDEn - Dept. of Architecture and Civil Eng., University of Bath-UK) amhf20@bath.ac.uk (12)
Dino IENCO (UMR TETIS, Montpellier) dino.ienco@teledetection.fr (13)
Romain JULLIARD (Vigie-Nature, CESCO - Centre d'Écologie et des Sciences de la Conservation Muséum national d'Histoire naturelle) romain.julliard@mnhn.fr (10)
Julien LECUBIN (SIP OSU Pytheas, CNRS), julien.lecubin@osupytheas.fr (8)
Yannick LEGRE (European Grill Infrastructure) yannick.legre@egi.eu (14)
Michelle LEYDET (IMBE, CNRS, Aix-Marseille Université), michelle.leydet@imbe.fr (1)
Grégoire LOIS (Vigie-Nature, CESCO - Centre d'Écologie et des Sciences de la Conservation Muséum national d'Histoire naturelle) gregoire.lois@mnhn.fr (10)
Victor MENDEZ MUNOZ (Department of Computer Architecture and Operating Systems (CAOS) Area of Computer Architecture and Technology) victor.mendez@uab.es (15)
Jean-Charles MEUNIER (LAM / CeSAM) jean-charles.meunier@lam.fr (16)
Isabelle MOUGENOT (UMR Espace DEV, Montpellier) isabelle.mougenot@univ-montp2.fr (17)
Sophie PAMERLON (GBIF-France) pamerlon@gbif.fr (3)
Jean-Claude RAYNAL (ECCOREV FR3098, CNRS, Aix-Marseille Université), raynal@eccorev.fr (18)
Geneviève ROMIER (IdGC, CNRS, France Grilles) genevieve.romier@idgrilles.fr (9)
Dad ROUX-MICHOLLET (GRAIE, OHM Vallée du Rhône) dad.roux@graie.org (19)
Alison SPECHT (CESAB, FRB) alison.specht@fondationbiodiversite.fr (2)
Christian SURACE (LAM, CNRS, Aix-Marseille Université), christian.surace@lam.fr (16)
Thierry TATONI (IMBE, CNRS, Aix-Marseille Université), thierry.tatoni@imbe.fr (1)

Résumé (900 signes max)

Dans le domaine de la biodiversité, la fréquence des campagnes de collecte de données (missions de terrain, capteurs optiques ou radar, suivi de la qualité des eaux, recensement automatique ou semi-automatique des taxons, etc.) ont permis d'acquérir un volume considérable de données. Les journées du GRAAL organisées en 2016 par le consortium IndexMed ont mis en évidence le potentiel des approches basées sur les graphes pour fouiller et visualiser les données, ainsi que les lacunes en termes de compétences de la communauté des écologues pour adapter et utiliser ces approches. Ce séminaire a notamment permis de montrer que les graphes sont pertinents pour modéliser plus efficacement les systèmes écologiques malgré l'hétérogénéité des contextes. Les participants à ces journées (Informaticiens et écologues - environnementalistes) ont sollicité l'organisation de nouvelles rencontres, et développent de nouvelles collaborations interdisciplinaires.

Summary (1500 max)

Data produced by biodiversity research projects to evaluate and monitor the "Good Environmental Status" have a high potential to be used by several stakeholders involved in environmental management; but new and accessible analyzing tools are needed. However, it remains that specific scientific objectives, organizational logic of projects and collection of information are leading to an unavoidable decentralized data distribution, which may hamper environmental research development. The newly created consortium IndexMEED, who task is to index biodiversity data, makes it possible to build graphs in order to analyze the data and develop new ways for data mining in ecology. This communication presents the 2016 IndexMEED kick-off seminar results and recent actions of the consortium: new approaches make possible the investigation of complex research questions and the emergence of new scientific hypotheses. With one day of plenary sessions and two days of practical workshops, this event was dedicated to the acquisition of practical methods to construct graphs and value data through metadata and "data papers". Recent developments in data mining based on graphs, contributions to environmental research, standard formats and specific protocols used to interconnect databases were exposed. IndexMEED project is now exploring the ability of two scientific communities (ecology *sensu lato* and computer sciences) to work together.

Keywords: interdisciplinarity, data qualification, graph, thesaurus, Decision support tools

Mots clefs : interdisciplinarité, qualification de données, graphes, thesaurus, outil d'aide à la décision

Contexte

La production de données[1] concernant les Systèmes Socio-Ecologiques (SSE)[2] coûte cher, et est peu automatisée. Les longues séries temporelles et/ou les études à large emprise spatiale sont difficiles à mener, dès lors qu'il faut recourir à plusieurs observateurs, et la robustesse de l'observation est plus difficile à obtenir lorsqu'il s'agit de données « d'interprétation » (c'est-à-dire une donnée résultant d'une interprétation, par exemple une reconnaissance d'espèce). La reproductibilité est aujourd'hui souvent difficile à obtenir. La production de données dans chaque discipline est discontinue, peu précise et mal répartie. Pourtant, toutes les variables de ces systèmes interagissent dans le temps et à chaque échelle spatiale (variables biotiques, abiotiques, pressions anthropiques et naturelles, services rendus et ressentis…).

Dans un cadre de production de données multi-sources, les problématiques d'équivalence de systèmes d'observations et d'inter-calibration d'observateurs deviennent cruciales. La meilleure compréhension de SSE complexes demande de développer des approches intégratives pluri- voire transdisciplinaires.

Le défi scientifique d'une meilleure compréhension globale des équilibres des SSE et de leur influence sur la biodiversité doit passer par la construction et le test de méthodes de co-interprétation de ces données hétérogènes (Gimenez O. *et al.*, 2014). La science de la fouille de données apporte de nouvelles perspectives aux recherches disciplinaires sur ces systèmes complexes qui finalement prennent en compte des objets intimement liés (la chimie environnementale, la génomique, la transcriptomique, la protéomique, la métabolomique, l'écologie des peuplements, les systèmes socio-écologiques, l'écologie du paysage… en sont quelques exemples). Pour répondre à ce défi, le consortium IndexMed a été créé il y a trois ans. Par essence pluridisciplinaire, il est animé par l'axe Gestion de la biodiversité et des espaces naturels de l'IMBE (Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale), et soutenu activement par le Pôle national de données biodiversité (ECOSCOPE 2017°[3] et le GBIF[4] (Global Biodiversity Information Facility).

---

[1] D'après The Royal Society, académie scientifique qui regroupe des chercheurs du monde entier, la donnée est un énoncé ou un nombre qualitatif ou quantitatif qui est (ou est supposé être) factuel (The Royal Society, 2012).

[2] Les SSE correspondent à des systèmes intégrés couplant les sociétés et la nature (Liu, J., *et al.*, 2007), ce qui vise finalement à redéfinir les écosystèmes en considérant explicitement l'ensemble des acteurs, en intégrant donc l'homme comme une composante active du système (Lagadeuc, Y. & Chenorkian, R., 2009).

[3] L'infrastructure de recherche "Pôle national de données biodiversité" inter-organismes (ECOSCOPE 2017), animée par la FRB (Fondation pour la Recherche sur la Biodiversité), déploie un ensemble de services pour la recherche et l'expertise afin de contribuer à mieux connaître et comprendre l'état et la dynamique de la biodiversité en s'appuyant sur la complémentarité des observations à différentes échelles et niveaux d'organisation du vivant.

[4] Le GBIF agrège notamment les données d'occurrences et de fréquences d'espèces dans le monde entier (plus de 800 millions d'enregistrements à ce jour).

IndexMed a pour ambition de développer la culture des bases de données et leur utilisation efficace dans le milieu de la recherche en écologie et biodiversité. Ce consortium s'est étendu à plusieurs Unités Mixtes de Recherche (UMR) de disciplines différentes, afin de catalyser son action grâce aux compétences plus développées dans d'autres disciplines (notamment dans celles des STIC - Sciences et Technologies de l'Information et de la Communication - pour l'expertise qualitative de la donnée, et de l'astronomie pour l'expertise en matière de gestion des grosses masses de données). Il répond à des appels à projets dans le domaine des bases de données en écologie en favorisant l'interdisciplinarité et les collaborations avec d'autres entités et instituts. Les projets qui y sont développés doivent s'appuyer sur les différentes démarches nationales et internationales et promouvoir un travail partenarial international. IndexMed sert notamment de relais aux réseaux et démarches en place aux niveaux national et international, et promeut la mise en place de processus de gestion des données conformes aux règles européennes auxquelles les laboratoires de recherche travaillant dans les domaines de l'environnement et des sciences humaines sont et seront de plus en plus soumis (par exemple, la rédaction d'un plan de gestion des données, qui définit ce que les chercheurs feront de leurs données pendant et après le projet, explicitant notamment les moyens mis en œuvre pour partager les données).

Le consortium IndexMed développe et anime une plateforme open source d'indexation des données[5] sur la biodiversité et des paramètres environnementaux ayant un intérêt pour la recherche (David R. *et al.*, 2016). Cette indexation utilise les outils et méthodes préconisés au niveau national (SINP - Système national d'Information sur la Nature et les Paysages, MNHN - Muséum National d'Histoire Naturelle, SPN - Service du Patrimoine Naturel, RBDD Réseau Bases De Données du CNRS) ou internationalement (OBIS, GBIF, LifeWatch, GEOBON, CoL, WoRMS…) et s'appuie sur les catalogues préexistants (IDCNP - Inventaire des Dispositifs de Collecte sur la Nature et les Paysages du SINP, Réseaux d'acteurs de la FRB - Fondation pour la Recherche sur la Biodiversité).

Depuis le séminaire de juin 2016, IndexMed est devenu IndexMEED (Indexing for Mining Ecological and Environmental Data), un nouveau consortium ayant pour objectif d'indexer les données sur la biodiversité, issues de bases de données hétérogènes et distantes, afin de construire des graphes et de les analyser. Ces approches basées sur la fouille de donnée sont à développer en écologie/environnement.

[5] L'index est une identification des données utilisé et entretenu par le système de gestion de base de données pour lui permettre de retrouver rapidement les données via un identifiant unique.

Quelques notions sur les graphes

Un graphe est un ensemble de points que l'on appelle des **nœuds** (sommets en mathématique ou entités en informatique) reliés par des traits (segments en mathématique ou relations en informatique) ou flèches nommées **liens** (ou arrêtes ou arcs). L'ensemble des liens entre les nœuds forme une figure similaire à un réseau (Aggarwal, C. & Wang, H., 2010). La représentation de données sous forme de graphes permet de relier des objets (champs/entité de la base de donnée ou valeurs de ces champs/attributs/relation) ayant des natures différentes (valeurs quantitatives, qualitatives ordonnés ou non ordonnés) ; les attributs contenus dans un second champ décrivant une qualité de l'objet permettent de créer les liens entre ces objets et/ou de les pondérer. Les liens sont matérialisés par des descripteurs, c'est-à-dire des variables ayant plus d'une valeur possible. Les objets ayant le plus de liens en commun sont les plus proches, ceux ayant les liens les plus ténus (c'est à dire le moins de chemins possibles pour les relier entre eux et beaucoup de nœuds intermédiaires) sont les plus éloignés dans la représentation. On peut traiter les champs un à un ou bien en groupe de valeurs pour former - selon la combinaison de leurs valeurs respectives - un motif appelé patron (pattern en anglais). Ces patrons peuvent décrire des objets et/ou des liens et/ou des contextes. Les champs de « contextes », sont ensuite utilisés pour différentier les nœuds entre eux (couleur, forme, grosseur des nœuds). Ils ne participent pas à la topologie du graphe (c'est-à-dire à sa forme et ses propriétés). Les motifs ainsi projetés dans le graphe peuvent être (i) dispersés, auquel cas les liens qui organisent le graphe ne sont pas liés aux éléments de contexte ; ou bien (ii) regroupés dans une ou plusieurs parties du graphe auquel cas il existe un lien entre la façon dont les nœuds sont organisés et un ou plusieurs contextes.

Analyser les regroupements de nœuds pour aller un peu plus loin :

Le clustering (classification non supervisée en français, mais c'est le terme anglais qui est le plus usité à la place de "classification", "groupe", ou "regroupement") consiste à regrouper des éléments. Cette agrégation est un élément-clé pour l'analyse de grands graphes. Une fois les groupes de nœuds obtenus, on peut réappliquer l'opération pour obtenir un clustering hiérarchique (basé sur une autre variable par exemple). Cette décomposition hiérarchique (ou multi-échelle) permet de modifier la complexité des algorithmes de fouille, de faciliter l'exploration des données, et de proposer une visualisation paramétrable : on parle aussi de navigation multi-échelle (Auber D. *et al.*, 2014 ; Lambert, A. *et al.*, 2013). Les descripteurs quantitatifs sont en général transformés en classes de valeurs. L'analyse des fréquences relatives des "motifs" et des redondances entre "plus proches voisins" par rapport à leur fréquence dans tout ou partie du graphe montre l'importance des corrélations entre certains contextes et des clusters du graphe. La **significativité de ces motifs peut ensuite être testée par des méthodes statistiques spécifiques** analysant les qualités des clusters de graphes. Dans des graphes plus complexes où le nombre de combinaisons et de liens peut croître exponentiellement, l'étude de la corrélation entre fréquence de contextes et "clusters" de nœuds peut demander de paralléliser les calculs nécessaires à une investigation des parcours possibles. Selon la question scientifique sous-jacente aux objets représentés par un graphe, certains éléments dans les liens ou les nœuds peuvent être ignorés ou simplifiés.

Utilisation en écologie/environnement :

Les graphes permettent de représenter tout ou partie d'un système observé (par exemple, un ensemble de sites plus ou moins ressemblants en terme de composition d'espèces ou les nœuds sont des sites, les liens sont les observations d'un taxon commune à differents sites et les clusters de nœuds correspondent aux sites les plus identiques et donc ayant le plus de liens entre eux. De la même manière, on peut représenter des individus reliés par la fréquence de leurs contacts, des groupes de taxons reliés par des traits, des groupes de personnes reliés par des réponses d'enquêtes sociologiques…) en intégrant des données de contextes de format différents (température, altitude, âges des individus, ensoleillement représentés par la grosseur ou la forme du noeud…). Ils permettent de mélanger des objets (graphes bipartites ou tripartites) ou de représenter de nouveaux objets complexes en combinant les valeurs de différents champs. Ils peuvent aussi être utilisés pour étudier le système d'observation ainsi que les efforts de prospection, ou pour avoir une approche visuelle de la répartition des compétences utilisées dans un projet ou évaluer un système d'information.

Les journées du GRAAL : compte rendu et perspectives

Le séminaire s'est déroulé sur trois journées à l'Université d'Aix-Marseille et au Centre de Synthèse et d'Analyse sur la Biodiversité (CESAB) à Aix-en-Provence.

Un séminaire de démonstration sur le potentiel des graphes en écologie

La première journée, plénière du séminaire, a permis de faire le point sur les perspectives possibles en terme de visualisation sous forme de graphes, l'utilisation des graphes à des fins d'analyse et de fouille de données, et les cas d'études disponibles au sein du consortium pour faire évoluer ces approches (David R. *et al.*, 2016). Cette journée a aussi été l'occasion de présenter les prérequis en terme d'architecture et de compétences à organiser pour développer l'utilisation des graphes sur de grands jeux de données hétérogènes et multi sources.

L'architecture d'indexation a elle-même été définie lors du séminaire « Design your infrastructure » organisé par European Grid Infrastructure - http://www.egi.eu/ - en avril 2016 (David R. *et al.*, 2016). Cette architecture a l'avantage de pouvoir soit laisser les jeux de données dans l'entrepôt de données officiel (ce sont alors des flux paramétrables qui sont interrogés), soit de se baser sur des imports (aux formats CSV, XML, JSON…). Les graphes sont construits via un prototype de visualisation de graphes en ligne sur internet à partir d'informations agrégées à partir de répertoires distants (David R. *et al.*, 2015).

Ce prototype de visualisation et de fouille de ces données hétérogènes est en cours de développement. Son objectif est d'utiliser les modèles sous forme de graphe pour proposer de nouvelles questions scientifiques, issues du croisement de données de différentes origines. Les démonstrations actuelles permettent de visualiser plusieurs milliers de nœuds reliés par plus de dix milles liens. (Figure 1.).

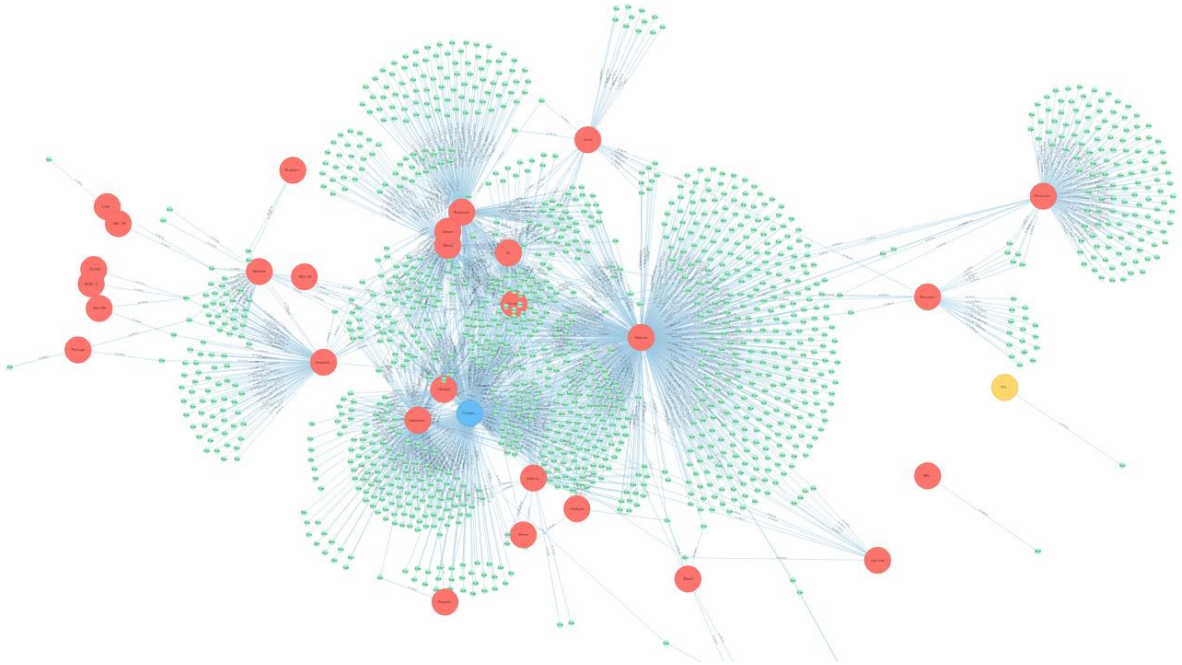


Fig. 1 Graphe bipartite issu de l'interface IndexMEED, utilisant les données exportées d'une sélection géographique sur ArkeoGIS[6], représentant 1 492 sites archéologiques et 4 950 liens matérialisant les types d'objets trouvés sur ces sites. Les sites sont représentés par les petits nœuds verts. Ces points sont proches (et forment donc des clusters) si les sites archéologiques ont les mêmes topologies/ patrons de descripteurs, matérialisés par des nœuds rouges. Une sélection a été faite sur deux valeurs de descripteurs : "Bleu" pour les types d'objet de niveau 1 qualifiés de "céramique" (le site contient donc au moins un objet de type céramique s'il est lié à ce nœud), "Jaune" : le nœud correspond à un objet daté de la période exactement égale à "-900 -à -726".

Des défis à venir

Les futurs défis concernant le développement de cette interface sont le passage à l'échelle (100 000 nœuds à 1 000 000 de nœuds ou plus) et l'augmentation des services aux usagers. La prochaine étape consistera à développer les cas d'utilisation du prototype et à améliorer les fonctionnalités selon les retours utilisateurs. La temporalisation de ces graphes sous forme de graphes évolutifs, devra permettre de générer des scénarios dépendants de contextes mesurés et utilisables dans les domaines de l'aide à la décision.

Le traitement des problématiques de la qualité et de l'interopérabilité des données nécessite encore de prévoir de nombreux échanges entre les différents participants. Les verrous scientifiques à lever dans un projet de recherche interdisciplinaire dans le cadre

[6] "ArkeoGIS » est un projet porté par des archéologues, permettant d'agréger des données issues de bases aussi bien archéologiques que paléo-environnementales. Les données unifiées ont permis suite aux premières journées IndexMEED de produire le graphe de la figure 1.

d'IndexMEED ont été identifiés par les experts du domaine des STIC, et concernent notamment i) les volumes croissants de données dus à l'augmentation des fréquences et de la densité d'acquisition des observations (liées notamment au développement des méthodes de reconnaissance automatique et au déploiement d'outils d'acquisition moins onéreux), ii) la diversification des objets et des descripteurs d'objet intégrés dans les graphes, iii) la normalisation des descripteurs de la donnée et les méthodes permettant d'intégrer les différents niveaux de qualité des données.

Une fois les méthodes développées, les bases de données de chaque participant (MNHN[7], ECOSCOPE[8], GBIF-France[9] et IMBE[10] notamment), représentant parfois plusieurs Téraoctets de données, sont utilisées comme cas d'étude. Cet exercice permet de rendre les approches analytiques plus génériques et capables de s'adapter à l'évolution des données (volume, complexité ou fréquences accrues liée à l'automatisation de l'acquisition via des drones ou des systèmes d'observation permanents par exemple).

Un effort est porté sur les fonctions d'analyse de la qualité des données et d'exploration des possibilités d'interopération entre bases de données de différents champs disciplinaires (écologie, sciences humaines et sociales, économie). De nouvelles qualifications (apportées par chaque nouvel usager) doivent permettre de lier des objets décrits par des attributs hétérogènes mais de même nature et de proposer de nouveaux concepts permettant des approches plus intégratives. Le prototype développé est un élément majeur du projet porté par le consortium IndexMEED dont un des objectifs est de faciliter l'utilisation les graphes et hypergraphes paramétrables basés sur ces données hétérogènes et distantes (qui peuvent concerner les molécules, les écosystèmes, en passant par les traits de vie, jusqu'aux paysages et aux interactions Homme-milieux) pour alimenter des systèmes d'aide à la décision. Pour atteindre cet objectif, il est prévu d'adapter une bibliothèque d'algorithmes d'analyse de grands graphes déjà utilisée par d'autres disciplines et développée en propre par les chercheurs intéressés dans le domaine des STIC.

Mieux connaître le potentiel des données pour mieux inter-opérer

Les métadonnées[11] favorisent la réutilisation des données et les collaborations scientifiques à des fins de recherche ou d'expertise. Les métadonnées peuvent être importées et exportées d'un système d'information à l'autre grâce à l'utilisation de standards, limitant ainsi les

[7] https://www.mnhn.fr
[8] http://www.fondationbiodiversite.fr/fr/recherche/programmes-frb/ecoscope.html
[9] www.gbif.fr
[10] https://www.imbe.fr
[11] Les métadonnées sont des informations sur les données et les jeux de données, permettant de les décrire et les cataloguer, ainsi que les dispositifs et structures dans le cadre desquels les données sont produites.

efforts des chercheurs pour valoriser leurs travaux aux niveaux national et international, voire à travers des "data papers[12]". L'amélioration des synergies entre bases de données (qui induit une amélioration des synergies entre acteurs) s'appuie entre autres sur l'infrastructure de recherche « Pôle national de données biodiversité » qui contribue à l'objectif global de consolidation des connaissances pour améliorer la compréhension de l'état et de la dynamique de la biodiversité. Celle-ci déploie un portail de métadonnées avec les observatoires de recherche sur la biodiversité qui travaillent à différents niveaux d'organisation, de l'infra-spécifique aux écosystèmes, des ressources génétiques au fonctionnement des milieux. L'objectif premier de ce portail est de porter à connaissance les jeux de données existants, leurs contextes d'acquisition, la qualité ainsi que les conditions d'accès et d'utilisation. Dans un contexte d'hétérogénéité et de dispersion des sources de données, ce portail catalogue les jeux de données et les ressources biologiques sous des formats standards (ISO/INSPIRE) ou en usage (EML, NCD) parmi les chercheurs en écologie. Les métadonnées permettent également d'analyser le paysage de la recherche - ses forces et lacunes - et de contribuer aux initiatives globales, comme le développement des Variables Essentielles de Biodiversité (EBV), proposées par GEO BON (Group on Earth Observation – Biodiversity Observation Network). Ces bases de métadonnées, dont l'exploitation sous forme de graphes a été récemment testée (Muñoz *et al.*, 2016), permettront d'estimer les opportunités d'interopération entre systèmes d'information (Loi pour une République numérique, 2016).

Développer la culture des données et leur « réutilisabilité »

D'une manière générale les données, lorsqu'elles sont rassemblées, sont souvent au mieux « empilées » et n'utilisent pas les mêmes standards. Les typologies de champs ne sont la plupart du temps pas uniformisées lorsque ceux-ci contiennent le même type d'information (géographiques, temporelles, noms d'auteurs, objets, constructions humaines…) ; cependant certains référentiels sont petit à petit institutionnalisés. De facto, les correspondances entre les données générées par ces études de différentes disciplines, portant cependant sur les mêmes territoires, sont encore peu aisées à réaliser, surtout sur un temps long. Améliorer le potentiel de ces données (et donc ainsi leur valeur) nécessite de mettre en place une stratégie de curation des données, d'organiser la gestion de leur cycle de vie et leur accès (via des plans de gestion des données notamment), selon les grands principes FAIR (Findable, Accessible, Interoperable, and Reusable).

[12] Le data paper est une publication scientifique dont le but principal est de décrire un jeu de données ou un ensemble de jeux de données.

Des ateliers se sont également tenus lors des journées du GRAAL, au cours desquels des outils pour gérer et publier ses métadonnées ont été présentés (« Présentation du portail d'ECOSCOPE » et « Comment générer son data paper avec l'IPT du GBIF »), ont permis de montrer l'importance i) d'un bon niveau de qualité des métadonnées ii) de l'interopérabilité des portails de métadonnées iii) des plans de gestion des données pourtant quasiment systématiquement absents « pour le moins » dans les systèmes présentés lors des journées du GRAAL.

Cette année, l'utilisation de la théorie des graphes en environnement se précise

Les journées du GRAAL et les propositions qui en ont découlé en préfigurent d'autres ; elles s'insèrent dans un projet à long terme véritablement interdisciplinaire qui organise les interactions entre des compétences de personnels d'instituts spécialisés en Mathématiques et Informatique (INS2I[13], INRIA[14]…) et de l'INEE[15]. L'implication réelle et sur le long terme de scientifiques des STIC et des chercheurs en sciences de l'environnement dans cette démarche interdisciplinaire permet de résoudre des questions scientifiques sur les graphes qui sont propres à chaque discipline. La mise en place de ces méthodes de travail interdisciplinaires sur des bases de données environnementales est encore dans un état de développement préliminaire, mais le travail annoncé devrait avoir des effets vertueux car respectant les très reconnus mais si mal appliqués principes FAIR (Findable, Accessible, Interoperable, Reusable), principes reconnus par les instances nationales et internationales (comme au séminaire ministériel de l'Alliance nationale de recherche pour l'environnement – AllEnvi en 2014) comme incontournables pour les projets d'observatoire, surtout dans une dimension telle que les relations Homme-milieux. En effet, ceux-ci exploitent les bases de métadonnées présentes dans les répertoires investis ces dernières années par des acteurs ayant des périmètres thématiques différents (dont le Pôle national de données de biodiversité – ECOSCOPE 2017, le SINP ou le GBIF). Dans ce contexte, plusieurs actions ont été lancées en 2017 pour développer les approches de graphes en écologie et sciences de l'environnement.

Graminées, une action soutenue par le Groupement De Recherche (GDR) MaDICS visant à rapprocher deux communautés

[13] https://www.inria.fr/
[14] http://www.cnrs.fr/ins2i/
[15] http://www.cnrs.fr/inee

Le GDR MaDICS encourage les activités d'animation interdisciplinaires sur les masses de données scientifiques. Dans ce cadre, une action au nom de « GRAMINEES[16] » ayant pour objectif de mettre en place une dynamique d'échange entre des experts en écologie/biodiversité et des experts du domaine des STIC a été développée. Regrouper des représentants des deux champs disciplinaires permet notamment de mieux formaliser des besoins en terme d'analyses complexes de la part de la communauté écologie/biodiversité et de stimuler la recherche en STIC afin de proposer des solutions plus adéquates pour l'analyse et la gestion des données écologiques dans le contexte du Big Data (prise en compte de la dimension temporelle et spatiale, données multi-échelles et hétérogènes).

En ce qui concerne les techniques et approches STIC étudiées et développées, la recherche est prioritairement dirigée vers des techniques de gestion et d'analyse de graphes qui puissent prendre en compte la complexité des données hétérogènes et notamment passer à l'échelle sur des jeux de données volumineux sans détériorer la qualité des résultats obtenus.

L'atelier de travail organisé lors des rencontres MaDICS le 23 Juin 2017 à Marseille a permis la réalisation d'un premier « mind-mapping »[17] autour des compétences en écologie et informatique des participants. La définition de cette carte de compétences, qui doit être enrichie, permettra de détecter un ensemble de laboratoires (et les personnels associés) pour envisager des projets au plan international sur la thématique du Big Data, liés aux données de l'écologie et de la biodiversité.

Le projet GRAINE, une application concrète des travaux du consortium IndexMEED appliquée aux problématiques « Homme-milieux »

L'objectif de ce projet, retenu par le LabEx Dispositif de Recherche Interdisciplinaire sur les Interactions Hommes-Milieux (DRIIHM), est de construire une méthode de visualisation de données hétérogènes basée sur les graphes dans le cadre de trois « cas d'études » issus d'Observatoires Hommes-Milieux[18] (OHM), en utilisant le prototype développé par le consortium IndexMEED. Pour chaque OHM, cette méthode de visualisation sera appliquée i) au système d'observation, c'est-à-dire à l'organisation des métadonnées et données et à la manière dont celles-ci diffèrent (type, qualité, contenu…), et ii) au système observé ou à une partie de celui-ci (ce résultat sera fonction de la consistance des données dans chacun des OHM). Une pré-étude de faisabilité sera réalisée sur trois OHM (Bassin minier de Provence,

[16] http:// www.madics.fr/actions/actions-en-cours/graminees/

[17] Un mind-map est un graphique représentant des idées, des tâches, des mots, des concepts… liés entre eux autour d'un sujet central.

[18] Les Observatoires Hommes-Milieux (OHM) s'attachent à l'étude des socio-écosystèmes fortement anthropisés, systèmes complexes qui nécessitent une convergence interdisciplinaire pour être étudiés et interprétés par une démarche d'écologie globale.

Vallée du Rhône et Littoral méditerranéen). L'ambition est de mettre en évidence les particularités et points communs, points forts et axes d'amélioration de chaque système observé en considérant tout type de données (biodiversité, socio-écologie, économie, économétrie de l'environnement…).

Appliquer cette approche aux données produites dans le cadre des OHM permettra d'élaborer une méthode d'intégration puis de fouille de données dans le cas de données très hétérogènes liant environnement et société, ce qui constitue un des objectifs majeurs du DRIIHM. Ce travail permettra d'évaluer le potentiel des représentations par les graphes de données hétérogènes issues ou intéressant les OHM, en réalisant des ateliers de curation et d'amélioration de la qualité des données et des systèmes d'informations, de manière à produire ces premières représentations.

Les premiers exemples de graphes concerneront les données actuellement accessibles dans chacun des OHMs et produites dans les communautés de recherche ou dans le cadre de dispositifs régaliens en lien avec la problématique de l'OHM, L'objectif est de produire une carte des données existantes et de leur utilisabilité, et de les mettre en relation (sous forme de graphe) avec un système d'information « idéal » pour chacun des OHMs. Le but fixé pour la première année est d'obtenir une carte des données généralisable à tous les OHM, permettant de mettre en exergue les descripteurs de données soit équivalents, soit traitant des mêmes objets (cartographie du territoire, qualité de l'eau, biodiversité, liens entre homme et milieu…) mais avec un vocabulaire différent (ce qui arrive fréquemment dans les projets interdisciplinaires). Il permettra aussi de mettre en évidence les descripteurs ou valeurs de descripteurs polysémiques, qui induisent une ambiguïté dans l'analyse intégrée de plusieurs lots de données. La méthode mise en place et testée sur les trois OHM ciblés sera documentée pour être transposable aux autres OHM car **toutes les données produites ou utilisables dans le cadre des OHMs peuvent être visualisées sous forme de graphe**.

Deux ateliers préparatoires « cas d'études » puis un séminaire centré sur les algorithmes

Les questions scientifiques véritablement novatrices, pour l'instant peu explorées, doivent émerger des données (cet état de fait a été une des conclusions majeures lors des deux derniers séminaires IndexMEED). Le consortium priorise les cas d'études présentant les meilleurs niveaux d'accessibilité et d'utilisabilité des données (par exemple, données ArkeoGIS, GBIF, MNHN, ECOSCOPE ou IMBE), et prépare les outils d'audit de la donnée permettant d'organiser les étapes nécessaires au respect de ces principes FAIR pour les cas d'études. Cet objectif d'amélioration qualitative de jeux de données s'est traduit par

l’organisation de deux ateliers thématiques qui permettront de faire une démonstration par l’exemple, basée sur plusieurs cas d’études précités. Le premier atelier a porté sur la curation de données, en vue de les intégrer dans des interfaces de manipulation de graphes. Les ateliers de curation, organisés à Aix-en-Provence, Marseille et à Paris, ont permis d’identifier des verrous actuels et futurs concernant les données et l’organisation des compétences, pour permettre une amélioration significative du référencement, de l’accessibilité, de l’interopérabilité et de la réutilisabilité des données selon les principes FAIR, et ainsi définir les moyens nécessaires au respect de la nouvelle réglementation sur l’accessibilité numérique des données et travaux de la recherche (Loi pour une République numérique, ; Lambert A. *et al.*, 2013). Un guide étape par étape sur la curation des jeux de données est en cours de réalisation sur la base des conclusions des ateliers (mise en forme des données pour l'import, homogénéisation des descripteurs...).

Les premiers ateliers "Visualisation de données par les graphes pour les sciences de l'environnement et l'écologie" se sont déroulés à Paris les 15, 16 et 17 novembre 2017. Ils ont permis la représentation de données dans l'espace de huit cas d'étude aux thématiques et bases de données diverses (archéologie/paléo-environnement, écologie marine benthique et hauturière, palynologie, biologie des populations et métadonnées du portail ECOSCOPE). De premières observations des données sous forme de graphes ont permis de visualiser les liens entre des entités (taxons, cadrats photos, sites de prospection, carottages, individus de communautés…), de faire varier le poids relatif des descripteurs servant de lien entre ces entités, afin de faire émerger des clusters et de nouvelles hypothèses expliquant la structuration mise en évidence. Ces représentations permettent aussi d'identifier des erreurs ou manques dans le jeu de données. A titre d’exemple, la figure 2 montre 2 graphes produits lors de ces ateliers (Figure 2) reliant des photos de plaquettes de récifs artificiels.

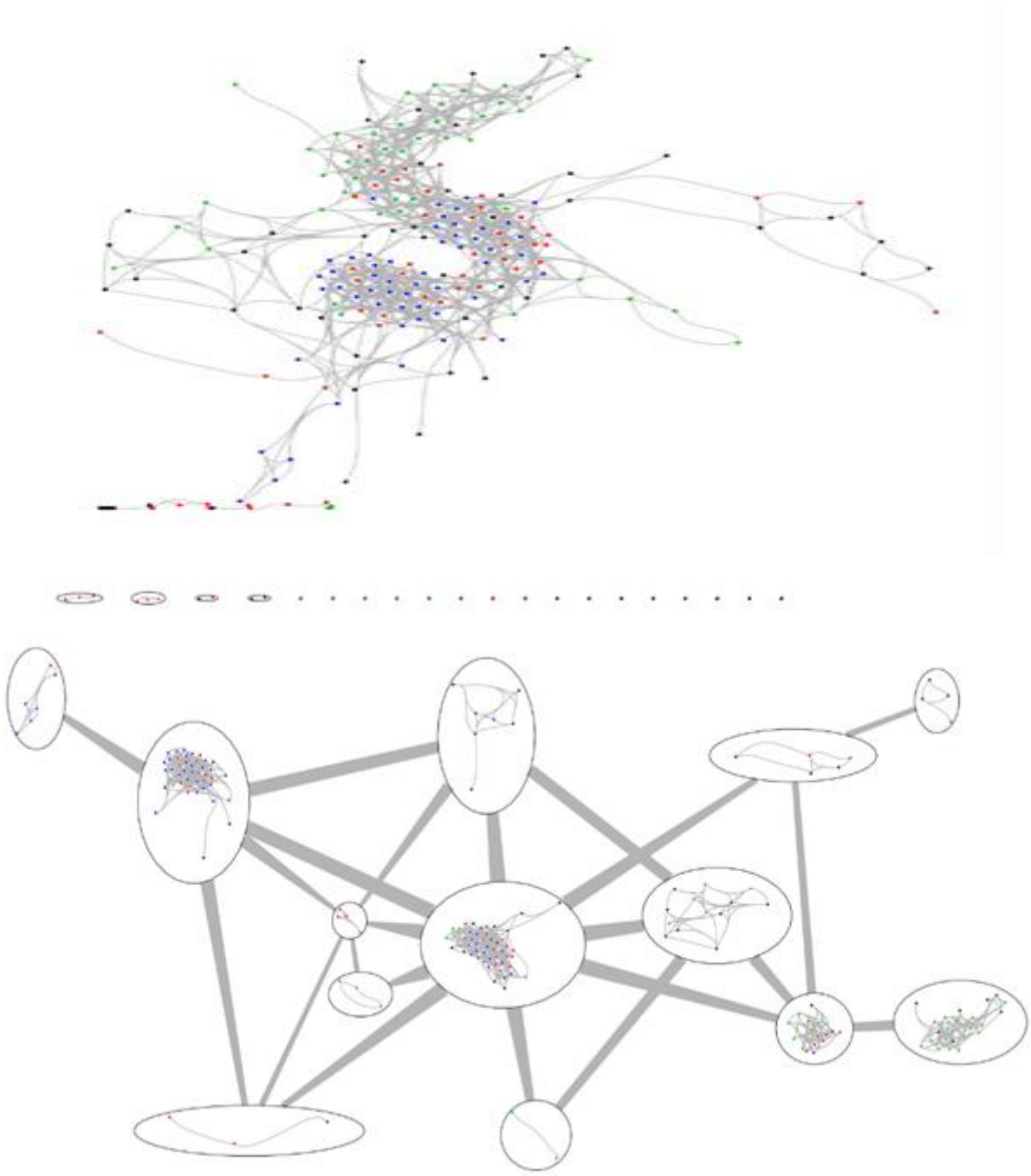


Fig 2 : Ce graphe a été construit avec le logiciel Tulip[19] à partir des fréquences relatives de taxons recueillis sur des éléments de récifs artificiels en PVC conçus par la NOAA[20] (ARMS : Artificial Reef Monitoring System) répartis à différents endroits du récif (et donc exposés différemment aux facteurs de contextes comme la lumière, le courant…). Dans cette visualisation préliminaire, les fréquences relatives ont été déterminées à partir de photos. 6 facettes de plaquettes ont été analysées dans 3 récifs, pour 3 sites pour chacune des 4 mers régionales (Golfe de Gascogne, méditerranée nord occidentale, Adriatique et Mer Rouge). Les nœuds de ce graphe représentent des facettes de plaques en PVC, et les liens relient les

[19] http://tulip.labri.fr/TulipDrupal/
[20] http://www.noaa.gov/

facettes pour chaque pourcentage d'un taxon en commun avec une autre facette. Chaque couleur représente une mer régionale différente. Dans la première figure, on visualise les premiers clusters, ou les plaques avec les compositions en taxons les plus similaires sont les plus proches, et les sites des différentes mers sont colorés de manière différentes avec des regroupements plus ou moins hétérogènes (Fig 2 en haut). La deuxième figure a été construite en sélectionnant uniquement des fréquences relatives en commun supérieures à 10%. Elle permet de mettre en exergue les regroupements en fonction de combinaisons de valeurs de paramètres environnementaux (en faisant varier la forme, la taille et/ou la couleur des nœuds en fonction des classes de valeurs des différents paramètres) avec un autre algorithme de visualisation mettant en évidence les "clusters" (Fig 2 en bas). Lorsque les formes, tailles et/ou couleurs des nœuds se regroupent dans un cluster, on visualise une relation entre certains regroupements et des groupes de valeurs de variables de contextes. Il faut alors tester la significativité de ces corrélations grâce à des méthodes statistiques propres aux graphes.
Perspectives : Ces représentations incluront prochainement des analyses méta génomique (des clusters de gènes en tant que nœuds ou liens).
Pour en savoir plus : programme DEVOTES : DEVelopment Of innovative Tools for understanding marine biodiversity and assessing good Environmental Status http://www.devotes-project.eu/

Ces cas d'étude, une fois les résultats publiés, seront autant d'exemples de l'utilisation de la théorie des graphes en écologie, et feront l'objet d'une diffusion dans les communautés de recherche concernées.

Ce travail doit servir de base au développement d'un réseau au niveau européen, qui permettra de continuer à rapprocher les communautés des écologues et des STIC autour des challenges proposés par la fouille de données en écologie et environnement basée sur la théorie des graphes. Le cercle vertueux de la curation et de la visualisation de données promet de les rendre réutilisables et intégrables dans divers modèles, dans un objectif de compréhension et d'analyse des systèmes socio-écologiques.

Le séminaire de restitution des travaux engagés s'intitule « Sciences and Algorithms around Graphs in Environment and Societies (Journées des SAGES) ». Il se déroulera **à Paris, au siège du CNRS**, (au début du printemps 2018). Il est gratuit et ouvert à tous, mais l'inscription est obligatoire : https://indexmeed2017.sciencesconf.org/

Remerciements

Remerciements aux intervenants STIC (D. Auber, R. Bourqui, A. Dias, J. Lecubin) et aux CESAB, ECOSCOPE, FRB, GBIF, IMBE, LAM, Labex DRIIHM (OHM Bassin Minier de Provence, OHM Vallée du Rhône, OHM Littoral méditerranéen), Fédération ECCOREV FR 3098, OSU Pythéas, et LabEx OT Med pour leur soutien humain et financier à l'organisation des journées du GRAAL en 2016.

Références bibliographiques

Aggarwal, C. & Wang, H., 2010. Managing and Mining Graph Data, Springer, 1st Edition, XXII, 600 p.

Auber, D., Archambault, D., Bourqui, R., Delest, M., Dubois, J., Pinaud, B., Lambert, A., Mary, P., Mathiaut, M., Melançon, G., 2014. Tulip III, *Encyclopedia of Social Network Analysis and Mining*, pp 2216-2240, doi: 10.1007/978-1-4614-6170-8_315.

David, R., Féral, J.-P., Tatoni, T., 2016. Architecture, concepts et services d'un système d'indexation de données distribuées pour l'observation à large échelle en écologie marine dans le cadre du consortium IndexMEED (Indexing for Mining Ecological and Environmental Data), *36ème conférence sur la Gestion de Données - Principes, Technologies et Applications (BDA 2016),* 15-18 novembre 2016, Poitiers.

David, R., Féral, J.-P., Archambeau, A.-S., Bailly, N., Blanpain, C., Breton, V., De Jode, A., Delavaud, A., Dias, A., Gachet, S., Guillemain, D., Lecubin, J., Romier, G., Surace, C., Thierry de Ville d'Avray, L., Arvanitidis, C., Chenuil, A., Çinar, M.-E., Koutsoubas, D., Sartoretto, S., Tatoni, T., July 2016. IndexMed projects: new tools using the CIGESMED DataBase on Coralligenous for indexing, visualizing and data mining based on graphs, in : Sauvage S, Sánchez-Pérez J-M., Rizzoli, AE (Eds.), *Proceedings of the 8th International Congress on Environmental Modelling and Software, Environmental modelling and software for supporting a sustainable future*, Vol. 3, pp.656-665, Toulouse, France. ISBN : 978-88-9035-745-9.

David, R., Bernard, L., Blanpain, C., Dias, A., Féral, J.-P., Gachet, S., Lecubin, J., Surace, C., Tatoni, T., 2016. Visualisation de données sous forme de graphes en archéologie : rencontre opérationnelle des archéologues d'ArkeoGIS et des écologues d'IndexMed, in Costa, L., Giligny, F., Djindjian, F. (Eds.), *Actes des 5èmes Journées d'informatique et archéologie de Paris, JIAP 2016*, Paris, 7-10 juin 2016, Archéologies numériques, volume 1, à paraître 2017 (https://www.openscience.fr/Archeologies-numeriques).

David, R., Féral, J.-P., Blanpain, C., Diaconu, C., Dias, A., Gachet, S., Gibert, K., Lecubin, J., Surace, C., 2015. A first prototype for indexing, visualizing and mining heterogeneous data in Mediterranean ecology within the IndexMed consortium interdisciplinary framework, in *SITIS 2015, 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Bangkok, Thailand, pp. 232-239, nov. 2015, doi: 10.1109/SITIS.2015.119.

Gimenez, O., Buckland, S.T., Morgan, B.J.T, Bez, N., Bertrand, S., Choquet, R., Dray, S., Etienne, M.P., Fewster, R., Gosselin, F., Mérigot, B., Monestiez, P., Morales, J., Mortier, F., Munoz, F., Ovaskainen, O., Pavoin, S., Pradel, R., Schurr, F.M., Thomas, L., Thuiller, W., Trenkel, V., de Valpine, P., Rexstad, E., 2014. Statistical ecology comes of age. *Biology Letters,*10: 20140698.

Lagadeuc, Y. & Chenorkian, R., 2009. Les systèmes socio-écologiques : vers une approche spatiale et temporelle. *Natures Sciences Sociétés*, vol. 17,(2), 194-196. (http://www.cairn.info/revue-natures-sciences-societes-2009-2-page-194.htm).

Lambert, A., Bourqui, R., Auber, D., 2013. Graph Visualization For Geography, in Rozenblat C., Melançon G. (eds), *Methods for Multilevel Analysis and Visualisation of Geographical Networks. Methodos Series (Methodological Prospects in the Social Sciences)*, vol 11. Springer, Dordrecht, doi :10.1007/978-94-007-6677-8_6.

LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique Art. L. 533-4 https://www.legifrance.gouv.fr/eli/loi/2016/10/7/ECFI1524250L/jo#JORFARTI000033202841 .

Liu, J., Dietz, T., Carpenter, S.R., Alberti, M., Folke, C., Moran, E., Pell, A.N., Deadman, P., Kratz, T., Lubchenco, J., Ostrom, E., Ouyang, Z., Provencher, W., Redman, C.L., Schneider, S.H., Taylor,W.W., 2007. Complexity of coupled human and natural systems, Science, 317, 5844, 1513-1516.

Muñoz, V., Cohen Nabeiro, A., Couvet, D., David, R., Delavaud, A., Feral, J.-P., Ivars, V.J., Nonell-Canals, A., Senar, M.A. and Tatoni, T., 2016. Analysis on the graph techniques for data-mining and visualization of heterogeneous biodiversity data sets, *Complexis conference*.

The Royal Society, 2012. Science as an open enterprise. Summary report. The Royal Society Science Policy Centre, London

Lessons from photo analyses of Autonomous Reef Monitoring Structures, as tools to detect geographical, spatial, and environmental effects

Romain DAVID[1*], Maria C. UYARRA[2], Susana CARVALHO[3], Holger ANLAUF[3], Angel BORJA[2], Abigail E. CAHILL[1,7], Laura CARUGATI[4], Roberto DANOVARO[4,5], Aurélien DE JODE[1], Jean-Pierre FERAL[1], Dorian GUILLEMAIN[1], Marco LO MARTIRE[4, 6], Laure THIERRY DE VILLE D'AVRAY [1], John K. PEARMAN[3], Anne CHENUIL[1]

(1) Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE), CNRS, Aix Marseille Université, IRD, and Université d'Avignon, Station Marine d'Endoume, Chemin de la Batterie des Lions, 13007 Marseille, France.
(2) AZTI, Marine Research Division, Herrera Kaia, Portualdea s/n, 20100 Pasaia, Spain
(3) King Abdullah University of Science and Technology (KAUST), Red Sea ReSearch Center (RSRC), Biological and Environmental Sciences & Engineering Division (BESE), Thuwal 23955-6900 Saudi Arabia
(4) Department Life and Environmental Sciences, Polytechnic University of Marche, Ancona, Italy.
(5) Stazione Zoologica Anton Dohrn, Naples, Italy.
(6) Ecoreach ltd, spin off Polytechnic University of Marche, Ancona, Italy
(7) Biology Department, Albion College, Albion MI 49224 USA

* corresponding author

Email: romain.david@imbe.fr

Acknowledgements:

This manuscript is a result of DEVOTES (DEVelopment Of innovative Tools for understanding marine biodiversity and assessing good Environmental Status) project, funded by the European Union under the 7th Framework Program, 'The Ocean of Tomorrow' Theme (grant agreement no. 308392), www.devotes-project.eu. Maria C. Uyarra was partially funded through the Spanish program for Talent and Employability in R+D+I "Torres Quevedo." This work is a contribution to the Labex OT-Med (no. ANR-11-LABX-0061) funded by the French Government 'Investissements d'Avenir' programme of the French National Research Agency (ANR) through the A*MIDEX project (no. ANR-11-IDEX-0001-02). We acknowledge all the field helpers and students who participated in data collection in the field and in the lab, as well as in data management: Frédéric Zuberer, Christian

Marschal, Sandrine Chenesseau, Anne Haguenauer, Caroline Rocher and Marjorie Selva (service ‘Plongée’ from the OSU Pytheas and service ‘Biologie moléculaire' from the IMBE).

Abstract

In situ monitoring of community composition provides crucial and irreplaceable information to characterize the ecological status of an area but is particularly complex for hard substrata in marine environments. Our goal was to assess the validity of Autonomous Reef Monitoring Structures (ARMS) as monitoring tools for hard bottoms by analyzing the patterns of change in biodiversity across a wide range of geographic regions and environments. We investigated their ability to detect differences in community composition related not only to geography but also to local environmental factors. We deployed ARMS in the Bay of Biscay (northeast Atlantic), the northwest Mediterranean, the Adriatic and the Red Sea at depths ranging between 7 and 17 m. Thirty-six ARMS were distributed in triplicates at three sites with distinct environmental conditions in each of the four seas. After 12-16 months, they were recovered, and their community composition was analyzed by photo-analyses, via broad taxonomic identification of random points drawn from photographs. Six plate faces (both the up and downside of three plates) were analyzed for each ARMS. Overall, ecological community analyses revealed a highly significant effect of the sea, of the site (within seas), and of plate-face on community composition. The distinct plate-faces of ARMS thus can be considered distinct micro-habitats and provide pseudo-replicates that increase the power of statistical analyses. In each sea region taken individually, there was also a highly significant effect of site and ARMS face on community composition. These strong effects were obtained despite taxonomic categories had been fused at high taxonomic ranks (class or phylum) to ensure comparability among biogeographic provinces. Thus, photo-analysis of communities colonizing ARMS appears a promising monitoring tool in the four sea regions, able to detect differences at geographic as well as much smaller spatial scales. Keeping three replicate ARMS per site appeared necessary; furthermore, assessing environmental effects on ARMS community composition would require the analysis of more numerous sites within a region than in this pilot study.

Keywords: hard substrata; biodiversity; settlement; colonization; monitoring; artificial reefs; scientific diving

1-Introduction

In an era when anthropogenic activities are having an impact on the marine environment, there is a desire to minimize these impacts and to improve the environmental status of marine habitats. To achieve this, over the last decades, several European directives have been implemented such as the Water Framework Directive and the Habitats Directive. More recently, the Marine Strategy Framework Directive (MSFD, 2008/56/EC) has incorporated new monitoring requirements to assess environmental quality with the aim of achieving the good environmental status (GES) of all European seas by 2020. This requires the monitoring and status assessment of a variety of descriptors, including some related to biological components, such as biodiversity or the presence of non-indigenous species [1]. However, the assessment of biodiversity at any spatial scale is a challenge [2,3] and thus innovative methods and approaches are required [1].

Owing to the availability of well-established and standardized sampling methods, which do not require scuba-diving [4–6], soft bottoms have benefitted from more monitoring studies than hard bottoms and their biodiversity patterns and dynamics are better understood. Additionally, a variety of sensors allow the automated collection of data for sediments [redox potential, organic matter, contaminants] as well as water quality which provides precious information in order to relate changes in community composition with these variables [7]. Some of these measurements could apply to regions of hard substrata as well but are so far to our knowledge not implemented.

In order to standardize the monitoring of benthic hard bottoms, ecologists often use settlement plates or other artificial sampling units. Once these units are colonized by marine organisms, they can be used to monitor or experimentally manipulate benthic communities (*e.g.* [8–15]). Although artificial structures have already been used to compare biodiversity of marine hard bottoms from distinct geographical regions, the artificial substrates varied in size and material while colonization time, processing and analytical protocols were also different, making large-scale comparisons difficult to establish.

To further standardize the sampling of benthic habitats, the Coral Reef Division (CRED) of the United States' National Oceanic and Atmospheric Administration (NOAA) developed Autonomous Reef Monitoring Structures (ARMS) [cf. Zimmerman et al., 2004]. The ARMS are composed of stacked PVC settlement plates and are designed to mimic the 3D structural complexity of coral reef habitats [16–18]. While ARMS were originally designed for coral reef habitats and have been used to monitor them in the Caribbean and Indo-Pacific [17,20] and in the Red Sea [21,22] they have also been deployed in other hard bottomed habitats including of the Atlantic coast of the US [19] and in the Adriatic Sea [Penessi and Danovaro 2017]. PVC, unlike other artificial substrates (e.g. limestone), can be manufactured to the same specifications globally, and thus enabled a standardized ARMS to be constructed (e.g. no geographical variations in construction materials). While other artificial construction materials (e.g. limestone) better mimic some natural environments the use of a single

material allowed for comparisons to be undertaken across a wide range of hard bottomed substrates in geographically separated regions.

One of the main issues associated with marine environmental monitoring is time and cost constraints. While Leray and Knowlton [19] assessed ARMS between temperate and subtropical regions they used a relatively expensive metabarcoding approach. In this study, we aim to test the potential of a photographic assessment of the sessile components of ARMS as a fast community screening tool across a range of environmental conditions. Towards this end, we analyzed the colonization of a subset of plates after more than one year of immersion under different environmental conditions across two regional European seas (Northeast Atlantic Ocean and Mediterranean Sea) as well as the Red Sea. We first tested whether community composition, inferred from photographs, was significantly different amongst the distinct plate surfaces of the ARMS, amongst sites (within seas) and amongst seas. Secondly, we also investigated the effect of various environmental factors reflecting both the level of anthropic pressure and the local diversity of habitats on the biodiversity patterns (although our experimental scheme provided a single environmental modality per site). Thirdly, since monitoring protocols must be as simple and cost-effective as possible, we performed some analyses considering each taxonomic group alone (rather than the whole community composition) and each plate face separately to see if these partial analyses provided useful results (in the aim of simplifying monitoring).

2-Material and Methods

2.1. Monitored sites

The sites sampled in this study were the southern part of the Bay of Biscay, which corresponds to the northeast Atlantic, the northwest Mediterranean coast of France, the Adriatic Sea, and the Red Sea (Fig. 1). Three replicate ARMS units were installed at three sites in each sea for a total of nine units per sea, at a depth of between 7 and 17 m (Table 1, Fig. 1). Information about times of deployment and sites is given in Table1. The sites were chosen based on natural and human pressures inferred by the expert knowledge of the authors (S1 File).

Table 1: Information about ARMS' deployment and sites monitored by four partners: AZTI (Bay of Biscay sites), CNRS-IMBE (northwest Mediterranean sites), CoNISMa (Adriatic Sea sites), and KAUST (Red Sea sites).

SEA REGION (code)	SITE (code)	DEPLOYMENT DATE	RECOVERY DATE	SITE ID [replicates]	LATITUDE	LONGITUDE	DEPTH (m)
Adriatic Sea (AdS)	Grotta Azzurra (Azz)	Jul-14	Jul-15	CONI_S1	N43 37.313	E13 31.691	7
	Due Sorelle (Sor)	Jun-14	Jul-15	CONI_S2	N43 32.953	E13 37.699	8.7
	La Scalaccia (Sca)	Jun-14	Jul-15	CONI_S3	N43 36.291	E13 33.102	8.8
NW Mediterranean (NWM)	Ile de l'Erevine (ELV)	Jun-13	Dec-14	CNRS_S1	N43 19.780	E05 14.210	17
	Ile Riou (RRS)	Jun-13	Dec-14	CNRS_S2	N43 10.370	E05 23.420	17
	Phare de Cassidaigne (CCA)	Jun-13	Dec-14	CNRS_S3	N43 08.740	E05 32.740	17
Bay of Biscay (BoB)	Lekeitio (Lek)	Jun-13	Jul-14	AZTI_S1	N43 22.311	W2 30.258	12.5
	Zumaia (Zum)	May-13	Jul-14	AZTI_S2	N43 18.748	W2 13.641	11
	Pasaia (Pas)	May-13	May-14	AZTI_S3	N43 20.230	W1 55.639	11
Red Sea_Jeddah (ReS)	Janib Sa'ara reef (JSR)	Apr-13	Jun-14	KAUS_S1	N21 27.253	E39 06.661	10
	South of Jeddah (SOJ)	Apr-13	Jun-14	KAUS_S2	N21 13.508	E39 07.237	10
	Qaham reef (QAR)	Apr-13	Jun-14	KAUS_S3	N21 04.921	E39 12.063	10

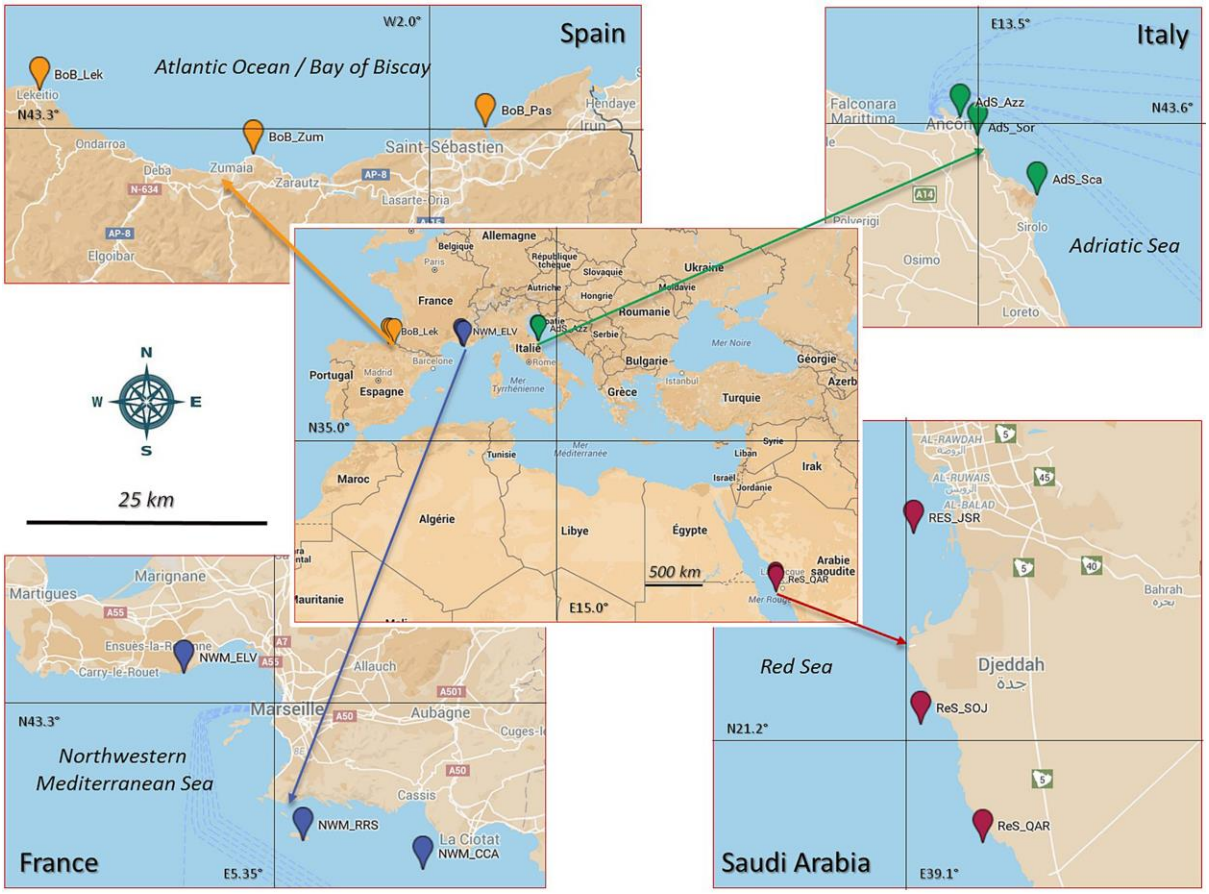


Figure 1: Geographic position of the DEVOTES ARMS sites. Their precise geographic positions are given table 1.

Table 2: Anthropogenic and environmental context of the study sites. Each site (column) is designated by two three-letter codes, one for Sea region and one for site separated by an underscore (full names in Table 1). The factors are Boolean: presence is indicated by Y, absence by N. In Row 2: for each of the four sea regions, we tested whether the community composition of the site (the column head) was significantly different from the other two sites (in a crossed design considering also the effect of plate face); there are three possible contrasts opposing one site with the other two sites ; numbers (1, 2 and 3) refer respectively to the most, second and least significant contrasts and significance are represented by usual symbols (NS: not significant, **: p<0.01, ***:p<0.001, ****:p<0.0001). For each environmental factor (row 3 to 12) we highlighted in green when the most contrasted site configuration singled out the site which differed most strongly from the other two for the environmental factor, in red when it did not. If environmental factors do not influence community composition (so site contrasts), 1/3 should correspond to each possible contrast, so 1/3 to the favourable case (in green). Our results do not depart from random expectation with 6 green, 14 red. Abbreviations are explained in Tab. 1.

	BoB_ Lek (1)	BoB_ Zum (3)	BoB_ Pas (2)	NWM_ ELV	NWM _RRS	NWM _CCA	AdS_ Azz	AdS_ Sor	AdS_ Sca	ReS_ JSR	ReS_ SOJ	ReS_ QAR
Most contrasted site versus 2 other ones for each Sea region (Single sea PERMANOVA: Face x Site)	1****	3***	2***	1***	3**	2***	1****	2**	3 NS	3 NS	1****	2**
Protection status	N	N	N	N	Y	N	N	N	N	N	N	N
Marine debris	N	N	N	N	N	N	Y	Y	Y	Y	N	Y
Sewage output	Y	Y	Y	N	N	N	Y	N	Y	Y	N	N
Chemical pollution	N	N	Y	Y	Y	Y	Y	Y	Y	Y	N	N
Urbanization	N	N	N	Y	N	N	Y	N	Y	Y	N	N
Harbor	Y	Y	Y	Y	N	N	N	N	N	Y	N	N
Nearby Seagrass meadows	N	N	N	Y	N	Y	N	N	N	N	N	N
Nearby sand	Y	N	N	Y	N	Y	N	N	Y	Y	Y	Y
Nearby mud	N	N	N	Y	N	N	N	N	Y	N	N	N

2.2. ARMS implementation and recovery

Each ARMS unit is composed of nine 22.5 cm x 22.5 cm PVC plates and spacers stacked in an alternating series of open and closed formats, attached to a 35 cm x 45 cm base plate (Fig. 2). Further details on the standard assembly, deployment and recovery of the ARMS are available on the NOAA's website[1] and in González-Goñi et al. [23].

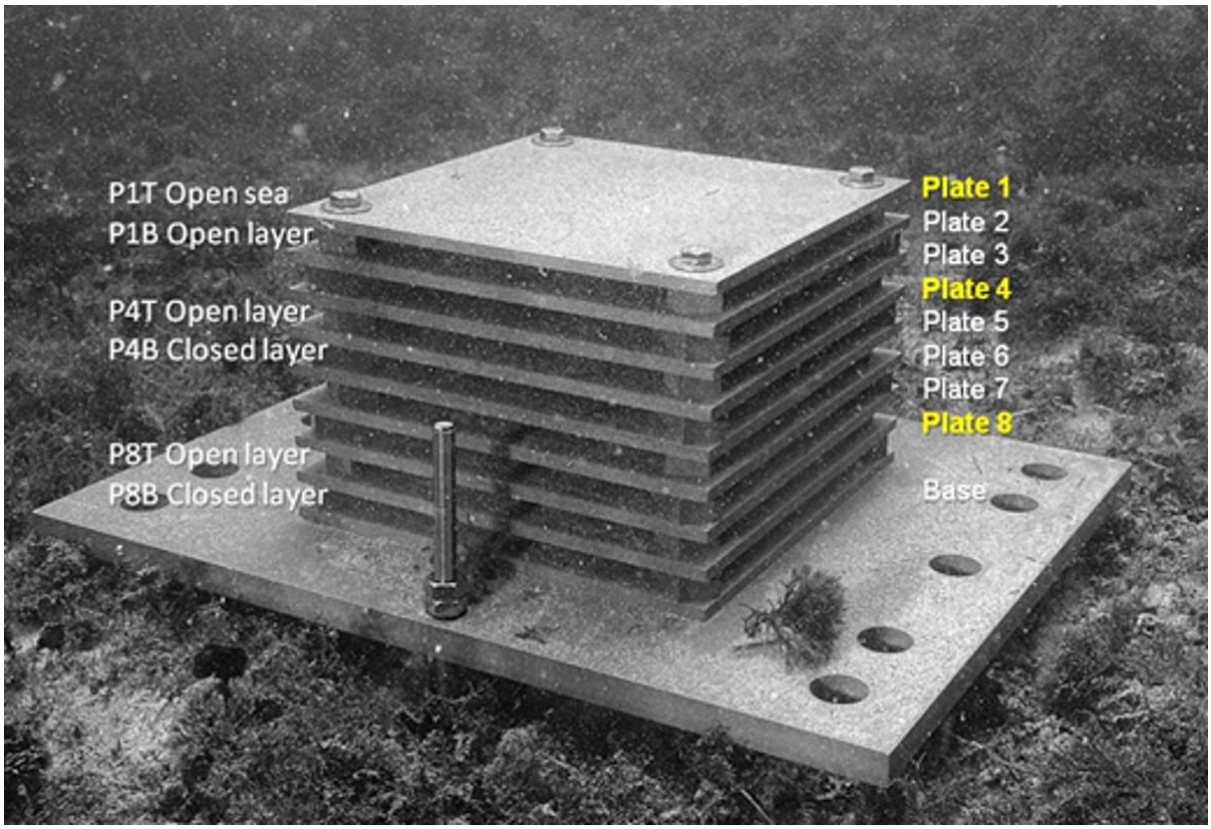


Figure 2: A newly deployed ARMS (Ile de l'Erevine, NW Mediterranean). The alternative use of long and short PVC cross spacers give a tower of four open and four closed layers. © photo CNRS / F. Zuberer.

The ARMS units were installed by divers and submerged for 12 to 16 months, depending on the sea (Table1). Subsequently, ARMS were recovered and returned to the laboratory, where they were dismantled and processed. Each plate surface was gently brushed to remove mobile fauna without

[1] https://www.pifsc.noaa.gov/cred/survey_methods/arms/overview.php

detaching sessile organisms. Plates were kept in seawater aerated with bubblers until photographs were taken (Fig. 3).

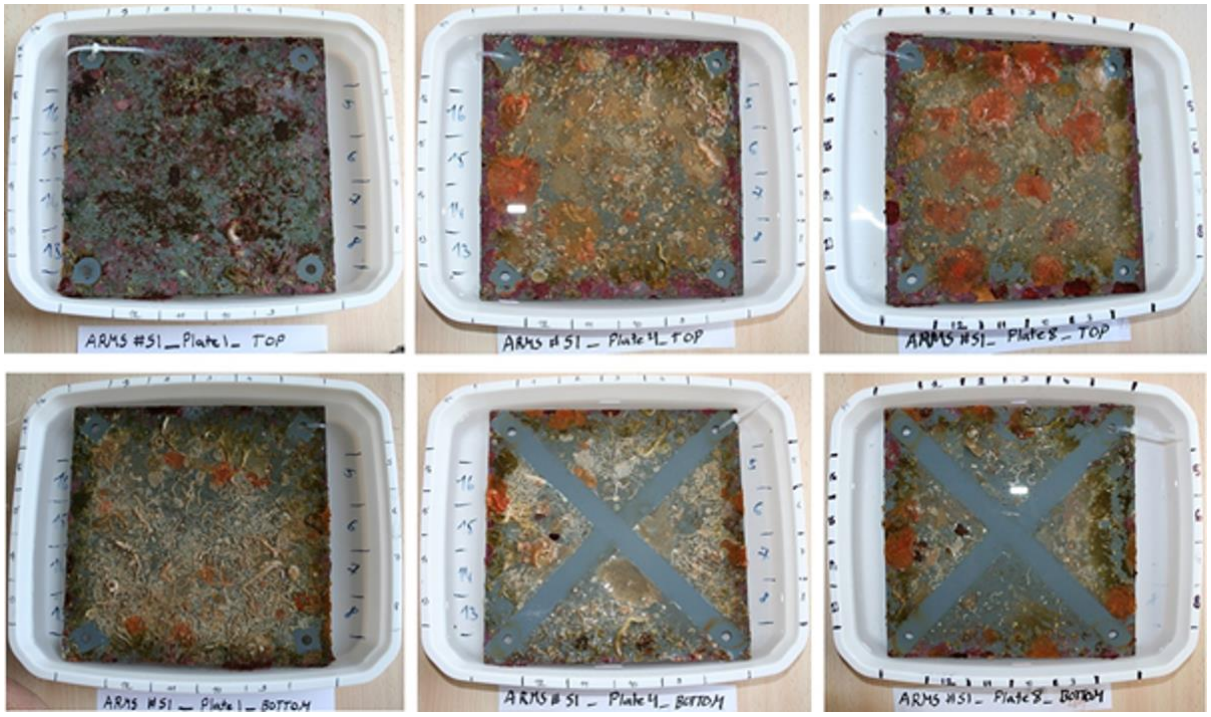


Figure 3: Sampled ARMS face plates (from left to right: 1, 4 and 8; upper pictures: top faces; lower pictures: bottom faces) after being recovered from the Sea and having collected the mobile fraction [NW Mediterranean, Ile Riou]. Cross markings in two of the bottom plates result from the cross spacers that alternate between some plates.

2.3. Photo analyses

Three plates (P) (plates 1, 4, and 8) (Fig. 2) were selected for analysis and for each one, the top (T) and bottom (B) surfaces were analyzed individually (i.e., 6 plate faces analyzed per ARMS). These faces were selected as representative of the different habitats found within an ARMS, which may represent different conditions experienced by the organisms in situ. The top surface of plate 1 is exposed to direct light and without any protection from predators, while the other five faces are not. Among these shaded plate faces, P1B, P4T, P8T are open to the current, while faces P4B and P8B represent less hydrodynamic niches due to compartmentalization, which does not allow the current to flow through the space in between the plates (Figs 2 and 3).

Photographs were analyzed using Photoquad® software [25]. Each photograph was divided in 64 squares and one point was randomly selected within each square. The organism present at each point was identified to the lowest possible taxonomic level by scientists from the four seas (scientists analyzed

ARMS from their own region). Prior to the global analysis, some taxonomic categories were merged in order to be compatible among the four regions and to minimize possible observer effect. The initial categories are provided in supplementary material (S2 File). The final (merged) categories were: Annelida, Bryozoa, Mollusca, Cnidaria, Porifera, Crustacea, colonial Tunicata, Tunicata, calcareous coralline algae (hereafter CCA), other Rhodophyta, Chlorophyta, Phaeophyta, other Algae, Foraminifera, "undetermined", and "not alive". Points that fell on uncolonizable parts of the plate due to the presence of the compartmentalizing cross or screws (Fig. 3) were considered as "not alive" (since not all partners had created a category "uncolonizable" while analyzing their photos). Community analyses were performed both with and without including the categories "not alive" and "undetermined" in the dataset.

2.4. Environmental factors

The simplicity and operability required for monitoring also explains our use of "broadly defined" context factors. On the European coasts, the three sites were chosen to reflect contrasting environmental situations and to respect the conditions of being on hard bottoms and at a reasonable distance apart to make field work feasible and to ensure that the potential pool of colonizer species was shared among sites within a sea. Nine binary environmental factors were assessed in a way to allow hypothesis testing: presence of a protection status (such as national park or marine protected area), marine debris (i.e. reported presence of visible plastic objects, litter, and, abandoned, lost and otherwise discarded fishing gear**s**), wastewater discharges, chemical pollution (influence evaluated after expert opinion), urbanization (influence evaluated after expert opinion), harbor (influence evaluated after expert opinion), nearby seagrass meadows, nearby sand, and nearby mud (at less than 15 m, and with a probable influence on ARMS (diver expert opinion)).

2.5. Statistical analyses

We used the PRIMER package (version 7) [26,27] for all community analyses that were performed on the whole data set (with samples from all faces, sites and seas). For all the analyses we used the Bray-Curtis resemblance measure. All analyses were performed twice, both with and without fourth-root transformation of abundance data (transformations are highly recommended to reduce the effect of abundant taxa on the dissimilarity matrices, but we wanted to check that our results were not biased by a particular transformation).

Multivariate analysis was undertaken using both PERMANOVA (fixed effects, type III sums of squares except for nested designs where type I was used) and, to enable comparisons with earlier studies, ANOSIM (all factors were unordered). We did 9999 permutations for each test. We first tested the effect of sea, sites (nested within sea), and plate face. We then tested the effect of each environmental

factor in a three-way crossed design containing also the factors sea and plate face (in these cases, we used partial data sets including only the sea regions for which the environmental factor was varying among sites). We also performed a nested PERMANOVA with the factors sea and site (within sea) but without the factor face to show the consequences of not-distinguishing plate faces, as was undertaken in a recent metabarcoding study [21].

We compared the dispersion of community compositions between seas, sites and plate faces (and tested the null hypothesis that it was not varying) using PERMDISP.

Additional analyses were also carried out on partial datasets (in order, among other things, to estimate the power of simpler monitoring protocols, that would focus restrict the efforts to single taxonomic groups, or a single plate faces). Taxon abundances were log transformed before performing ANOVA for testing whether effect of sea and site were significant on the abundances of each taxonomic category. ANOVAs were performed in R version 3.2.4 [28]. Similarly, we tested whether the effects of sea and site on community composition were significant for each plate face individually using a nested PERMANOVA (six distinct analyses). We also tested separately for each sea and each of two plate faces (P1T, P4B) whether the effect of site on community composition was significant (thus eight separate analyses were performed on small data sets, one for each selected plate face in each of the four seas). We computed, for each sea, the effect of site in each of the three possible contrasted designs opposing one site versus the two other sites, in a two-way crossed PERMANOVA with factors site and plate face. This allows checking whether the most contrasted site configuration corresponds to contrasted environmental factors (Table 3).

3-Results

Table 3 : The proportion of taxa which were identified at the species level was generally low except for Mollusca.

Number of identified taxa	Specie level	Other levels	Total
Annelida	5	15	20
Bryozoa	5	14	19
Cnidaria	1	6	7
Crustacea	2	2	4
Mollusca	8	2	10
colonial *Tunicata*	2	0	2
non-colonial Tunicata	1	3	4
CCA	1	3	4
Rhodophyta	2	5	7
Chlorophyta	1	1	2

The proportion of taxa which were identified at the species level was generally low (see Table XXX). The list of all taxa initially identified in each sea (i.e. prior to merging) is available in S2 File. The row-data used for statistical community analyses however correspond to the merged taxonomical category from the six plate faces of all ARMS (S3 File). Unless specified, the results presented below were obtained with transformed abundance data from the data set containing the "not alive" and "undetermined" categories because the other datasets (i.e. untransformed abundances, and without "not alive" and "undetermined" categories) yielded similar results.

On average, the percentage of ARMS area colonized by an identifiable taxonomic category ranged from 50% in the Adriatic Sea and 60% in the Red Sea to over 70-75% in the Bay of Biscay and the Mediterranean Sea. There was no consistent pattern between top and bottom faces across seas for plate 1 (the comparison was not possible for plates 4 and 8 because of the compartmentalization of bottom, but not top faces): in the Adriatic and Red Sea, bottom faces (P1B) were more colonised than top ones (P1T), with 72% and 52% for the Adriatic Sea and 84% and 71 % for the Red Sea, but differences were small (and reversed) in the two other seas. These are underestimates of biological colonization because of undetermined points (on average 2% of points were undetermined in AdS, 4% in ReS and BoB, and

7% in NWM). Most abundant groups represented on the plates of ARMS (Fig. 4) were Annelida, Bryozoa, Porifera and Mollusca for animals, CCA and other Rhodophyta for algae. Groups like Tunicata (colonial and non-colonial Tunicata) and Cnidaria for animals and Chlorophyta and Phaeophyta for algae were much less abundant or widespread. Animals represented the largest part of the colonization with up to 54% for the northwest Mediterranean Sea, 48% for the Bay of Biscay, 38% for the Adriatic Sea and 32% for the Red Sea, whereas algae represented more than 20% of the total colonization in northwest Mediterranean Sea and Bay of Biscay, 11% for the Adriatic Sea and 27% for the Red Sea. Across all seas, the relative abundances of the most abundant groups differed between the top and bottom faces of plates: Annelida and Bryozoa were more frequent on bottom faces whereas CCA and other Rhodophyta preferred top faces (Fig. 4).

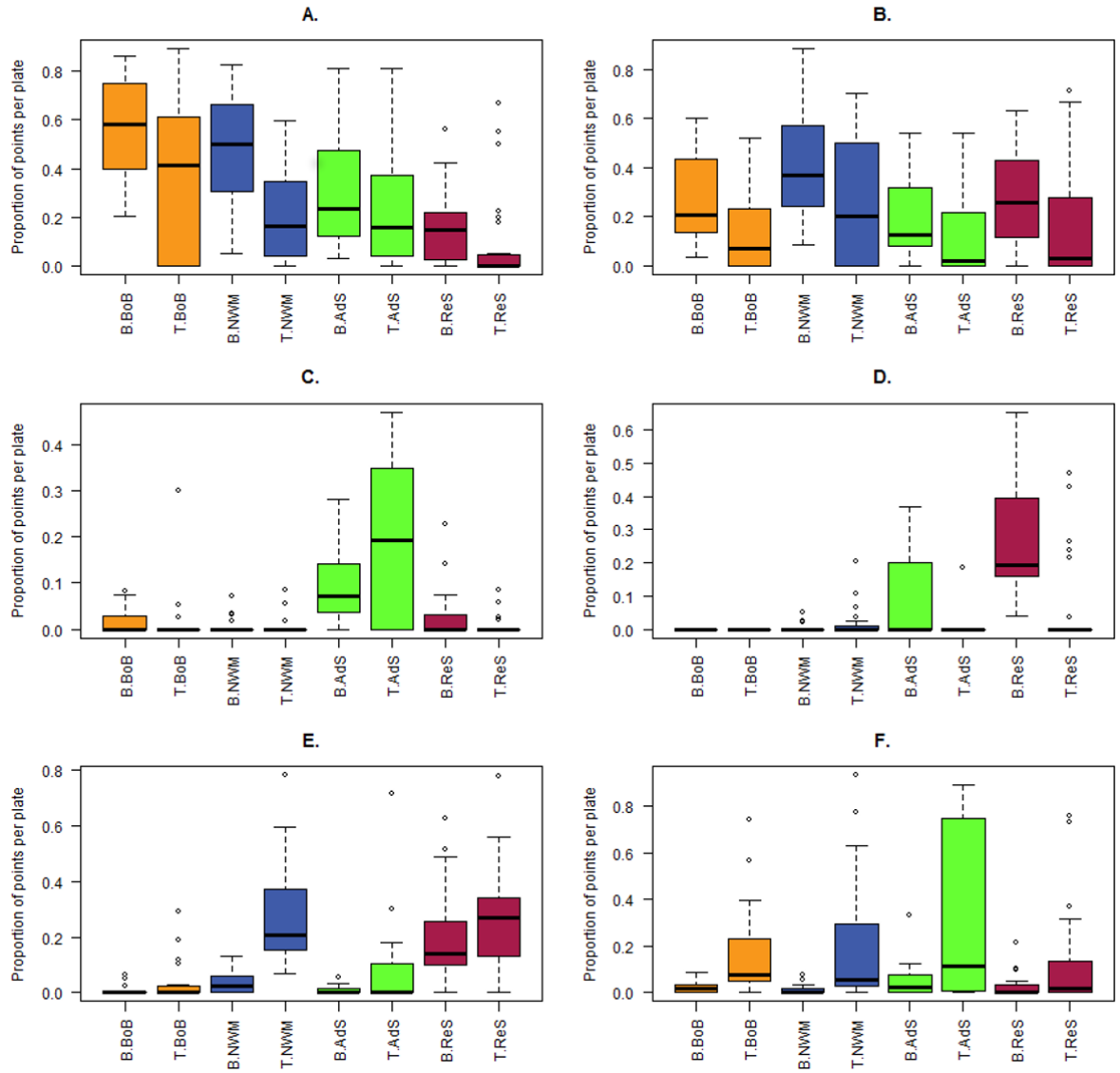


Figure 4: Boxplots showing median [and quartiles 1 and 3] abundances of main animal and algal groups on ARMS from the four Sea regions. Annelida (A), Bryozoa (B), Mollusca (C), Porifera (D), Calcareous Coralline Algae (E), other Rhodophyta (F), on ARMS from four Seas. Seas

are ordered from west to east and named as in Table 1. Top (T) and bottom (B) plates were analyzed separately.

In general taxa (log transformed abundances) showed significant differences among seas but not sites within each sea (Table 4). In particular, Annelida and Bryozoa were more abundant in the Bay of Biscay and northwest Mediterranean compared to the Adriatic Sea and the Red Sea. Mollusca were more abundant in the Adriatic Sea and Porifera were especially abundant in the Adriatic and Red Sea (particularly on the bottom faces of the plates), while CCA were more abundant in the northwest Mediterranean Sea and in the Red Sea (Fig. 4, Table 4). For both the top and bottom faces of the plates, seas could be differentiated based on the abundances of CCA, Mollusca and, despite their low abundances, Phaeophyta. Only Chlorophyta in top plates revealed significant intra-sea variability.

Table 4: P-values for each taxon from nested ANOVAs (4 Regions, 3 sites within each region) based on log-transformed abundances and written in bold when significant (<0.05). CCA: Calcareous Coralline Algae.

	Top Faces		Bottom Faces	
Taxon	P-value Sea	P-value Site	P-value Sea	P-value Site
Annelida	**0.029**	0.197	**<0.001**	0.997
Bryozoa	**0.031**	0.331	**0.032**	0.544
Mollusca	**<0.001**	0.615	**0.009**	0.588
Porifera	0.163	0.115	**<0.001**	0.993
CCA	**<0.001**	0.103	**0.008**	0.737
Other Rhodophyta	**0.009**	0.961	0.094	0.904
Chlorophyta	0.661	**<0.001**	**0.029**	0.677
Phaeophyta	**0.005**	0.683	**0.005**	0.947

PERMANOVA and ANOSIM performed on the whole dataset revealed highly significant effects of sea, site (nested within sea), and plate face on community composition (see S4 File for detailed PERMANOVA results). There was also a highly significant effect of the interaction between sea and face, and between site (nested within sea) and face, but less significant.

All pairwise sea comparisons (two-way crossed design of plate face and sea) were highly significant. PERMDISP did not show significant differences in dispersion between seas, nor between sites. All pairwise plate face comparisons displayed highly significant differences in community composition, except between faces P4T and P8T (not significant, P=0.1483) and between the untransformed data of the two compartmentalized faces P4B-P8B (fourth-root transformed data, P= 0.0118). The dispersion was highly differentiated among plate faces (PERMDISP P-values = 0.0001) and pairwise PERMDISP tests indicated three levels of dispersion: highly dispersed faces (P1T, P8T), moderately dispersed faces (P4T, P1B), and less dispersed faces (P4B, P8B) (detailed values not shown).

Non-metric multidimensional scaling (nMDS) (Fig. 5) illustrates the uniqueness of plate P1T communities with respect to the other ones within each sea and, although less clearly and not for all seas, differences between top and bottom plate faces. Algal taxa (Chlorophyta, Phaeophyta, other Rhodophyta, and other algae, but not CCA) tend to be more abundant on the exposed faces (P1T) in all four regions as reflected by the position of these variables on the nMDS plot (S1 Fig). By contrast, all animal variables are distant from P1T samples in the nMDS plot. Variables of species composition (taxa) are also represented in S1 Fig. to visualize possible important associations of some taxa with some plates (e.g. various algal taxa are close to the P1T in the 2D-representation).

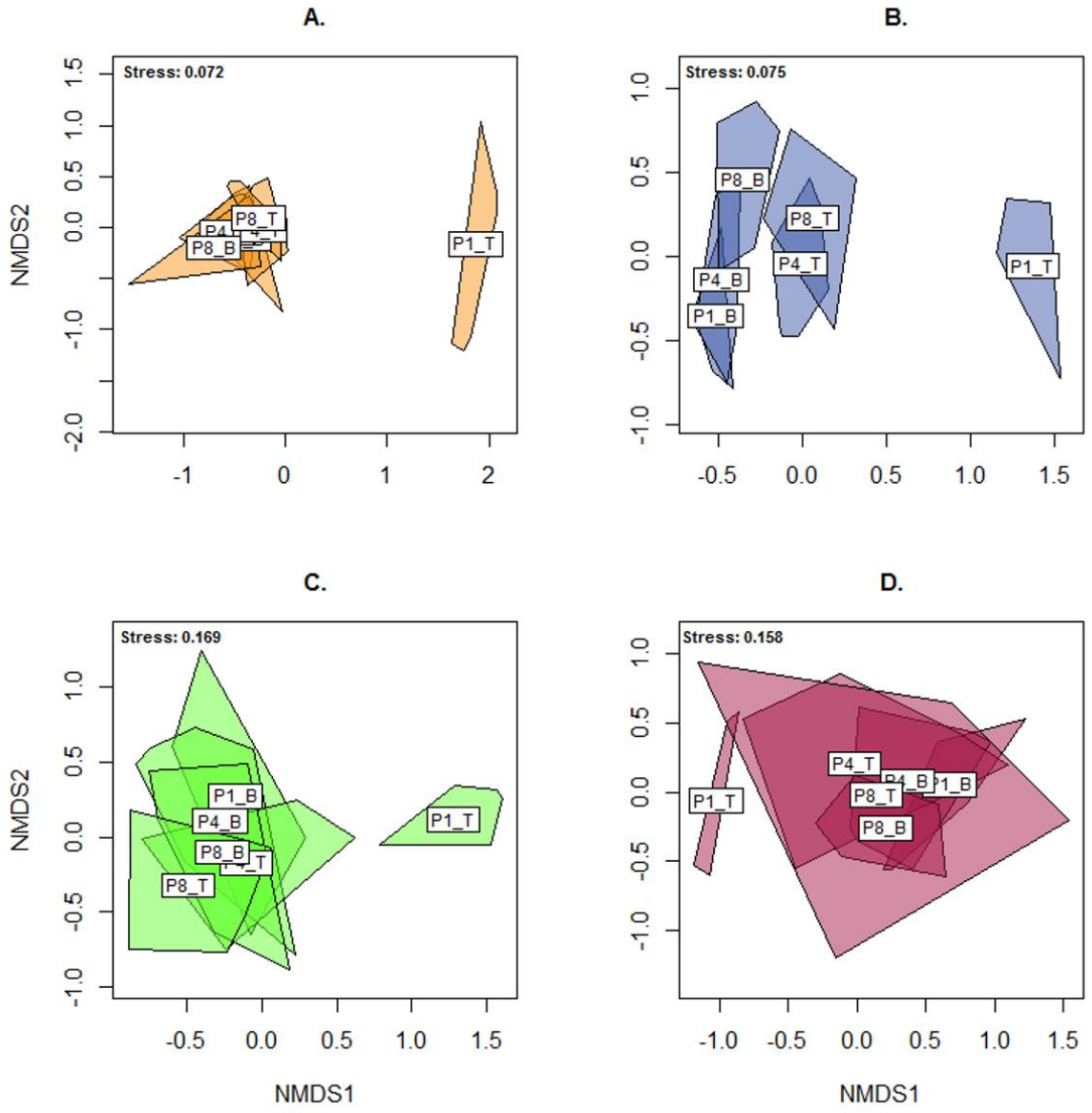


Figure 5: Non-metric multidimensional scaling (nMDS) representation of community for each plate face within each Sea region. A: Bay of Biscay, B: Northwest Mediterranean Sea; C: Adriatic Sea, D: Red Sea. Each dot represents one ARMS for a given plate face, the stress value is indicated.

For each of the six plate faces analyzed separately, there was a highly significant effect of both sea and site (Table 5) with the single exception of P8T for which site was not significant, but this is probably explained by missing data since P8T pictures were missing for two ARMS in the Bay of Biscay whereas photographs were available for all the other plate faces of these ARMS. When analyzed within each sea separately, both faces P1T and P4B were (individually) able to detect significant effects of sites except in the Red Sea (Table 6) but the number of distinct permutations was reduced in these small data sets (S4 File) consisting of at most 9 samples (three ARMS in each of three sites).

Table 5: Summary of PERMANOVA results for each plate face: Plate faces are denominated by PiT or PiB, with i for the plate number, and T for top face, or B for bottom face. P-values for both regions and sites (within regions). Plate peculiarities are indicated in parenthesis (plate 1 top is exposed to the exterior environment, and plates 4 bottom and 8 bottom are compartmentalized by a central cross). All P-values are highly significant except for P8T (there are two missing pictures, for P8T, both in BoB).

Plate face	Sea region (P-value)	Site (P-value)
P1T (exposed)	**<0.001**	**<0.001**
P4T	**<0.001**	**0.004**
P8T	**<0.001**	0.749
P1B	**<0.001**	**<0.001**
P4B (comp.)	**<0.001**	**<0.001**
P8B (comp.)	**<0.001**	**0.002**

Table 6: Results of one-way PERMANOVA testing the effect of sites for two plate faces (in columns) : P-values corresponding to the site effect are reported, written in bold when significant (P< 0.05). The nomenclature for plate faces is explained in the legend of Table 4.

Sea region	P1T (P-value)	P4B (P-value)
Bay of Biscay	**0.047**	**0.047**
Northwestern Mediterranean	**0.007**	**0.047**
Adriatic Sea	**0.035**	**0.024**
Red Sea	0.670	0.051

Among the six Boolean environmental factors which could be tested in multiple seas, four were significant, generally between the 2%-5% probability level but there was a much more significant effect of the interaction of the environmental factor with the sea (generally $p < 0.01$ or 0.001), suggesting that the observed effect of the environmental factor indeed reflected an effect of site (remember that the effect of site within sea was highly significant with the lowest p-value possible for the 9999 permutations of the data: 0.0001). Significant factors (but collinear with sites) were urbanization, chemical pollution and presence of a harbor. Nearby mud, nearby sand and sewage output were not significant. Indeed, "harbor presence", with fourth root transformed data, was more significant (P=0.0285) than the interaction sea x harbor (P=0.0541 NS), but this result was not obtained with untransformed abundances, or with square-roots of abundances (harbor became not significant, while the interaction "harbor x sea" became significant). For the environmental factors which varied only within a sea, the only way to investigate their potential effect on community composition was to check whether the site which was singled out with respect to the Boolean environmental factor was also the best contrast for the site effect (among 3 possibilities). Globally we obtained 6 significant correspondences out of 20, which is very close to the expected proportion (20/3) under the null hypothesis that there is no effect of environmental factors on community composition (Table 2). Thus we have no reliable evidence of an influence of any tested environmental factor on community composition, probably due to our experimental design with few sites per sea, no variation of environmental factors within sites, and strong effects of both sea and site.

In the nested PERMANOVA with the factors sea and site (within sea), but without face, both sea and sites remained significant (S4 File, n°26 to be changed, remove non-useful PERMANOVAs, check discussion about need to separate "face").

4-Discussion

Despite the fact that we used high taxonomic categories, ARMS photo-analyses appeared to be a powerful way to compare marine benthic communities. We detected significant effects at all the levels of our experimental design: sea, site (within sea), and plate face.

The ability of ARMS photo analysis to discriminate among seas is not surprising because the seas correspond to well-differentiated biogeographical units [29] with substantial distinctions in a variety of environmental parameters (such as salinity, light or nutrient availability). Differences among regions cannot be rigorously interpreted since photographs from the distinct seas were analysed by distinct observers and the ARMS were not installed for identical durations (and in the case of NWM, also depth) in the four seas. Obvious seasonal effects are observable in temperate seas, including successions of organisms, as opposed to more equatorial latitudes (where little variation in day-length or temperature

occurs over the year) [30,31] therefore the deployment and removal date discrepancies between seas probably contribute to the difference in community composition imputed to the "sea" factor in our statistical analyses.

Nevertheless, our study is to our knowledge the first one encompassing such a variety of non-tropical regions. What is particularly relevant for a pilot study aimed at assessing the potential usefulness of monitoring devices is their ability to discriminate among sites within a sea or a region, because distinct sites are susceptible to correspond to distinct environmental conditions. Indeed, the distinct sites in a given region share a common pool of species). Sites were chosen *a priori* to correspond to contrasted environmental contexts and levels of pressures in each sea. The statistical design of this pilot study however was not primarily intended to test an effect of the environmental factors but rather to check the feasibility of community monitoring in very distinct sea regions. Since within each sea, ARMS communities significantly differed among sites and among plate faces, ARMS photography analysis appears a promising monitoring tool in each of these four regions. A more appropriate design to investigate regional environmental effects would use substantially more sites within a region in order to separate pure spatial effects from environmental effects. For instance, with at least 6-12 sites linearly placed parallel to the coast line, and variations in environmental factors interspersed in each part of this transect, it should be possible to isolate a purely spatial effect (computing the correlation of the distances in community composition with the spatial distances between sites) from the filtering effect of the environmental context. By focusing on a given region (for instance the Bay of Biscay, or the French Mediterranean Coast), it would also be possible to use finer taxonomic categories, which would probably provide more power to detect environmental effects on species composition. For illustration, in the Bay of Biscay, North West Mediterranean, Adriatic Sea and Red Sea, we initially had reported 31, 36, 33 and 34 distinct taxa respectively, although our global analysis used only 14 categories (due to merging).

Indeed, we found no clearcut evidence of an effect of the environmental factors investigated. We cannot conclude whether the significant differences found among sites (within sea) are due to typical spatial effects (the fact that distinct species couldcolonize ARMS in distinct sites, due to connectivity patterns) or to some purely environmental effects (that is, involving the filter of natural selection) that differed among sites. For the North West Mediterranean region, we have evidence from population genetics data that gene flow and dispersal are limited between the sites, despite their proximity (Cahill et al 2017; De Jode et al. in prep), so a purely spatial effect is likely to explain at least part of the differences among sites. In the same area, a metabarcoding study of established coralligenous communities from 19 sites also revealed very strong spatial effects (in addition to smaller but significant environmental effects) (De Jode, David and Chenuil, personal communication). However, this sea region was also the single one where most environmental factors (4 out of 7) matched the highest contrasts among sites (Table 1).

Our results support the proposition that photo identification of ARMS plates can be utilized as a fast-screening tool to detect changes in the community composition at relatively small spatial scales (tens of km). This result is expected to be conservative because we used very broad taxonomic categories. Indeed, taxonomic categories defined by scientists from each distinct sea had to be merged to limit the risk of errors among observers. As already underlined, discriminatory power could be increased in future studies focusing on a single sea by using finer taxonomic categories and including more numerous sites. Some taxa, when considered alone, proved more powerful than others to discriminate among seas. However, single taxa abundances, contrary to community composition, were unable to discriminate sites within seas, with the exception of Chlorophyta in top faces. By contrast, the community analyses (nested PERMANOVA of sites within seas) found significant effects for both sea and site even when the face factor was not considered. In addition, for five of the six plate faces analyzed (the exception, for P8T, being explained by two missing pictures), when taken individually, community compositions significantly differed both among regions and among sites (within regions). This is an illustration of the limit inherent to approaches based on single or few indicator taxa (at least for high rank taxa): community composition data are more powerful than single taxon abundance data to detect changes and thus to monitor ecological status [6]. The fact that the sole plate face (P8T) that was unable to differentiate sites within seas was the one for which two sites had missing data (and were therefore represented by duplicates rather than triplicates) suggests that triplicating ARMS in each monitored site is useful. When an ARMS unit is lost accidentally, the fact of having multiple faces in other ARMS units may partially compensate the lack of information, since single faces within a sea were able to differentiate sites in three out of four seas (Table 5). Indeed, one striking feature of our study is the strong differences in community composition among plate faces of the ARMS: this establishes that the faces represent distinct micro-habitats. They display obvious differences in terms of light exposure, predation exposure, sedimentation and water flow, for instance. The fact that several algal taxa appear to contribute strongly to the uniqueness of plate P1T (nMDS, Fig. 5) is a clear illustration, probably explained by light exposure. A recent study conducted on microalgae colonizing ARMS provided similar conclusions [32]. For compartmentalized faces (P4B and P8B) each such face represents 4 independent colonization units and/or is less susceptible to be affected by a given random event such as predation by a grazer. This interpretation is supported by the fact that these faces display the lowest levels of data dispersion, and significantly less than the other ones (PERMDISP analyses). The sandwich-like structure of ARMS thus appears a positive feature of these systems, as compared with single layer colonization plates, because it allows sampling distinct micro-habitats and guarantees a balanced, thus more powerful design (each plate face being present in each sampling unit).

Recommendations can be made for further ARMS implementations to make them an efficient monitoring tool which could be used to detect effects of environmental conditions changes and/or

anthropogenic pressure. We already invoked the necessity of using more numerous sites within regions because a very strong effect of site on community composition was detected and may be due to purely spatial (available species pool), not environmental (i.e. related to natural selection or ecological niche), effects.

Considering plate faces separately is likely to improve statistical power but analyzing all 16 faces of each ARMS is probably not necessary. We recommend however that future studies will compare all 16 faces in order to provide a complete picture of the pairwise differences and dispersion of an ARMS. Based on the results, a subset of faces could be selected for repeated temporal monitoring according to their ability to reveal differences associated with particular environmental parameters.

Because ARMS are left submerged for at least 1 year, they are not suitable for monitoring on short time scales (e.g. perturbations on time scales of days, weeks or months), but are more appropriate for monitoring of longer term trends. For this reason, ARMS units should be replaced at the same sites regularly in order to provide a time series, which would enable the detection of changes in the community composition to chronic perturbations (e.g. eutrophication, climate change). At last, the use of ARMS is complementary to studies investigating the established benthic community as the ARMS only represent pioneer or early successional stages of the community which may show differential responses to impacts compared with the established community [21]. Because of the early successional status of the community on the ARMS, these units could possibly be used to monitor the arrival of non-indigenous species [11,33]. Lastly, in order to make the data as reproducible and verifiable as possible, data should be placed in an open access repository along with appropriate metadata. This would then improve future monitoring efforts as future studies will have a reference dataset usable for comparison.

Typically, monitoring programs are not undertaken on pan-regional scales and instead focus on smaller regional scales. As already stated, in order to better understand how distinct environmental pressures are affecting the marine benthic community, a larger number of sites should be assessed in each region. Whenever possible, the regular measure of temperature data as well as the analysis of nutrient, chlorophyll and chemical contaminant concentrations would enable to relate changes in biodiversity and community composition to physical-chemical variables. Temperature and light sensors (although light sensors require regular maintenance or recalibration due to immediate biofilm formation) are now affordable and can be attached to at least one unit per site. Benthic communities strongly vary according to physical parameters (depth, orientation, slope, rugosity) (De Jode et al, in prep.) so these factors should be kept constant across sites, or, but at the cost of increased amount of work, varied within each site to investigate their influence on community composition.

Benthic substrata, when monitored with photographic approaches, only reveal a superficial part of the local biodiversity [24]. Photo-inferred communities may not accurately represent the full diversity in highly complex 3D-habitats, such as coralligenous reefs. This is especially true as they only focus on the sessile component of the community. Further, as this analysis was undertaken by various investigators, taxonomic resolution was lost when the data was combined into a single dataset. While the use of a single expert to analyze all the plates would negate this issue, this is not always possible for large datasets production incorporating cross-border institutions. Molecular techniques (including barcoding and metabarcoding as proposed by Leray and Knowlton [19] in their analysis of ARMS), although with their own limitations [34] could provide new opportunities to make the analysis of ARMS more efficient and standardized [20]. Indeed, Pearman et al. [35] showed that for ARMS in the Red Sea, a higher diversity, encompassing a broader range of taxa was observed, when using molecular techniques compared with morphological approaches. In their analyses, metabarcoding was able to differentiate sites while morphological approaches did not show a significant difference in the composition between sites (despite the sessile fraction of all plate faces had been pooled for metabarcoding). An ongoing project will present the results of metabarcoding analyses carried out on the ARMS studied here (and including additional European regions).

As the conclusion, this study established that ARMS photo-analyses provide an efficient tool to reveal the effect of seas and sites about 10-30 km distant one another with a protocol simple enough to be generalized for monitoring applications.

To summarize, the key points of this study are:

- ARMS photo-analyses provide relevant information in each sea
- The distinct ARMS faces correspond to different micro-habitats
- 3 replicate ARMS appear necessary in each site
- 64 points are sufficient to detect consistent differences in community composition
- The use of broad taxonomic categories did not impede to detect differences among sites

5-**References cited**

1. Danovaro R, Carugati L, Berzano M, Cahill AE, Carvalho S, Chenuil A, et al. Implementing and innovating marine monitoring approaches for assessing marine environmental status. Front Mar Sci. 2016;3. Available: http://plymsea.ac.uk/id/eprint/7299

2. Borja Á, Elliott M, Andersen JH, Berg T, Carstensen J, Halpern BS, et al. Overview of Integrative Assessment of Marine Systems: The Ecosystem Approach in Practice. Front Mar Sci. 2016;3. doi:10.3389/fmars.2016.00020

3. Selig ER, Longo C, Halpern BS, Best BD, Hardy D, Elfes CT, et al. Assessing global marine biodiversity status within a coupled socio-ecological perspective. PloS One. 2013;8: e60284.

4. Borja Á, Franco J, Pérez V. A marine biotic index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. Mar Pollut Bull. 2000;40: 1100–1114.

5. Patrício J, Little S, Mazik K, Papadopoulou N, J. Smith C, Teixeira H, et al. European Marine Biodiversity Monitoring Networks: strengths, weaknesses, opportunities and threats. Front Mar Sci. 2016;3. doi:10.3389/fmars.2016.00161

6. Borja Á, Marín SL, Muxika I, Pino L, Rodríguez JG. Is there a possibility of ranking benthic quality assessment indices to select the most responsive to different human pressures? Mar Pollut Bull. 2015;97: 85–94.

7. Birchenough SNR, Parker RE, McManus E, Barry J. Combining bioturbation and redox metrics: potential tools for assessing seabed function. Ecol Indic. 2012;12: 8–16.

8. Altman S, Whitlatch RB. Effects of small-scale disturbance on invasion success in marine communities. J Exp Mar Biol Ecol. 2007;342: 15–29. doi:10.1016/j.jembe.2006.10.011

9. Bowden DA, Clarke A, Peck LS, Barnes DK. Antarctic sessile marine benthos: colonisation and growth on artificial substrata over three years. Mar Ecol Prog Ser. 2006;316: 1–16.

10. Judge ML, Craig SF. Positive flow dependence in the initial colonization of a fouling community: results from in situ water current manipulations. J Exp Mar Biol Ecol. 1997;210: 209–222.

11. Marraffini ML, Ashton GV, Brown CW, Chang AL, Ruiz GM. Settlement plates as monitoring devices for non-indigenous species in marine fouling communities. Management. 2017;8: 559–566.

12. Moura A, da Fonseca LC, Cúrdia J, Carvalho S, Boaventura D, Cerqueira M, et al. Is surface orientation a determinant for colonisation patterns of vagile and sessile macrobenthos on artificial reefs? Biofouling. 2008;24: 381–391. doi:10.1080/08927010802256414

13. Piola RF, Johnston EL. Pollution reduces native diversity and increases invader dominance in marine hard-substrate communities. Divers Distrib. 2008;14: 329–342.

14. Sorte CJ, Fuller A, Bracken ME. Impacts of a simulated heat wave on composition of a marine community. Oikos. 2010;119: 1909–1918.

15. Féral J-P, Saucède T, Poulin E, Marschal C, Marty G, Roca J-C, et al. PROTEKER: implementation of a submarine observatory at the Kerguelen Islands (Southern Ocean). Underw Technol. 2016;34.

16. Brainard, RE, Asher, J, Blyth-Skyrme, V, Coccagna, EF, Dennis, K, Donovan, MK, et al. Coral reef ecosystem monitoring report of the Mariana Archipelago: a 2003 – 2007 Pacific Islands Fisheries Science Center. 2012.

17. Knowlton N, Brainard RE, Fisher R, Moews M, Plaisance L, Caley MJ. Coral reef biodiversity. Life World's Oceans Divers Distrib Abundance. 2010; 65–74.

18. Plaisance L, Caley MJ, Brainard RE, Knowlton N. The diversity of coral reefs: what are we missing? PLoS One. 2011;6: e25026.

19. Leray M, Knowlton N. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proc Natl Acad Sci. 2015;112: 2076–2081.

20. Ransome E, Geller JB, Timmers M, Leray M, Mahardini A, Sembiring A, et al. The importance of standardization for biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. PLOS ONE. 2017;12: e0175066. doi:10.1371/journal.pone.0175066

21. Pearman JK, Anlauf H, Irigoien X, Carvalho S. Please mind the gap–Visual census and cryptic biodiversity assessment at central Red Sea coral reefs. Mar Environ Res. 2016;118: 20–30.

22. Al-Rshaidat MM, Snider A, Rosebraugh S, Devine AM, Devine TD, Plaisance L, et al. Deep COI sequencing of standardized benthic samples unveils overlooked diversity of Jordanian coral reefs in the northern Red Sea. Genome. 2016;59: 724–737.

23. González-Goñi, L., Borja, A., Uyarra, M.C. Comparación de herramientas de análisis de imagen: eficiencia y uso en ecología bentónica de sustrato duro. Rev Investig Mar. 2017;24: 1–12.

24. Sini M, Kipson S, Linares C, Koutsoubas D, Garrabou J. The Yellow Gorgonian *Eunicella cavolini*: demography and disturbance levels across the Mediterranean Sea. PloS One. 2015;10: e0126253.

25. Trygonis V, Sini M. photoQuad: a dedicated seabed image processing software, and a comparative error analysis of four photoquadrat methods. J Exp Mar Biol Ecol. 2012;424: 99–108.

26. Clarke K.R, Gorley R.N, Somerfield P.J., Warwick R.M. Change in marine communities: an approach to statistical analysis and interpretation. 3rd edition. PRIMER-E, Plymouth. 2014.

27. Clarke K.R., Gorley R.N. PRIMER v7: User Manual/Tutorial. PRIMER-E, Plymouth. 2015.

28. R Core Development Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Internet]. 2016. Available: URL https://www.R-project.org/

29. Spalding MD, Fox HE, Allen GR, Davidson N, Ferdaña ZA, Finlayson MAX, et al. Marine ecoregions of the world: a bioregionalization of coastal and shelf areas. BioScience. 2007;57: 573–583.

30. Mellin C, Mouillot D, Kulbicki M, Mcclanahan TR, Vigliola L, Bradshaw CJA, et al. Humans and seasonal climate variability threaten large-bodied coral reef fish with small ranges. Nat Commun. 2016;7.

31. van Hoytema N, Bednarz VN, Cardini U, Naumann MS, Al-Horani FA, Wild C. The influence of seasonality on benthic primary production in a Red Sea coral reef. Mar Biol. 2016;163: 52.

32. Pennesi C, Danovaro R. Assessing marine environmental status through microphytobenthos assemblages colonizing the Autonomous Reef Monitoring Structures (ARMS) and their potential in coastal marine restoration. Mar Pollut Bull. 2017;

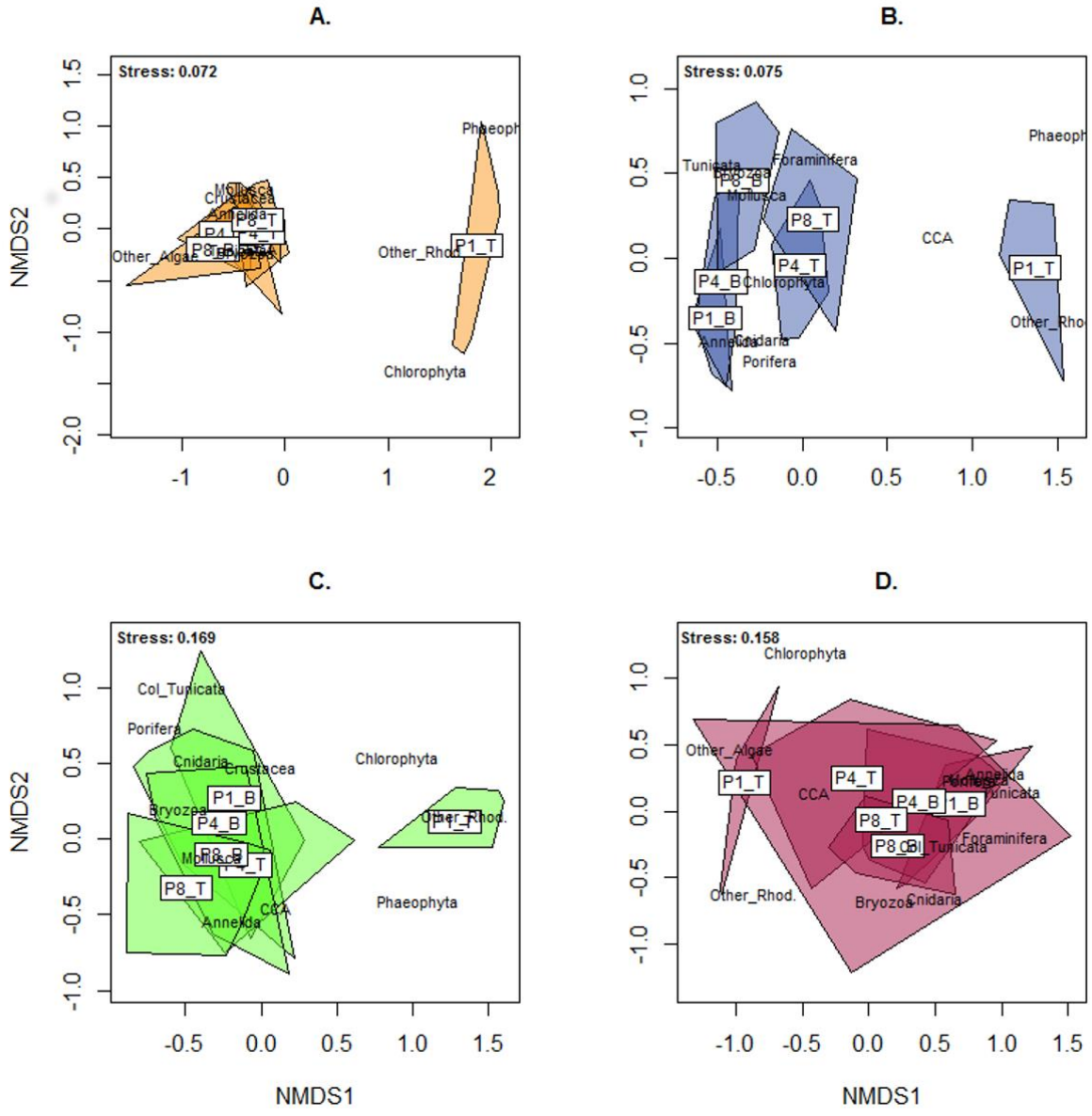
33. Hayes KR, Cannon R, Neil K, Inglis G. Sensitivity and cost considerations for the detection and eradication of marine pests in ports. Mar Pollut Bull. 2005;50: 823–834.

34. Carugati L, Corinaldesi C, Dell'Anno A, Danovaro R. Metagenetic tools for the census of marine meiofaunal biodiversity: An overview. Mar Genomics. 2015;24: 11–20. doi:10.1016/j.margen.2015.04.010

35. Pearman JK, Irigoien X, Carvalho S. Extracellular DNA amplicon sequencing reveals high levels of benthic eukaryotic diversity in the central Red Sea. Mar Genomics. 2016;26: 29–39.

36. Zimmerman, T. L. and J. W. Martin. 2004. Artificial Reef Matrix Structures (Arms): An Inexpensive and Effective Method for Collecting Coral Reef-Associated Invertebrates. Gulf and Caribbean Research 16 (1): 59-64.

6-Supporting information



S1 Fig. Non-metric Multidimensional scaling (nMDS) plots of the plate faces for each sea.

Non-metric multidimensional scaling (nMDS) representation of community for each plate face, within each Sea. A: Bay of Biscay, B: Northwest Mediterranean Sea; C: Adriatic Sea, D: Red Sea. Each dot represents one ARMS for a given plate face, the stress value is indicated. Variables of species composition (taxa) are also represented to visualize possible important associations of some taxa with some plates (e.g. various algal taxa are close to the P1T in the 2D-representation).

S1 Tab. Results of three-way ANOSIM with various environmental factors : Symbols for significance levels are : : #:P<0.1; *:P<0.05; **:P<0.01; ***:P<0.001; NS: not significant

Environmental factor tested	R parameter (ANOSIM)	Significance level ANOSIM (R & significance level)
Protection_status	0.188	*
General_anthropization	0.077	#
Marine_debris	0.133	#
Sewage_output	0.104	*
Chemical_pollution	-0.09	#
Urbanization	0.086	*
Harbor	0.043	NS
Nearby_Seagrass_meadows	0.188	*
Nearby_sand	0.26	***
Nearby_mud	0.13	*

S1 File. Description of the 12 study sites (three in each sea).

S2 File. List of taxa and number of occurrences in data from the four Sea regions, prior to merging taxonomic categories.

S3 File. Raw-data excel file. For each sample, the number of points affected to each category is given.

S4 File. Detailed results of PERMANOVA analyses. The outputs of 36 PERMANOVA analyses performed in the Primer7 package are listed below.

Annexe 2 : liste des programmes et responsabilités associées

Programmes européens :

- CIGESMED - Coralligenous based Indicators to evaluate and monitor the "Good Environmental Status" of the MEDiterranean coastal waters (ANR conventions n° 12-SEAS-0001-01, 02 and 03 for France; GSRT - 12SEAS-12-C2 for Greece; TUBITAK Project No: 112Y393 for Turkey) - Responsable du Work Package 6 (bases de données et data mining)
- DEVOTES - DEVelopment Of innovative Tools for understanding marine biodiversity and assessing good Environmental Status www.devotes-project.eu (European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°308392) : Participation à la mise en oeuvre de suivis basés sur les structures ARMs et ASUS.

Financements obtenus dans le cadre du CNRS et responsabilités associées :

L'ensemble des financements présentés ci-dessous ont été utilisés pour développer des projets dans le cadre du consortium IndexMed (devenu en 2017 "IndexMEED"). J'ai été responsable depuis sa création de l'instruction des demandes de financements, de l'animation de ce consortium, de la réalisation des projets financés et de la rédaction des rapports demandés :

- GRAMINÉES - GRAphe data Mining In Natural, Ecological and Environmental Sciences (Financement GDR MaDICS) (Financement 2017 FRB et GDR MaDICS, financement reconduit en 2018 par le GDR MaDICS)
- IndexMEED - Indexing for Mining Ecological and Environmental Data : www.indexmed.eu (Financement 2016 - MI du CNRS)
- VIGI-GEEK - VIsualisation of Graph In transdisciplinary Global Ecology, Economy and Sociology data-Kernel (Défi Imag'in 2015, Financement MI du CNRS)
- CHARLIE - CHAnger de Regard En Liant dans IndexMed l'Environnement et les Etoiles (PEPS Blanc, Financement INEE - CNRS)

Ce consortium a organisé de nombreux ateliers et notamment 4 rencontres annuelles cofinancées par les programmes présentées ci-dessus et différents organismes sur présentation d'un dossier de demande de subvention :

- **2018** (décalé de 2017) : Séminaire des SAGES (Sciences and Algorithms around Graphs in Environment and Societies) - financements FRB, OHM Littoral

Méditerranéen, OHM Bassin Minier de Provence, Labex OTMed, Labex DRIIHM, CNRS GDR MaDICS, https://indexmeed2017.sciencesconf.org/

- **2016** : GRAphs and datamIning for environmentaL research - Data, Research questions and New hypotheses (GRAIL days - journées du GRAAL) financements FRB, OSU Pytheas, OHM Littoral Méditerranéen, Labex OTMed, https://indexmed2016.sciencesconf.org/
- **2015** : Méthodes et outils pour la fouille de données hétérogènes et multi-sources en écologie, financements IMBE, OSU Pytheas, GBIF France, OHM Littoral Méditerranéen, CNRS, http://www.indexmed.eu/-Deuxieme-seminaire-Methodes-et-.html
- **2014** : financements IMBE, OSU Pytheas, GBIF France, OHM Littoral Méditerranéen, CNRS, Interopérabilité des bases de données en écologie, http://www.indexmed.eu/-Premier-seminaire-Interoperabilite-.html

Autres implications dans des programmes de recherche :

EVACOR et EVACOR2 - EVAluation des services écosystémiques des habitats CORalligènes (1 et 2)

Annexe 3 : Programme CIGESMED (Féral *et al.* 2016)

Présentation

Il existe très peu de programmes et de réseaux pour la surveillance des habitats coralligènes. CIGESMED (indicateurs coralligènes basés sur l'évaluation et le suivi du « bon état écologique » des eaux côtières méditerranéennes[130]) est un programme européen « ERANET » de recherche à visée d'aide à la gestion des habitats coralligènes, écosystèmes marins patrimoniaux méditerranéens qui a été le cadre principal de mes travaux de thèse. Ce programme appuie la mise en œuvre de la directive 2008/56 / CE du Parlement européen et du Conseil du 17 juin 2008. Elle participe à la mise en place d'un cadre d'action communautaire pour les acteurs dans le domaine de la politique environnementale marine (MSFD), et apporte une contribution en termes de connaissances pour les descripteurs 1 (diversité biologique), 2 (espèces non indigènes) et 6 (intégrité des fonds marins). Il a impliqué trois pays (France, Grèce et Turquie) de 2013 à 2016 (Figure 56).

Figure 56 : laboratoires investis dans la réalisation du programme CIGESMED

Dans ce projet, une approche écologique (développement et test d'indicateurs, prélèvements en plongée sous-marine, transects et techniques « *visual sensus* », analyses de photographies) est couplée au développement d'approches génétiques (barcoding,

[130] http://www.cigesmed.eu

metabarcoding, phylogéographie[131] , génétique des populations). Ces approches permettront de décrire la composition en espèces d'algues rouges bio-constructrices de ce milieu. L'approche génétique vise également à fournir des outils innovants pour la bio-indication, basés sur la diversité génétique intraspécifique d'un panel d'espèces. Utilisant les technologies de séquençage nouvelle génération (Next-Generation Sequencing, NGS), ce projet caractérisera, par metabarcoding, la composition en organismes (plusieurs phylums d'animaux et algues) de différents profils écologiques de coralligène.

Les objectifs de CIGESMED étaient (1) de combler les lacunes concernant les connaissances scientifiques actuelles sur les habitats coralligènes qui seraient les plus utiles pour formuler des recommandations pour les protéger ; à cette fin, lors de la définition de ce programme, il a été décidé de développer les approches de « barcoding », d'améliorer l'identification fiable des espèces clefs pour la matrice biogène à des fins de conservation et de protection (en considérant leur sensibilité aux espèces envahissantes et cryptiques) et en étudiant la structuration génétique et le potentiel de dispersion des espèces clés / habitat, (2) améliorer les connaissances sur les populations coralligènes en définissant à large échelle des états de référence et en établissant un réseau de spécialistes méditerranéens (ceci en espérant initier des séries à long terme et à large échelle), (3) animer ces réseaux de scientifiques, les gérer localement et les coordonner à l'échelle régionale, normaliser les protocoles qui pourraient être appliqués à l'ensemble de la Méditerranée et les indices et indicateurs applicables aux habitats coralligènes. Cette normalisation a permis de les comparer avec les données d'origine moléculaires (Phylogénie / génétique des populations) et d'évaluer l'efficacité relative de chacun d'entre eux comme outils de suivi du GES de la mer côtière méditerranéenne, (5) mettre en place un réseau «science citoyenne» et (6) utiliser les approches par les graphes comme outils pour trier, organiser et fouiller les grands ensembles de données hétérogènes produites et (7) développer un système d'information utilisable à différents niveaux par des scientifiques, des décideurs, des gestionnaires de l'environnement et par le grand public.

[131] Étude des processus qui expliquent la distribution des lignées généalogiques au sein de la même espèce (processus allant éventuellement jusqu'à la spéciation).

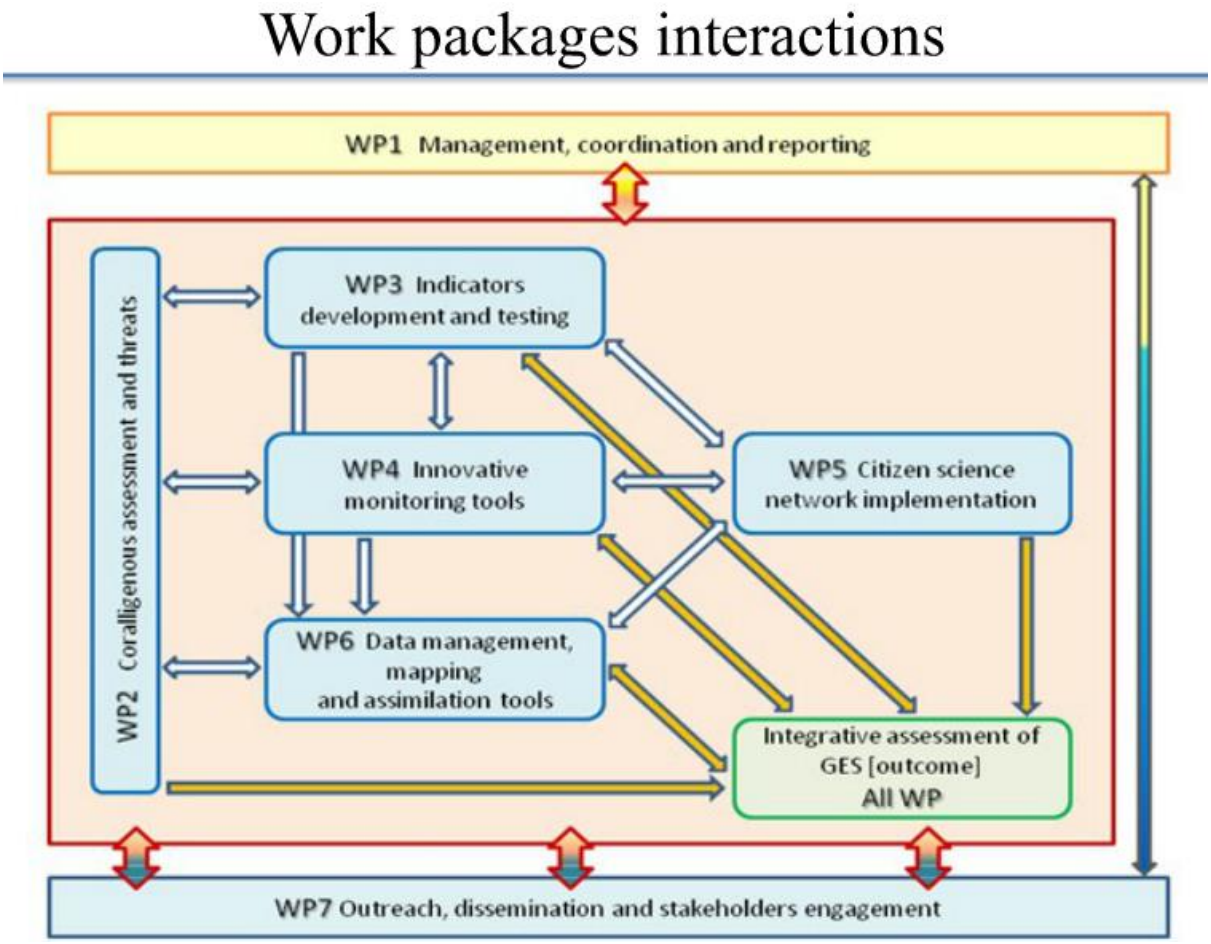


Figure 57 : interactions entre Work Packages dans le cadre de CIGESMED

Comme le montre la figure 57, ces différents objectifs, organisés en Work Packages sont fortement liés, et de nombreuses interactions ont été nécessaires notamment entre les travaux concernant la conception du système d'information / la gestion des données et les autres work packages.

Librairie de taxons utilisée avec photoquad pour l'analyse des quadrats photo CIGESMED en France

Species name	Species ID	Group name	Group ID
indeterminate	158	indeterminate	0
Chlorophyta indet.	173	Chlorophyta	1
Rhodophyta indet.	175	Rhodophyta	2
Calcareous Rhodophyta indet.	176	Calcareous Rhodophyta	3
Phaeophyta indet.	120	Phaeophyta	4
Porifera indet.	146	Porifera	6
Cnidaria indet.	40	Cnidaria	7
Hydraire indet.	162	Cnidaria	7
Cnidaria Octocorallia indet.	179	Cnidaria Octocorallia	8
Cnidaria Madreporia indet.	180	Cnidaria Madreporia	9
Madreporaires coloniaux indet.	100	Cnidaria Madreporia	9
Annelida Polychaeta indet.	177	Annelida Polychaeta	10
Serpulidae indet.	142	Annelida Polychaeta	10
Echiuridae indet.	178	Echiuridae	11
Encr bryozoa indet.	63	Encrusting Bryozoa	12
Erect bryozoa indet.	22	Erect Bryozoa	13
Ascidiacea indet.	160	Ascidiacea	14
Echinodermata indet.	164	Echinodermata	15
Crustacea indet.	165	Crustacea	18
Nudibranchia indet.	171	Nudibranchia	19
Foraminifera indet.	168	Foraminifera	20
Mollusca indet.	181	Mollusca	21
Acanthella acuta	1	Porifera	6
Adeonella calveti	2	Erect Bryozoa	13
Agelas oroides	3	Porifera	6

Aglaophenia spp.	4	Cnidaria	7
Aiptasia mutabilis	5	Cnidaria	7
Alcyonium spp.	6	Cnidaria Octocorallia	8
Alcyonium coralloides	7	Cnidaria Octocorallia	8
Antedon mediterraneus	8	Echinodermata	15
Astrospartus mediterranea	9	Echinodermata	15
Aplidium fuscum	10	Ascidiacea	14
Aplidium spp.	11	Ascidiacea	14
Aplysilla sulfurea	12	Porifera	6
Aplysina cavernicola	13	Porifera	6
Axinella damicornis	14	Porifera	6
Axinella polypoides	15	Porifera	6
Axinella verrucosa	16	Porifera	6
Balanophyllia europea	17	Cnidaria Madreporia	9
Bare rock	18	Abiotic	16
Beania hirtissima *cf.* cylindrica	19	Erect Bryozoa	13
Bispira volutacornis	20	Annelida Polychaeta	10
Bonellia viridis	21	Echiuridae	11
Cacospongia spp.	23	Porifera	6
Caryophyllia inornata	24	Cnidaria Madreporia	9
Caryophyllia smithii	25	Cnidaria Madreporia	9
Ceberea boryi	26	Erect Bryozoa	13
Cellaria sp.	27	Erect Bryozoa	13
Centrostephanus longispinus	28	Echinodermata	15
Cerianthus spp.	29	Cnidaria	7
Cereus pedunculatus	30	Cnidaria	7
Chaetaster longipes	31	Echinodermata	15
Chondrosia reniformis	32	Porifera	6
Chrysimenia ventricosa	33	Rhodophyta	2

Cidaris cidaris	34	Echinodermata	15
Ciona spp.	35	Ascidiacea	14
Cladocora caespitosa	36	Cnidaria Madreporia	9
Clathrina spp.	37	Porifera	6
Clavelina spp.	38	Ascidiacea	14
Cliona viridis	39	Porifera	6
Codium bursa	41	Chlorophyta	1
Codium effusum	42	Chlorophyta	1
Corallium rubrum	43	Cnidaria Octocorallia	8
Corticum candelabrum	44	Porifera	6
Corynactis viridis	45	Cnidaria	7
Crambe crambe	46	Porifera	6
Crella pulvinar	47	Porifera	6
Cribrinopsis crassa	48	Cnidaria	7
Cystodytes dellechiazei	49	Ascidiacea	14
Cystoseira zosteroides	50	Phaeophyta	4
Cystoseira spp.	51	Phaeophyta	4
Organic Detritus	52	Abiotic	16
Dentiporella sardonica	53	Encrusting Bryozoa	12
Diazona violacea	54	Ascidiacea	14
Dictyonella sp.	55	Porifera	6
Dictyopteris polypodioides	56	Phaeophyta	4
Didemnum drachi	57	Ascidiacea	14
Diporula verrucosa	58	Erect Bryozoa	13
Peltodoris atromaculata	59	Nudibranchia	19
Dysidea avara	60	Porifera	6
Echinaster sepositus	61	Echinodermata	15
Echinus melo	62	Echinodermata	15
Eudendrium spp.	64	Cnidaria	7

Eunicella cavolini	65	Cnidaria Octocorallia	8
Eunicella singularis	66	Cnidaria Octocorallia	8
Eunicella verrucosa	67	Cnidaria Octocorallia	8
Eupolymnia nebulosa	68	Annelida Polychaeta	10
Flabellina affinis	69	Nudibranchia	19
Flabellia petiolata	70	Chlorophyta	1
Frondipora lichenoides	71	Erect Bryozoa	13
Galathea strigosa	72	Crustacea	18
Gloiocladia repens	73	Rhodophyta	2
Gregarinidra gregaria	74	Encrusting Bryozoa	12
Hacelia attenuata	75	Echinodermata	15
Haliclona fulva	76	Porifera	6
Haliclona mediterranea	77	Porifera	6
Haliclona mucosa	78	Porifera	6
Halimeda tuna	79	Chlorophyta	1
Halocynthia papillosa	80	Ascidiacea	14
Halopteris filicina	81	Phaeophyta	4
Halymenia elongata	82	Rhodophyta	2
Hemimycale columella	83	Porifera	6
Hexadella pruvoti	84	Porifera	6
Hexadella racovitzai	85	Porifera	6
Holothuria tubulosa	86	Echinodermata	15
Holothuria polii	87	Echinodermata	15
Holothuria forskali	88	Echinodermata	15
Homarus gammarus	89	Crustacea	18
Hoplangia durothrix	90	Cnidaria Madreporia	9
Hornera lichenoides	91	Erect Bryozoa	13
Hypselodoris picta	92	Nudibranchia	19
Hypselodoris spp.	93	Nudibranchia	19

Idmidronea sp.	94	Erect Bryozoa	13
Ircinia spp.	95	Porifera	6
Janolus cristatus	96	Nudibranchia	19
Leptogorgia sarmentosa	97	Cnidaria Octocorallia	8
Leptosamnia pruvoti	98	Cnidaria Madreporia	9
Madracis pharensis	99	Cnidaria Madreporia	9
Marstasteria glacialis	101	Echinodermata	15
Mesophyllum foliaces	102	Calcareous Rhodophyta	3
Lithophyllum spp. Foliacees	103	Calcareous Rhodophyta	3
Miniacina miniacea	104	Foraminifera	20
Myriapora truncata	105	Erect Bryozoa	13
Ophiothrix fragilis	106	Echinodermata	15
Oscarella lobularis	107	Porifera	6
Oscarella tuberculata	108	Porifera	6
Oscarella spp.	109	Porifera	6
Palinurus elephas	110	Crustacea	18
Palmophyllum crassum	111	Chlorophyta	1
Paramuricea clavata	112	Cnidaria Octocorallia	8
Parazoanthus axinellae	113	Cnidaria	7
Pentapora fascialis	114	Erect Bryozoa	13
Petrosia ficiformis	115	Porifera	6
Peyssonnelia foliacee	116	Rhodophyta	2
Peyssonnelia encroutant	117	Rhodophyta	2
Phallusia fumigata	118	Ascidiacea	14
Phallusia mamillata	119	Ascidiacea	14
Phorbas tenacior	121	Porifera	6
Phyllangia mouchetzi	122	Cnidaria Madreporia	9
Phyllariopsis brevipes	123	Phaeophyta	4
Pleraplysilla spinifera	124	Porifera	6

Pleurobranchus testudinarius	125	Nudibranchia	19
Polysyncraton spp.	126	Ascidiacea	14
Protula tubularia	127	Annelida Polychaeta	10
Raspaciona aculeata	128	Porifera	6
Reteporella grimaldii	129	Erect Bryozoa	13
Rhynchozoon neapolitanum	130	Encrusting Bryozoa	12
Sabella spallanzanii	131	Annelida Polychaeta	10
Salmacina spp./Filograna implexa	132	Annelida Polychaeta	10
Scalarospongia/Sarcotragus	133	Porifera	6
Schizomavella spp.	134	Encrusting Bryozoa	12
Schizotheca serratimargo	135	Erect Bryozoa	13
Scrupocellaria sp.	136	Erect Bryozoa	13
Scyllarides latus	137	Crustacea	18
Scyllarus arctus	138	Crustacea	18
Sphaerechinus granularis	139	Echinodermata	15
Sphaerococcus coronopifolius	140	Rhodophyta	2
Sediment	141	Abiotic	16
Shadow/holes	143	indeterminate	0
Smittina cervicornis	144	Erect Bryozoa	13
Spirastrella cunctatrix	145	Porifera	6
Spongia lamella	147	Porifera	6
Spongia/Hippospongia spp.	148	Porifera	6
Stoloniferes	149	Cnidaria Octocorallia	8
Stylocidaris affinis	150	Echinodermata	15
Suberites sp.	151	Porifera	6
Tethya aurantium	152	Porifera	6
Turbicellepora avicularis	153	Erect Bryozoa	13
Turf	154	TURF	5
Umbraculum umbraculum	155	Nudibranchia	19

Valonia macrophysa	156	Chlorophyta	1
Zanardinia typus	157	Phaeophyta	4
Macrodechet	159	Abiotic	16
Didemnidae	161	Ascidiacea	14
Lithophyllum spp. Encroutant	169	Calcareous Rhodophyta	3
Mesophyllum spp. Encroutant	170	Calcareous Rhodophyta	3
Polyclinum aurantium	172	Ascidiacea	14
Callistoma spp.	182	Mollusca	21
Luria lurida	183	Mollusca	21
Euthria spp.	184	Mollusca	21
Bittium spp.	185	Mollusca	21
Limoida spp.	186	Mollusca	21
Pterioida spp.	187	Mollusca	21
Pycnoclavella spp.	188	Ascidiacea	14
Dysidea spp.	189	Porifera	6
Haliclona aquaeductus	190	Porifera	6
Cratena peregrina	191	Nudibranchia	19
Felimare orsinii	192	Nudibranchia	19
Aplysia spp.	193	Nudibranchia	19
Thuridilla hopei	194	Nudibranchia	19
Flabellina spp.	195	Nudibranchia	19
Flabellina ischitana	196	Nudibranchia	19
Flabellina pedata	197	Nudibranchia	19
Felimare spp.	198	Nudibranchia	19
Felimare picta	199	Nudibranchia	19
Peyssonelia spp.	200	Rhodophyta	2
Pseudobiceros splendidus	201	Platyhelminthe	22
Yungia aurantica	202	Platyhelminthe	22
Prostheceraeus roseus	203	Platyhelminthe	22

Prostheceraeus vittatus	204	Platyhelminthe	22
Prostheceraeus giesbrechtii	205	Platyhelminthe	22
Prostheceraeus moseleyi	206	Platyhelminthe	22
Prostheceraeus spp.	207	Platyhelminthe	22
Diffuse image	208	indeterminate	0
Jujubinus striatus	209	Mollusca	21
Spongia officinalis	210	Porifera	6
Dysidea fragilis	211	Porifera	6
Reteporella spp.	212	Erect Bryozoa	13
Holothuria spp.	213	Echinodermata	15
Aplysilla rosea	215	Porifera	6
Alcyonidium spp.	216	Erect Bryozoa	13
Padina pavonica	163	Phaeophyta	4
Arbacia lixula	166	Echinodermata	15

Données prétraitées (extrait)

L'ensemble des données CIGESMED est téléchargeable sur le site internet :
http://www.cigesmed.eu/-Data-upload-

IMAGE	SITE_TRANSECT	OPE	OBS	LUM	CADRE	CAMERA	METHOD	SP_NOM	SP_ID	PTS_SP	PCT_SP	REF_PTS	GROUP_ID	GROUP_NOM	LEVEL_DETER	REGROUP
CIGESMED_FTF_20140115_D28_T01_Q01_FZ	FTF_T01	LD	FZ	100%	simple	NikonD300s	permanent_continu	Axinella damicornis	14	4	4	100	6	Porifera	sp	Porifera
CIGESMED_FTF_20140115_D28_T01_Q01_FZ	FTF_T01	LD	FZ	100%	simple	NikonD300s	permanent_continu	Axinella verrucosa	16	1	1	100	6	Porifera	sp	Porifera
CIGESMED_FTF_20140115_D28_T01_Q01_FZ	FTF_T01	LD	FZ	100%	simple	NikonD300s	permanent_continu	Organic Detritus	52	24	24	100	16	Abiotic	abio	Abiotique
CIGESMED_FTF_20140115_D28_T01_Q01_FZ	FTF_T01	LD	FZ	100%	simple	NikonD300s	permanent_continu	Eunicella cavolini	65	2	2	100	8	Cnidaria Octocorallia	sp	Cnidaria
CIGESMED_FTF_20140115_D28_T01_Q01_FZ	FTF_T01	LD	FZ	100%	simple	NikonD300s	permanent_continu	Haliclona fulva	76	3	3	100	6	Porifera	sp	Porifera
CIGESMED_FTF_20140115_D28_T01_Q01_FZ	FTF_T01	LD	FZ	100%	simple	NikonD300s	permanent_continu	Hoplangia durothrix	90	1	1	100	9	Cnidaria Madreporia	sp	Cnidaria
CIGESMED_FTF_20140115_D28_T01_Q01_FZ	FTF_T01	LD	FZ	100%	simple	NikonD300s	permanent_continu	Leptosamnia pruvoti	98	3	3	100	9	Cnidaria Madreporia	sp	Cnidaria
CIGESMED_FTF_20140115_D28_T01_Q01_FZ	FTF_T01	LD	FZ	100%	simple	NikonD300s	permanent_continu	Myriapora truncata	105	3	3	100	13	Erect Bryozoa	sp	Bryozoa
CIGESMED_FTF_20140115_D28_T01_Q01_FZ	FTF_T01	LD	FZ	100%	simple	NikonD300s	permanent_continu	Oscarella spp.	109	17	17	100	6	Porifera	genre	Porifera
CIGESMED_FTF_20140115_D28_T01_Q01_FZ	FTF_T01	LD	FZ	100%	simple	NikonD300s	permanent_continu	Peyssonnelia foliacée	116	2	2	100	2	Rhodophyta	sp	Rhodophyta
CIGESMED_FTF_20140115_D28_T01_Q01_FZ	FTF_T01	LD	FZ	100%	simple	NikonD300s	permanent_continu	Peysonnelia encroutant	117	1	1	100	2	Rhodophyta	genre	Rhodophyta
CIGESMED_FTF_20140115_D28_T01_Q01_FZ	FTF_T01	LD	FZ	100%	simple	NikonD300s	permanent_continu	Schizomavella spp.	134	1	1	100	12	Encrusting Bryozoa	genre	Bryozoa

Annexe 4 : Programme DEVOTES (Borja, 2017)

Présentation

DEVOTES est un projet collaboratif coordonné par AZTI-Tecnalia, Pasaia (Espagne) qui a été financé pendant 4 ans (2012-2016), majoritairement par le 7ème programme-cadre de l'Union européenne. La présentation qui suit est un résumé de la présentation officielle du programme (http://www.devotes-project.eu/)

L'objectif global de DEVOTES était d'améliorer la compréhension des relations entre les pressions exercées par les activités humaines et les influences climatiques et leurs effets induits sur les écosystèmes marins, y compris la diversité biologique. Ces nouvelles connaissances doivent permettre d'améliorer la gestion écosystémique et d'atteindre le bon état écologique des écosystèmes marins et des eaux « marines ».

Il a impliqué 23 partenaires de 15 pays de l'UE, dont deux partenaires non européens (d'Arabie Saoudite et d'Ukraine) et quatre PME, ainsi que deux observateurs (EPA et N.O.A.A.) des États-Unis. Un groupe de scientifiques indépendants formait le Conseil consultatif (AB) qui a fourni des orientations stratégiques et contrôlait que les résultats du projet atteignent les objectifs.

Les principaux objectifs de DEVOTES étaient les suivants: i) améliorer notre compréhension de l'impact des activités humaines et du changement climatique sur la biodiversité marine; ii) identifier les obstacles et les goulets d'étranglement qui empêchent l'établissement d'un bon état écologique; iii) tester des indicateurs et développer de nouveaux indicateurs innovants pour évaluer la biodiversité de manière harmonisée dans les 4 mers régionales concernées par ce programme; iv) développer, tester et valider des outils novateurs de modélisation et de suivi intégratifs pour améliorer notre compréhension des changements dans les écosystèmes et la biodiversité, pour les intégrer dans une évaluation unique et holistique; v) proposer et diffuser des stratégies et des mesures pour la gestion adaptative des écosystèmes, en tenant compte du rôle actif de l'industrie et d'autres parties prenantes économiques.

DEVOTES a relevé trois défis principaux dans la détermination du statut environnemental : (i) l'évaluation des pressions anthropiques, y compris le changement climatique, qui ont un impact sur la biodiversité; (ii) la sélection d'indicateurs appropriés pour évaluer le statut; et (iii) l'intégration de ces indicateurs à travers un certain nombre d'échelles écologiques, pour une évaluation intégrative unique de la biodiversité.

Les objectifs du programme DEVOTES s'appuyaient sur les activités menées dans sept « work packages » opérationnels, en plus du management: Pressions Humaines et

Changement Climatique (WP1), Implications socio-économiques pour réaliser le GES (WP2), Test et développement d'indicateurs (WP3) Outils de modélisation innovants (WP4); Techniques de surveillance innovantes (WP5); Évaluation intégrative de la biodiversité (WP6); Sensibilisation, engagement des parties prenantes et diffusion des produits (WP7).

Librairie de taxons utilisée avec PhotoQuad pour l'analyse des quadrats photo DEVOTES

Kingdom	Phylum	Class	Order	Family	Species
Plantae	Chlorophyta	Ulvophyceae	Dasycladales	Polyphysaceae	*Acetabularia acetabulum* (Linnaeus) P.C. Silva, 1952
Plantae	Chlorophyta	Ulvophyceae	Bryopsidales	Caulerpaceae	Caulerpa cylindracea Sonder, 1845
Plantae	Chlorophyta	Ulvophyceae	Bryopsidales	Codiaceae	*Codium bursa* (Olivi) C.Agardh, 1817
Plantae	Chlorophyta	Ulvophyceae	Bryopsidales	Codiaceae	*Codium coralloides* (Kützing) P.C. Silva, 1960
Plantae	Chlorophyta	Ulvophyceae	Bryopsidales	Codiaceae	*Codium effusum* (Rafinesque) Delle Chiaje, 1829
Plantae	Chlorophyta	Ulvophyceae	Bryopsidales	Udoteaceae	*Flabellia petiolata* (Turra) Nizamuddin, 1987
Plantae	Chlorophyta	Ulvophyceae	Bryopsidales	Halimedaceae	*Halimeda tuna* (J. Ellis & Solander) J.V. Lamouroux, 1816
Plantae	Chlorophyta	Incertae sedis	Palmophyllales	Palmophyllaceae	*Palmophyllum crassum* (Naccari) Rabenhorst, 1868
Plantae	Rhodophyta	Florideophyceae	Rhodymeniales	Rhodymeniaceae	*Chrysymenia ventricosa* (J.V. Lamouroux) J. Agardh, 1842
Plantae	Rhodophyta	Florideophyceae	Corallinales	Corallinaceae	*Lithophyllum* sp. Philippi, 1837
Plantae	Rhodophyta	Florideophyceae	Corallinales	Hapalidiaceae	*Mesophyllum* sp. Me. Lemoine, 1928
Plantae	Rhodophyta	Florideophyceae	Ceramiales	Rhodomelaceae	*Osmundaria volubilis* (Linnaeus) R.E. Norris, 1991
Plantae	Rhodophyta	Florideophyceae	Peyssonneliales	Peyssonneliaceae	*Peyssonnelia* sp. Decaisne, 1841
Plantae	Rhodophyta	Florideophyceae	Gigartinales	Phyllophoraceae	*Phyllophora* sp. Greville, 1830
Plantae	Rhodophyta	Florideophyceae	Gigartinales	Sphaerococcaceae	Sphaerococcus coronopifolius Stackhouse, 1797
Plantae	Rhodophyta	Florideophyceae	Ceramiales	Rhodomelaceae	*Womersleyella setacea* (Hollenberg) R.E. Norris, 1992
Plantae	Ochrophyta	Phaeophyceae	Dictyotales	Dictyotaceae	*Dictyopteris polypodioides* (A.P. De Candolle) J.V. Lamouroux, 1809
Plantae	Ochrophyta	Phaeophyceae	Dictyotales	Dictyotaceae	*Dictyota dichotoma* (Hudson) J.V. Lamouroux, 1809
Plantae	Ochrophyta	Phaeophyceae	Dictyotales	Dictyotaceae	*Dictyota fasciola* (Roth) J.V. Lamouroux, 1809
Plantae	Ochrophyta	Phaeophyceae	Sphacelariales	Stypocaulaceae	*Halopteris filicina* (Grateloup) Kützing, 1843
Plantae	Ochrophyta	Phaeophyceae	Dictyotales	Dictyotaceae	*Padina pavonica* (Linnaeus) Thivy, 1960
Plantae	Ochrophyta	Phaeophyceae	Tilopteridales	Phyllariaceae	*Phyllariopsis brevipes* (C. Agardh) E.C. Henry & G.R. South, 1987
Plantae	Ochrophyta	Phaeophyceae	Cutleriales	Cutleriaceae	*Zanardinia typus* (Nardo) P.C. Silva, 2000
Animalia	Porifera	Demospongiae	Bubarida	Dyctionellidae	Acanthella acuta Schmidt, 1862
Animalia	Porifera	Demospongiae	Agelasida	Agelasidae	*Agelas oroides* (Schmidt, 1864)
Animalia	Porifera	Demospongiae	Dendroceratida	Darwinellidae	Aplysilla rosea (Barrois, 1876)
Animalia	Porifera	Demospongiae	Dendroceratida	Darwinellidae	Aplysilla sulfurea Schultze, 1878
Animalia	Porifera	Demospongiae	Verongiida	Aplysinidae	Aplysina cavernicola (Vacelet, 1959)
Animalia	Porifera	Demospongiae	Axinellida	Axinellidae	Axinella damicornis (Esper, 1794)
Animalia	Porifera	Demospongiae	Axinellida	Axinellidae	Axinella polypoides Schmidt, 1862
Animalia	Porifera	Demospongiae	Axinellida	Axinellidae	*Axinella* sp. Schmidt, 1862
Animalia	Porifera	Demospongiae	Axinellida	Axinellidae	Axinella vaceleti Pansini, 1984
Animalia	Porifera	Demospongiae	Axinellida	Axinellidae	Axinella verrucosa (Esper, 1794)
Animalia	Porifera	Demospongiae	Dictyoceratida	Thorectidae	*Cacospongia* sp. Schmidt, 1862
Animalia	Porifera	Demospongiae	Chondrillida	Chondrillidae	Chondrilla nucula Schmidt, 1862
Kingdom	Phylum	Class	Order	Family	Species
Animalia	Porifera	Demospongiae	Chondrisiida	Chondrisiidae	Chondrosia reniformis Nardo, 1847
Animalia	Porifera	Calcarea	Clathrinida	Clathrinidae	*Clathrina* sp. Gray, 1867
Animalia	Porifera	Demospongiae	Clionaida	Clionaidae	Cliona celata Grant, 1826
Animalia	Porifera	Demospongiae	Clionaida	Clionaidae	Cliona schmidti (Ridley, 1881)
Animalia	Porifera	Demospongiae	Clionaida	Clionaidae	*Cliona viridis* Schmidt, 1826
Animalia	Porifera	Demospongiae	Poecilosclerida	Crambeidae	*Crambe crambe* (Schmidt, 1862)
Animalia	Porifera	Demospongiae	Poecilosclerida	Crellidae	Crella (Grayella) pulvinar (Schmidt, 1868)
Animalia	Porifera	Demospongiae	Bubarida	Dictyonellidae	*Dictyonella* sp. Schmidt, 1868
Animalia	Porifera	Demospongiae	Dictyoceratida	Dysideidae	*Dysidea* sp. Jonhston, 1842
Animalia	Porifera	Demospongiae	Tetractinellida	Geodiidae	Erylus deficiens Topsent, 1927
Animalia	Porifera	Demospongiae	Haplosclerida	Chalinidae	Haliclona (Halichoclona) fulva (Topsent, 1893)
Animalia	Porifera	Demospongiae	Haplosclerida	Chalinidae	Haliclona (Reniera) mediterranea Griessinger, 1971
Animalia	Porifera	Demospongiae	Haplosclerida	Chalinidae	Haliclona (Soestella) mucosa (Griessinger, 1971)
Animalia	Porifera	Demospongiae	Haplosclerida	Chalinidae	*Haliclona* sp. Grant, 1836
Animalia	Porifera	Demospongiae	Poecilosclerida	Hymedesmiidae	Hemimycale columella (Bowerbank, 1874)
Animalia	Porifera	Demospongiae	Verongiida	Ianthellidae	Hexadella pruvoti Topsent, 1896
Animalia	Porifera	Demospongiae	Verongiida	Ianthellidae	Hexadella racovitzai Topsent, 1896
Animalia	Porifera	Demospongiae	Verongiida	Ianthellidae	*Hexadella* sp. Topsent, 1896
Animalia	Porifera	Demospongiae	Dictyoceratida	Spongiidae	*Hippospongia* sp. Schulze, 1879
Animalia	Porifera	Demospongiae	Dictyoceratida	Irciniidae	*Ircinia* sp. Nardo, 1833
Animalia	Porifera	Demospongiae	Dictyoceratida	Irciniidae	Ircinia variabilis (Schmidt, 1862)
Animalia	Porifera	Demospongiae	Homosclerophorida	Oscarellidae	Oscarella lobularis (Schmidt, 1862)
Animalia	Porifera	Demospongiae	Homosclerophorida	Oscarellidae	*Oscarella* sp. Vosmaer, 1884
Animalia	Porifera	Demospongiae	Homosclerophorida	Oscarellidae	Oscarella tuberculata (Schmidt, 1868)
Animalia	Porifera	Demospongiae	Haplosclerida	Petrosiidae	Petrosia ficiformis (Poiret, 1789)

Animalia	Porifera	Demospongiae	Poecilosclerida	Hymedesmiidae	*Phorbas* sp. Duchassaing & Michelotti, 1864
Animalia	Porifera	Demospongiae	Poecilosclerida	Hymedesmiidae	Phorbas tenacior (Topsent, 1925)
Animalia	Porifera	Demospongiae	Dictyoceratida	Dysideidae	Pleraplysilla spinifera (Schulze, 1879)
Animalia	Porifera	Demospongiae	Axinellida	Raspailiidae	Raspaciona aculeata (Jonhston, 1842)
Animalia	Porifera	Demospongiae	Dictyoceratida	Thorectidae	*Scalarispongia* sp. Cook & Bergquist, 2000
Animalia	Porifera	Demospongiae	Dictyoceratida	Irciniidae	*Sarcotragus* sp. Schmidt, 1862
Animalia	Porifera	Demospongiae	Clionaida	Spirastrellidae	Spirastrella cunctatrix Schmidt, 1868
Animalia	Porifera	Demospongiae	Dictyoceratida	Spongiidae	Spongia (Spongia) lamella (Schulze, 1879)
Animalia	Porifera	Demospongiae	Dictyoceratida	Spongiidae	Spongia (Spongia) officinalis Linnaeus, 1759
Animalia	Porifera	Demospongiae	Dictyoceratida	Spongiidae	*Spongia* sp. Linnaeus, 1759
Animalia	Porifera	Demospongiae	Suberitida	Suberitidae	*Suberites* sp. Nardo, 1833
Animalia	Porifera	Demospongiae	Tethyida	Tethyidae	Tethya aurantium (Pallas, 1766)
Animalia	Cnidaria	Hydrozoa	Leptothecata	Aglaopheniidae	Aglaophenia elongata Meneghini, 1845
Animalia	Cnidaria	Hydrozoa	Leptothecata	Aglaopheniidae	*Aglaophenia* sp. Lamouroux, 1812
Animalia	Cnidaria	Anthozoa	Actiniaria	Aiptasiidae	*Aiptasia mutabilis* (Gravenhorst, 1831)
Animalia	Cnidaria	Anthozoa	Alcyonacea	Acyoniidae	Alcyonium coralloides (Pallas, 1766)
Animalia	Cnidaria	Anthozoa	Alcyonacea	Acyoniidae	*Alcyonium* sp. Pallas, 1766
Animalia	Cnidaria	Anthozoa	Actinaria	Aliciidae	Alicia mirabilis Jonhson, 1861
Kingdom	Phylum	Class	Order	Family	Species
Animalia	Cnidaria	Anthozoa	Actinaria	Actiniidae	Anemonia viridis (Forsskål, 1775)
Animalia	Cnidaria	Anthozoa	Scleractinia	Dendrophylliidae	Balanophyllia (Balanophyllia) europaea (Risso, 1826)
Animalia	Cnidaria	Anthozoa	Scleractinia	Caryophylliidae	Caryophyllia (Caryophyllia) inornata (Duncan, 1878)
Animalia	Cnidaria	Anthozoa	Scleractinia	Caryophylliidae	Caryophyllia (Caryophyllia) smithii Stokes & Broderip, 1828
Animalia	Cnidaria	Anthozoa	Actinaria	Sagartiidae	Cereus pedunculatus (Pennant, 1777)
Animalia	Cnidaria	Anthozoa	Spirularia	Cerianthidae	*Cerianthus* sp. Della Chiaje, 1830
Animalia	Cnidaria	Anthozoa	Scleractinia	Scleractinia incertae sedis	Cladocora caespitosa (Linnaeus, 1767)
Animalia	Cnidaria	Anthozoa	Alcyonacea	Clavulariidae	*Clavularia* sp. Blainville, 1830
Animalia	Cnidaria	Anthozoa	Alcyonacea	Coralliidae	Corallium rubrum (Linnaeus, 1758)
Animalia	Cnidaria	Anthozoa	Actinaria	Actiniidae	Cribrinopsis crassa (Andrès, 1881)
Animalia	Cnidaria	Anthozoa	Anthoathecata	Eudendriidae	*Eudendrium* sp. Ehrenberg, 1834
Animalia	Cnidaria	Anthozoa	Alcyonacea	Gorgoniidae	Eunicella cavolini (Koch, 1887)
Animalia	Cnidaria	Anthozoa	Alcyonacea	Gorgoniidae	Eunicella singularis (Esper, 1791)
Animalia	Cnidaria	Anthozoa	Alcyonacea	Gorgoniidae	Eunicella verrucosa (Pallas, 1766)
Animalia	Cnidaria	Anthozoa	Scleractinia	Caryophylliidae	Hoplangia durotrix Gosse, 1860
Animalia	Cnidaria	Anthozoa	Alcyonacea	Gorgoniidae	Leptogorgia sarmentosa (Esper, 1789)
Animalia	Cnidaria	Anthozoa	Scleractinia	Dendrophylliidae	*Leptopsammia pruvoti* Lacaze-Duthiers, 1897
Animalia	Cnidaria	Anthozoa	Scleractinia	Astrocoeniidae	Madracis pharensis (Heller, 1868)
Animalia	Cnidaria	Anthozoa	Alcyonacea	Gorgoniidae	Paramuricea clavata (Risso, 1826)
Animalia	Cnidaria	Anthozoa	Scleractinia	Caryophyllidae	Paracyathus pulchellus (Philippi, 1842)
Animalia	Cnidaria	Anthozoa	Zoantharia	Parazoanthidae	Parazoanthus axinellae (Schmidt, 1862)
Animalia	Cnidaria	Anthozoa	Scleractinia	Caryophylliidae	Phyllangia americana mouchezii (Lacaze-Duthiers, 1897)
Animalia	Cnidaria	Anthozoa	Actinaria	Phymanthidae	Phymanthus pulcher (Andrès, 1883)
Animalia	Platyhelminthes	Rhabditophora	Polycladidia	Euryleptidae	Prostheceraeus giesbrechtii Lang, 1884
Animalia	Annelida	Polychaeta	Sabellida	Sabellidae	Bispira volutacornis (Montagu, 1804)
Animalia	Annelida	Polychaeta	Echiuroidea	Bonellidae	Bonellia viridis Rolando, 1821
Animalia	Annelida	Polychaeta	Terebellida	Terebellidae	Eupolymnia nebulosa (Montagu, 1818)
Animalia	Annelida	Polychaeta	Sabellida	Serpulidae	Filograna implexa Berkeley, 1835
Animalia	Annelida	Polychaeta	Sabellida	Serpulidae	Protula tubularia (Montagu, 1803)
Animalia	Annelida	Polychaeta	Sabellida	Sabellidae	Sabella pavonina Savigny, 1822
Animalia	Annelida	Polychaeta	Sabellida	Sabellidae	Sabella spallanzanii (Gmelin, 1791)
Animalia	Annelida	Polychaeta	Sabellida	Serpulidae	*Salmacina* sp. Claparède, 1870
Animalia	Annelida	Polychaeta	Sabellida	Serpulidae	*Serpula* sp. Linnaeus, 1758
Animalia	Annelida	Polychaeta	Sabellida	Serpulidae	Spirobranchus triqueter (Linnaeus, 1758)
Animalia	Mollusca	Gastropoda	Littorinimorpha	Apporrhaidae	Aporrhais pespelecani (Linnaeus, 1758)
Animalia	Mollusca	Gastropoda		Calliostomatidae	Calliostoma zizyphinum (Linnaeus, 1758)
Animalia	Mollusca	Gastropoda	Nudibranchia	Facelinidae	Cratena peregrina (Gmelin, 1791)
Animalia	Mollusca	Gastropoda	Nudibranchia	Chromodorididae	*Felimare picta* (Schulz in Philippi, 1836)
Animalia	Mollusca	Gastropoda	Nudibranchia	Chromodorididae	*Felimare* sp. Ev. Marcus & Er. Marcus, 1967
Animalia	Mollusca	Gastropoda	Nudibranchia	Flabellinidae	Flabellina affinis (Gmelin, 1791)
Animalia	Mollusca	Gastropoda	Nudibranchia	Flabellinidae	Flabellina pedata (Montagu, 1816)
Kingdom	Phylum	Class	Order	Family	Species
Animalia	Mollusca	Gastropoda	Nudibranchia	Chromodorididae	*Hypselodoris* sp. Stimpson, 1855
Animalia	Mollusca	Gastropoda	Nudibranchia	Proctonotidae	*Janolus cristatus (*Delle Chiaje, 1841)
Animalia	Mollusca	Bivalvia	Limida	Limidae	*Lima* sp. Bruguière, 1797
Animalia	Mollusca	Gastropoda	Nudibranchia	Discodorididae	Peltodoris atromaculata Bergh, 1880
Animalia	Mollusca	Gastropoda	Pleurobranchomorpha	Pleurobranchidae	Pleurobranchus testudinarius Cantraine, 1835
Animalia	Mollusca	Bivalvia	Pectinida	Spondylidae	Spondylus gaederopus Linnaeus, 1758
Animalia	Mollusca	Gastropoda	Sacoglossa	Plakobranchidae	Thuridilla hopei (Vérany, 1853)
Animalia	Mollusca	Gastropoda	Littorinimorpha	Vermetidae	Thylacodes arenarius (Linnaeus, 1758)
Animalia	Mollusca	Gastropoda	Umbraculida	Umbraculidae	Umbraculum umbraculum (Lightfoot, 1786)
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Adeonellidae	*Adeonella calveti* (Canu & Bassler, 1930)
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Beaniidae	Beania hirtissima cylindrica (Hincks, 1886)
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Candidae	*Caberea boryi* (Audouin, 1826)
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Cellariidae	*Cellaria* sp. Ellis & Solander, 1786

Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Phidoloporidae	Dentiporella sardonica (Waters, 1879)
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Microporellidae	Diporula verrucosa (Peach, 1868)
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Frondiporidae	Frondipora verrucosa (Lamouroux, 1821)
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Tubuliporidae	*Idmidronea atlantica* (Forbes, in Johnston, 1847)
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Myriaporidae	Myriapora truncata (Pallas, 1766)
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Bitectiporidae	Pentapora fascialis (Pallas, 1766)
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Phidoloporidae	*Reteporella* sp. Busk, 1884
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Phidoloporidae	*Rhynchozoon* sp. Hincks, 1895
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Bitectiporidae	Schizomavella (Schizomavella) mamillata (Hincks, 1880)
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Phidoloporidae	Schizoretepora serratimargo (Hincks, 1886)
Animalia	Bryozoa	Gymnolaemata	Cheilostomatida	Celleporidae	Turbicellepora avicularis (Hincks, 1860)
Animalia	Echinodermata	Crinoidea	Comatulida	Antedonidae	Antedon mediterranea (Lamarck, 1816)
Animalia	Echinodermata	Ophiuroidea	Euryalida	Gorgonocephalidae	Astrospartus mediterraneus (Risso, 1826)
Animalia	Echinodermata	Echinoidea	Diadematoida	Diadematidae	Centrostephanus longispinus (Philippi, 1845)
Animalia	Echinodermata	Asteroidea	Valvatida	Chaetasteridae	Chaetaster longipes (Retzius, 1805)
Animalia	Echinodermata	Echinoidea	Cidaroida	Cidaridae	*Cidaris cidaris* (Linnaeus, 1758)
Animalia	Echinodermata	Asteroidea	Forcipulatida	Asteriidae	Coscinasterias tenuispina (Lamarck, 1816)
Animalia	Echinodermata	Asteroidea	Spinulosida	Echinasteridae	Echinaster (Echinaster) sepositus (Retzius, 1783)
Animalia	Echinodermata	Echinoidea	Camarodonta	Echinidae	*Echinus melo* Lamarck, 1816
Animalia	Echinodermata	Asteroidea	Valvatida	Ophidiasteridae	Hacelia attenuata Gray, 1840
Animalia	Echinodermata	Holothuroidea	Aspidochirotida	Holothuriidae	Holothuria (Panningothuria) forskali Delle Chiaje, 1823
Animalia	Echinodermata	Holothuroidea	Aspidochirotida	Holothuriidae	Holothuria (Roweothuria) poli Delle Chiaje, 1824
Animalia	Echinodermata	Holothuroidea	Aspidochirotida	Holothuriidae	Holothuria (Holothuria) tubulosa Gmelin, 1791
Animalia	Echinodermata	Ophiuroidea	Ophiurida	Ophiotrichidae	*Ophiothrix fragilis* (Abildgaard, in O.F. Müller, 1789)
Animalia	Echinodermata	Echinoidea	Camarodonta	Parechinidae	Paracentrotus lividus (Lamarck, 1816)
Animalia	Echinodermata	Echinoidea	Camarodonta	Toxopneustidae	Sphaerechinus granularis (Lamarck, 1816)
Kingdom	Phylum	Class	Order	Family	Species
Animalia	Echinodermata	Echinoidea	Cidaroida	Cidaridae	Stylocidaris affinis (Philippi, 1845)
Animalia	Chordata	Ascidiacea	Aplousobranchia	Polyclinidae	*Aplidium* sp. Savigny, 1816
Animalia	Chordata	Ascidiacea	Aplousobranchia	Polyclinidae	*Aplidium undulatum* Monniot & Gail, 1978
Animalia	Chordata	Ascidiacea	Phlebobranchia	Cionidae	*Ciona* sp. Fleming, 1822
Animalia	Chordata	Ascidiacea	Aplousobranchia	Clavelinidae	Clavelina lepadiformis (Müller, 1776)
Animalia	Chordata	Ascidiacea	Aplousobranchia	Clavelinidae	*Clavelina* sp. Savigny, 1816
Animalia	Chordata	Ascidiacea	Aplousobranchia	Polycitoridae	Cystodytes dellechiajei (Della Valle, 1877)
Animalia	Chordata	Ascidiacea	Phlebobranchia	Diazonidae	Diazona violacea Savigny, 1816
Animalia	Chordata	Ascidiacea	Aplousobranchia	Didemnidae	Didemnum drachi Lafargue, 1975
Animalia	Chordata	Ascidiacea	Aplousobranchia	Didemnidae	*Didemnum* sp. Savigny, 1816
Animalia	Chordata	Ascidiacea	Aplousobranchia	Didemnidae	Diplosoma spongiforme (Giard, 1872)
Animalia	Chordata	Ascidiacea	Stolidobranchia	Pyuridae	Halocynthia papillosa (Linnaeus, 1767)
Animalia	Chordata	Ascidiacea	Phlebobranchia	Ascidiidae	Phallusia fumigata (Grube, 1864)
Animalia	Chordata	Ascidiacea	Phlebobranchia	Ascidiidae	Phallusia mammillata (Cuvier, 1815)
Animalia	Chordata	Ascidiacea	Aplousobranchia	Polycitoridae	Polycitor crystallinus (Renier, 1804)
Animalia	Chordata	Ascidiacea	Aplousobranchia	Polyclinidae	Polyclinum aurantium Milne Edwards, 1841
Animalia	Chordata	Ascidiacea	Aplousobranchia	Clavelinidae	*Pycnoclavella* sp. Garstang, 1891
Animalia	Chordata	Ascidiacea	Aplousobranchia	Didemnidae	*Polysyncraton* sp. Nott, 1892

Données prétraitées des fréquences de taxons issues des analyses ARMS

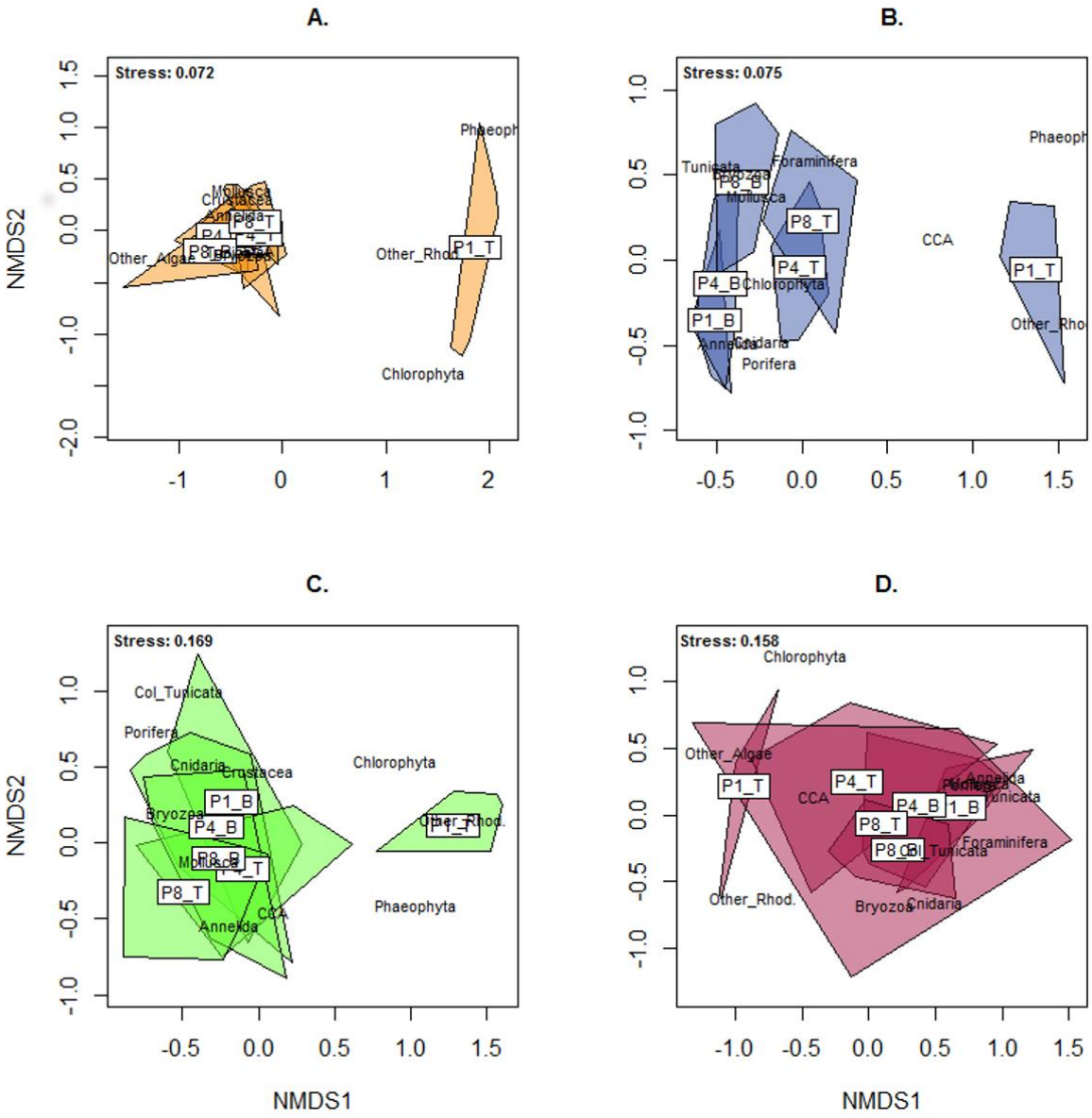
Somme de N pts per group	Annelida	Tunicata	Bryozoa	CCA	Chloroph	Cnidaria	Colonial	Crustace	Echinode	Foramini	Indeterm	Mollusca	Not alive	Other Al	Phaeoph	Porifera	Other Rh
BoB_Lek_R1_P1B	27	0	22	0	4	0	0	0	0	0	0	0	6	0	0	0	5
BoB_Lek_R1_P4B	11	0	16	2	0	0	0	0	0	0	2	1	31	0	0	0	1
BoB_Lek_R1_P8B	18	0	16	0	0	0	0	0	0	0	3	1	25	0	0	0	1
BoB_Lek_R2_P1B	7	0	16	0	11	0	0	0	0	0	0	0	30	0	0	0	0
BoB_Lek_R2_P4B	15	0	21	2	1	0	0	1	0	0	10	0	13	0	1	0	0
BoB_Lek_R2_P8B	14	0	22	0	0	0	0	0	0	0	2	0	22	0	1	0	3
BoB_Lek_R3_P1B	31	0	18	0	3	0	0	0	0	0	0	0	12	0	0	0	0
BoB_Lek_R3_P4B	23	0	10	0	3	0	0	0	0	0	4	1	21	0	0	0	2
BoB_Lek_R3_P8B	13	0	21	0	0	0	0	0	0	0	7	0	22	0	0	0	1
BoB_Zum_R1_P1B	31	0	19	0	0	0	0	0	0	0	8	0	5	0	0	0	1
BoB_Zum_R1_P4B	22	0	16	0	0	0	0	0	0	0	1	0	25	0	0	0	0
BoB_Zum_R1_P8B	17	2	22	0	0	0	0	0	0	0	2	0	20	0	0	0	1
BoB_Zum_R2_P1B	48	0	6	0	0	0	0	1	0	0	1	0	5	0	1	0	2
BoB_Zum_R2_P4B	34	0	6	0	0	0	0	0	0	0	3	0	21	0	0	0	0
BoB_Zum_R2_P8B	30	0	8	0	0	0	0	0	0	0	3	0	22	0	0	0	1
BoB_Zum_R3_P1B	43	0	5	0	0	0	0	2	0	0	2	2	5	0	4	0	1
BoB_Zum_R3_P4B	32	0	8	1	0	0	0	0	0	0	1	1	18	0	1	0	2
BoB_Zum_R3_P8B	32	1	5	0	0	0	0	1	0	0	3	2	18	0	0	0	2
BoB_Pas_R1_P1B	50	0	8	0	0	0	0	0	0	0	0	1	5	0	0	0	0
BoB_Pas_R1_P4B	21	0	3	0	0	0	0	0	0	0	2	2	35	0	0	0	1
BoB_Pas_R1_P8B	18	0	5	0	0	0	0	0	0	0	2	1	36	0	0	0	2
BoB_Pas_R2_P1B	24	0	7	0	0	0	0	0	0	0	3	0	14	16	0	0	0
BoB_Pas_R2_P4B	12	0	3	0	0	0	0	0	0	0	1	1	32	15	0	0	0
BoB_Pas_R2_P8B	6	0	1	0	0	0	0	0	0	0	3	0	32	22	0	0	0
BoB_Pas_R3_P1B	43	0	7	0	0	0	0	0	0	0	2	0	12	0	0	0	0
BoB_Pas_R3_P4B	14	0	4	0	0	0	0	0	0	0	2	2	38	4	0	0	0
BoB_Pas_R3_P8B	15	0	6	0	0	0	0	0	0	0	3	0	35	5	0	0	0
BoB_Lek_R1_P1T	0	0	0	0	31	0	0	0	0	0	0	0	13	0	0	0	20
BoB_Lek_R1_P4T	16	0	11	7	0	0	0	0	0	0	6	0	21	0	0	0	3
BoB_Lek_R1_P8T	16	0	7	11	0	0	0	0	0	0	6	1	20	0	1	0	2
BoB_Lek_R2_P1T	0	0	0	0	40	0	0	0	0	0	0	0	11	0	0	0	13
BoB_Lek_R2_P4T	14	0	24	9	0	0	0	0	0	0	5	0	11	0	0	0	1
BoB_Lek_R2_P8T	17	0	20	0	0	0	0	0	0	0	9	0	7	0	0	0	11
BoB_Lek_R3_P1T	0	0	1	0	26	0	0	0	0	0	8	0	13	0	0	0	16
BoB_Lek_R3_P4T	23	0	13	0	16	0	0	0	0	0	6	0	2	0	0	0	4
BoB_Zum_R1_P1T	0	0	0	0	0	0	0	0	0	0	1	0	0	0	59	0	4
BoB_Zum_R1_P4T	18	0	23	0	0	0	0	0	0	0	15	0	5	0	0	0	3
BoB_Zum_R2_P1T	0	0	0	0	0	0	0	0	0	0	0	0	1	0	54	0	9
BoB_Zum_R2_P4T	36	0	3	0	0	0	0	0	0	0	0	18	4	0	0	0	3
BoB_Zum_R1_P8T	36	0	20	0	0	0	0	0	0	0	2	0	3	0	0	0	3
BoB_Zum_R3_P1T	0	0	0	0	0	0	0	0	0	0	0	0	1	0	52	0	11
BoB_Zum_R3_P4T	41	0	8	0	0	0	0	0	0	0	4	3	2	0	1	0	5
BoB_Zum_R3_P8T	53	0	4	0	0	0	0	1	0	0	1	0	3	0	1	0	1
BoB_Pas_R1_P1T	0	0	0	0	0	0	0	0	0	0	1	0	2	0	37	0	24
BoB_Pas_R1_P4T	20	0	8	1	0	0	0	0	0	0	0	0	28	0	0	0	7
BoB_Pas_R1_P8T	12	0	9	3	0	0	0	0	0	0	2	0	36	0	0	0	2
BoB_Pas_R2_P1T	0	0	0	1	0	0	0	0	0	0	0	0	18	0	19	0	26
BoB_Pas_R2_P4T	25	0	7	0	0	0	0	0	0	0	5	0	26	1	0	0	0
BoB_Pas_R2_P8T	22	0	2	3	0	0	0	0	0	0	1	0	34	2	0	0	0
BoB_Pas_R3_P1T	0	0	1	0	0	0	0	0	0	0	0	0	14	0	12	0	37
BoB_Pas_R3_P4T	37	0	2	1	0	0	0	0	0	0	2	0	19	0	0	0	3
BoB_Pas_R3_P8T	25	0	2	0	0	0	0	0	0	0	2	0	34	1	0	0	0

NWM_ELV_R1_P1B	43	0	11	2	0	0	0	0	0	0	4	0	3	0	0	0	1
NWM_ELV_R1_P4B	31	0	4	1	0	0	0	0	0	0	1	0	24	0	0	2	1
NWM_ELV_R1_P8B	14	0	21	2	0	0	0	0	0	0	0	0	27	0	0	0	0
NWM_ELV_R2_P1B	41	0	14	1	0	0	0	0	0	0	5	0	2	0	0	0	1
NWM_ELV_R2_P4B	23	5	9	0	0	0	0	0	0	0	4	0	23	0	0	0	0
NWM_ELV_R2_P8B	19	2	15	0	0	0	0	0	0	0	3	0	23	0	0	0	2
NWM_ELV_R3_P1B	47	0	9	0	0	0	0	0	0	0	4	0	3	0	0	0	1
NWM_ELV_R3_P4B	33	2	10	0	0	0	0	0	0	0	0	0	17	0	0	1	1
NWM_ELV_R3_P8B	14	0	21	5	0	0	0	0	0	0	1	0	23	0	0	0	0
NWM_RRS_R1_P1B	30	0	14	0	9	1	0	0	0	0	1	0	8	0	0	0	1
NWM_RRS_R1_P4B	19	0	14	1	2	0	0	0	0	0	2	0	24	0	0	1	1
NWM_RRS_R1_P8B	9	0	32	1	0	0	0	0	0	0	2	0	20	0	0	0	0
NWM_RRS_R2_P1B	41	0	16	2	0	2	0	0	0	0	1	0	2	0	0	0	0
NWM_RRS_R2_P4B	27	0	12	3	0	1	0	0	0	0	1	0	20	0	0	0	0
NWM_RRS_R2_P8B	2	0	28	4	0	1	0	0	0	0	1	0	27	0	0	0	1
NWM_RRS_R3_P1B	47	0	5	3	0	3	0	0	0	0	1	0	5	0	0	0	0
NWM_RRS_R3_P4B	30	0	11	4	0	1	0	0	0	0	2	0	16	0	0	0	0
NWM_RRS_R3_P8B	11	3	12	4	0	0	0	0	0	0	6	1	27	0	0	0	0
NWM_CCA_R1_P1B	39	0	17	1	0	0	0	0	0	0	1	1	3	0	0	0	2
NWM_CCA_R1_P4B	6	4	17	2	0	0	0	0	0	0	1	1	33	0	0	0	0
NWM_CCA_R1_P8B	4	0	31	0	0	0	0	0	0	0	2	0	27	0	0	0	0
NWM_CCA_R2_P1B	22	1	31	1	0	0	0	0	0	0	4	0	5	0	0	0	0
NWM_CCA_R2_P4B	11	0	24	0	0	1	0	0	0	0	0	0	28	0	0	0	0
NWM_CCA_R2_P8B	3	11	23	3	0	0	0	0	0	0	1	0	23	0	0	0	0
NWM_CCA_R3_P1B	36	0	15	0	0	6	0	0	0	0	4	1	2	0	0	0	0
NWM_CCA_R3_P4B	13	0	26	0	0	0	0	0	0	0	2	3	20	0	0	0	0
NWM_CCA_R3_P8B	2	0	33	0	0	0	0	0	0	0	0	0	26	0	0	0	3
NWM_ELV_R1_P1T	1	0	0	12	0	0	0	0	0	0	24	0	5	0	0	0	22
NWM_ELV_R1_P4T	24	0	6	8	0	0	0	0	0	0	2	0	19	0	0	1	4
NWM_ELV_R1_P8T	12	0	29	8	0	0	0	0	0	0	4	0	8	0	0	0	3
NWM_ELV_R2_P1T	0	0	0	16	0	0	0	0	0	0	25	0	2	0	0	0	21
NWM_ELV_R2_P4T	20	0	9	6	0	0	0	0	0	0	6	0	14	0	0	9	0
NWM_ELV_R2_P8T	10	0	13	5	0	0	0	0	0	0	9	0	24	0	0	2	1
NWM_ELV_R3_P1T	0	0	0	25	0	0	0	0	0	0	29	0	3	0	0	0	7
NWM_ELV_R3_P4T	31	0	8	10	0	0	0	0	0	0	0	0	12	0	0	1	2
NWM_ELV_R3_P8T	14	0	30	12	0	0	0	0	0	0	2	0	4	0	0	0	2
NWM_RRS_R1_P1T	4	0	0	26	4	0	0	0	0	0	2	0	13	0	7	0	8
NWM_RRS_R1_P4T	11	0	28	17	2	0	0	0	0	0	3	0	3	0	0	0	0
NWM_RRS_R1_P8T	3	0	40	4	0	0	0	0	0	1	1	1	6	0	0	2	6
NWM_RRS_R2_P1T	0	0	0	12	0	1	0	0	0	0	1	0	5	0	0	0	45
NWM_RRS_R2_P4T	21	0	11	8	0	3	0	0	0	0	4	3	5	0	0	6	3
NWM_RRS_R2_P8T	2	0	17	13	0	1	0	0	0	1	24	0	5	0	0	0	1
NWM_RRS_R3_P1T	0	0	0	4	0	0	0	0	0	0	2	0	3	0	0	0	55
NWM_RRS_R3_P4T	15	0	9	15	0	1	0	0	0	0	17	0	7	0	0	0	0
NWM_RRS_R3_P8T	21	0	6	16	0	0	0	0	0	3	14	0	3	0	0	1	0
NWM_CCA_R1_P1T	2	0	0	33	0	0	0	0	0	0	3	0	1	0	0	0	25
NWM_CCA_R1_P4T	19	0	24	9	0	0	0	0	0	0	0	5	5	0	0	0	2
NWM_CCA_R1_P8T	8	0	33	10	0	0	0	0	0	6	2	0	3	0	0	0	2
NWM_CCA_R2_P1T	4	0	3	28	0	0	0	0	0	0	1	0	7	0	0	0	21
NWM_CCA_R2_P4T	22	0	27	11	0	0	0	0	0	0	1	0	3	0	0	0	0
NWM_CCA_R2_P8T	9	0	34	7	0	0	0	0	0	2	4	0	4	0	0	0	4
NWM_CCA_R3_P1T	0	0	0	32	0	1	0	0	0	0	5	0	5	0	0	0	21
NWM_CCA_R3_P4T	8	0	31	13	0	0	0	0	0	1	0	5	6	0	0	0	0
NWM_CCA_R3_P8T	19	0	32	7	0	0	0	0	0	0	0	0	4	0	0	0	2

Somme de N pts per group	Annelida	Tunicata	Bryozoa	CCA	Chloroph	Cnidaria	Colonial	Crustace	Echinodo	Foramini	Indeterm	Mollusca	Not alive	Other Alg	Phaeoph	Porifera	Other Rh
AdS_Azz_R1_P1B	19	0	3	1	0	0	6	0	0	0	0	3	29	0	1	0	2
AdS_Azz_R1_P4B	26	0	0	1	0	1	0	0	0	0	1	0	31	0	0	0	4
AdS_Azz_R1_P8B	16	0	2	0	0	0	0	0	0	0	0	2	44	0	0	0	0
AdS_Azz_R2_P1B	9	0	8	0	1	0	0	0	0	0	0	0	37	0	0	0	9
AdS_Azz_R2_P4B	18	0	2	0	0	0	1	0	0	0	1	1	38	0	0	0	3
AdS_Azz_R2_P8B	9	0	0	0	0	0	0	0	0	0	0	0	52	0	2	0	1
AdS_Sor_R1_P1B	32	0	6	0	0	0	0	0	0	0	0	0	23	0	0	0	3
AdS_Sor_R1_P4B	12	0	16	0	0	0	0	0	0	0	1	11	24	0	0	0	0
AdS_Sor_R1_P8B	7	0	15	1	0	0	0	0	0	0	0	5	33	0	0	3	0
AdS_Sor_R2_P1B	6	0	19	0	0	0	23	0	0	0	0	2	13	0	0	0	1
AdS_Sor_R2_P4B	1	0	16	2	0	0	5	0	0	0	0	4	29	0	0	7	0
AdS_Sor_R2_P8B	5	0	13	0	0	0	4	0	0	0	0	2	26	0	0	14	0
AdS_Sor_R3_P1B	11	0	11	0	0	0	10	1	0	0	0	8	9	0	0	11	3
AdS_Sor_R3_P4B	3	0	11	0	0	1	8	1	0	0	0	4	20	0	0	16	0
AdS_Sor_R3_P8B	8	0	19	0	0	0	6	0	0	0	1	2	28	0	0	0	0
AdS_Sca_R1_P1B	5	0	5	0	0	10	3	0	0	0	1	8	13	0	0	16	3
AdS_Sca_R1_P4B	3	0	2	2	1	6	5	0	0	0	4	5	24	0	0	12	0
AdS_Sca_R1_P8B	8	0	1	1	1	3	8	1	0	0	7	1	27	0	0	4	2
AdS_Sca_R2_P1B	2	0	7	0	1	0	50	0	0	0	0	3	1	0	0	0	0
AdS_Sca_R2_P4B	9	0	3	0	0	4	15	0	0	0	2	0	25	0	0	3	3
AdS_Sca_R2_P8B	6	0	6	0	0	6	2	0	0	0	3	4	34	0	0	0	3
AdS_Sca_R3_P1B	6	0	7	0	0	20	6	0	0	0	6	2	13	0	0	3	1
AdS_Sca_R3_P4B	10	0	3	0	0	7	0	0	0	0	2	3	27	0	0	11	1
AdS_Sca_R3_P8B	9	0	2	0	0	6	0	0	0	0	6	5	36	0	0	0	0
AdS_Azz_R1_P1T	1	0	0	5	0	0	0	0	0	0	9	0	3	0	2	0	44
AdS_Azz_R1_P4T	2	0	0	2	0	0	0	0	0	0	0	8	44	0	0	0	8
AdS_Azz_R1_P8T	3	0	3	2	0	0	0	0	0	0	0	8	47	0	0	0	1
AdS_Azz_R2_P1T	0	0	0	0	2	0	0	0	0	0	0	0	0	0	5	0	57
AdS_Azz_R2_P4T	2	0	0	2	0	0	0	0	0	0	2	4	51	0	0	0	3
AdS_Azz_R2_P8T	1	0	0	5	0	0	0	0	0	0	0	1	57	0	0	0	0
AdS_Sor_R1_P1T	0	0	0	0	1	0	0	1	0	0	0	0	38	0	1	0	23
AdS_Sor_R1_P4T	34	0	11	0	0	0	0	0	0	0	0	6	12	0	0	0	1
AdS_Sor_R1_P8T	17	0	3	0	0	0	0	0	0	0	0	15	24	0	3	0	2
AdS_Sor_R2_P1T	0	0	0	2	0	0	0	0	0	0	0	0	45	0	2	0	15
AdS_Sor_R2_P4T	3	0	13	0	0	0	0	1	0	0	0	7	40	0	0	0	0
AdS_Sor_R2_P8T	7	0	5	1	0	0	2	0	0	0	0	6	42	0	0	0	1
AdS_Sor_R3_P1T	0	0	2	1	1	0	0	0	0	0	0	0	35	0	0	0	25
AdS_Sor_R3_P4T	1	0	6	0	0	0	0	0	0	0	0	3	50	0	0	0	4
AdS_Sor_R3_P8T	1	0	5	0	0	0	0	0	0	0	0	7	48	0	0	3	0
AdS_Sca_R1_P1T	0	0	0	6	0	0	0	0	0	0	0	0	44	0	0	0	14
AdS_Sca_R1_P4T	32	0	0	0	0	0	0	0	0	0	1	10	21	0	0	0	0
AdS_Sca_R1_P8T	38	0	0	1	0	0	0	0	0	0	0	8	17	0	0	0	0
AdS_Sca_R2_P1T	6	0	1	0	1	0	0	0	0	0	0	0	42	0	0	0	14
AdS_Sca_R2_P4T	11	0	4	2	0	0	0	0	0	0	0	0	47	0	0	0	0
AdS_Sca_R2_P8T	3	0	2	0	0	0	0	0	0	0	0	5	52	0	0	0	2
AdS_Sca_R3_P1T	4	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0	28
AdS_Sca_R3_P4T	8	0	8	0	0	1	0	0	0	0	0	8	38	0	0	0	1
AdS_Sca_R3_P8T	10	0	0	0	0	1	0	0	0	0	1	6	45	0	0	0	1

ReS_JSR_R1_P1T	0	0	0	14	0	0	0	0	0	0	0	0	22	19	0	0	9
ReS_JSR_R1_P4T	1	0	19	15	0	0	1	0	0	0	3	1	11	1	0	12	0
ReS_JSR_R1_P8T	0	0	18	4	0	2	0	0	0	0	1	0	36	2	0	1	0
ReS_JSR_R2_P1T	0	0	0	19	0	0	0	0	0	0	0	0	29	3	0	0	13
ReS_JSR_R2_P4T	0	0	7	3	0	0	0	0	0	0	0	0	45	3	0	0	6
ReS_JSR_R2_P8T	0	0	5	0	0	0	0	0	0	0	0	0	57	0	0	0	2
ReS_JSR_R3_P1T	0	0	0	9	17	0	0	0	0	0	0	0	32	0	0	0	6
ReS_JSR_R3_P4T	7	0	1	4	5	0	0	0	0	0	0	3	29	0	0	15	0
ReS_JSR_R3_P8T	0	0	0	4	0	0	0	0	0	0	0	0	2	57	0	0	1
ReS_SOJ_R1_P1T	0	0	0	22	4	0	0	0	0	0	0	0	0	38	0	0	0
ReS_SOJ_R1_P4T	21	0	0	3	0	3	0	0	0	0	1	1	25	0	0	10	0
ReS_SOJ_R1_P8T	23	0	2	1	0	0	2	0	0	1	1	0	17	6	0	10	1
ReS_SOJ_R2_P1T	0	0	0	33	0	0	0	0	0	0	0	0	5	21	0	0	5
ReS_SOJ_R2_P4T	3	0	1	2	0	0	0	0	0	0	0	1	47	2	0	8	0
ReS_SOJ_R2_P8T	6	0	1	0	0	0	0	0	0	2	0	0	55	0	0	0	0
ReS_SOJ_R3_P1T	0	0	0	13	0	0	0	0	0	0	0	0	10	0	0	0	41
ReS_SOJ_R3_P4T	2	0	0	7	0	0	0	0	0	0	0	0	55	0	0	0	0
ReS_SOJ_R3_P8T	2	0	2	11	0	0	0	0	0	0	1	0	13	34	0	0	1
ReS_QAR_R1_P1T	0	0	0	11	0	0	0	0	0	0	0	0	23	0	0	0	30
ReS_QAR_R1_P4T	1	0	8	5	0	0	0	0	0	0	0	0	41	7	0	0	2
ReS_QAR_R1_P8T	0	0	1	8	1	0	0	0	0	0	0	0	44	10	0	0	0
ReS_QAR_R2_P1T	1	0	0	11	0	0	0	0	0	0	0	0	26	26	0	0	0
ReS_QAR_R2_P4T	1	0	0	10	0	0	0	0	0	0	0	0	32	20	0	0	1
ReS_QAR_R2_P8T	1	0	4	10	0	0	0	0	0	0	0	0	45	4	0	0	0
ReS_QAR_R3_P1T	0	0	0	15	0	0	0	0	0	0	0	0	21	25	0	0	3
ReS_QAR_R3_P4T	0	0	15	8	0	1	0	0	0	0	0	0	25	15	0	0	0
ReS_QAR_R3_P8T	0	0	16	11	0	0	1	0	0	0	0	0	27	9	0	0	0
ReS_JSR_R1_P1B	14	0	3	8	0	0	0	0	0	0	0	3	4	4	0	28	0
ReS_JSR_R1_P4B	7	0	4	7	0	0	0	0	0	3	6	0	23	0	0	14	0
ReS_JSR_R1_P8B	1	0	10	19	0	0	0	0	0	1	0	0	25	1	0	7	0
ReS_JSR_R2_P1B	28	0	7	7	0	1	0	0	0	0	3	1	11	0	0	6	0
ReS_JSR_R2_P4B	0	0	16	7	0	0	0	0	0	0	6	0	26	2	0	6	1
ReS_JSR_R2_P8B	5	0	9	0	0	0	0	0	0	0	18	0	27	0	0	3	2
ReS_JSR_R3_P1B	9	0	5	5	0	0	0	0	0	0	2	10	18	0	0	14	1
ReS_JSR_R3_P4B	2	1	16	5	0	0	4	0	0	3	1	1	22	1	0	8	0
ReS_JSR_R3_P8B	0	0	12	11	0	0	0	0	0	3	7	2	26	0	0	3	0
ReS_SOJ_R1_P1B	16	0	0	0	0	0	3	0	0	1	0	0	6	0	0	38	0
ReS_SOJ_R1_P4B	10	0	0	4	0	0	0	0	0	1	3	1	23	0	0	21	1
ReS_SOJ_R1_P8B	7	0	11	4	4	0	0	0	0	1	2	1	19	0	0	15	0
ReS_SOJ_R2_P1B	22	1	1	1	0	0	0	0	0	0	1	2	4	1	0	28	3
ReS_SOJ_R2_P4B	6	4	3	1	0	0	0	0	0	0	1	3	22	0	0	22	2
ReS_SOJ_R2_P8B	3	0	12	0	0	0	0	0	0	0	12	0	33	0	0	4	0
ReS_SOJ_R3_P1B	24	0	8	9	0	0	1	0	0	0	1	8	6	0	0	7	0
ReS_SOJ_R3_P4B	1	0	9	5	0	0	0	0	0	0	2	0	27	0	0	20	0
ReS_SOJ_R3_P8B	8	0	12	10	0	0	0	0	0	0	4	0	19	0	0	11	0
ReS_QAR_R1_P1B	0	0	9	6	0	0	3	0	0	0	9	0	9	0	0	18	10
ReS_QAR_R1_P4B	0	0	0	15	0	1	0	0	0	0	26	0	14	0	0	8	0
ReS_QAR_R1_P8B	3	0	6	16	0	0	0	0	0	0	8	0	25	0	0	5	1
ReS_QAR_R2_P1B	8	1	24	6	0	4	0	0	0	0	9	0	3	0	0	4	5
ReS_QAR_R2_P4B	2	0	8	6	0	15	0	0	0	0	4	0	23	0	0	6	0
ReS_QAR_R2_P8B	4	0	26	13	1	0	0	0	0	0	2	0	14	0	0	2	2
ReS_QAR_R3_P1B	0	0	29	5	0	0	5	0	0	1	4	2	2	5	0	11	0
ReS_QAR_R3_P4B	0	0	18	16	0	0	0	0	0	0	8	0	14	0	0	8	0
ReS_QAR_R3_P8B	1	0	19	16	0	0	0	0	0	0	4	0	16	6	0	2	0

Matériel Supplémentaire



S1 Fig. Non-metric Multidimensional scaling (nMDS) plots of the plate faces for each sea.

Non-metric multidimensional scaling (nMDS) representation of community for each plate face, within each Sea. A: Bay of Biscay, B: Northwest Mediterranean Sea; C: Adriatic Sea, D: Red Sea. Each dot represents one ARMS for a given plate face, the stress value is indicated. Variables of species composition (taxa) are also represented to visualize possible important associations of some taxa with some plates (e.g. various algal taxa are close to the P1T in the 2D-representation).

S1 Tab. Results of three-way ANOSIM with various environmental factors : Symbols for significance levels are : : #:P<0.1; *:P<0.05; **:P<0.01; ***:P<0.001; NS: not significant

Environmental factor tested	R parameter (ANOSIM)	Significance level ANOSIM (R & significance level)
Protection_status	0.188	*
General_anthropization	0.077	#
Marine_debris	0.133	#
Sewage_output	0.104	*
Chemical_pollution	-0.09	#
Urbanization	0.086	*
Harbor	0.043	NS
Nearby_Seagrass_meadows	0.188	*
Nearby_sand	0.26	***
Nearby_mud	0.13	*

Annexe 5 : IndexMed

Contexte

IndexMed est un consortium pluridisciplinaire créé dans le cadre de cette thèse à l'IMBE (Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale). Son objectif principal est de développer la culture des bases de données et leur utilisation efficace dans le milieu de la recherche en écologie et biodiversité. Ce consortium s'est étendu à plusieurs UMR de disciplines différentes (notamment, de l'environnement pour l'expertise qualitative de la donnée, et de l'astronomie pour l'expertise en matière de gestion des grosses masses de données). Il a permis de répondre à des appels à projet dans le domaine des bases de données en écologie méditerranéenne en favorisant l'interdisciplinarité et les collaborations avec d'autres entités du CNRS.

Des projets chaque année

Les projets qui y sont développés s'appuient sur les différentes démarches nationales et internationales et promeuvent un travail partenarial international. IndexMed sert notamment de relais aux réseaux et démarches en place nationalement et internationalement, et proposer une réponse aux obligations européennes (Aarhus, INSPIRE…) auxquelles les laboratoires de recherche travaillant dans les domaines de l'environnement et des sciences humaines sont de plus en plus soumis.

Objectif général

L'objectif à court terme d'IndexMed est de mettre en place une plateforme d'indexation des données sur la biodiversité méditerranéenne et des paramètres environnementaux ayant un intérêt pour la recherche (David et al 2015). Cette indexation utilise les outils et méthodes préconisés nationalement (SINP - Système national d'Information sur la Nature et les Paysages, MNHN - Muséum National d'Histoire Naturelle, SPN - Service du Patrimoine Naturel, RBDD Réseau Bases De Données du CNRS) ou internationalement (OBIS, GBIF, LifeWatch, GEOBON, CoL, WoRMS…) et s'appuie sur les catalogues développés à ce niveau (IDCNP - Inventaire des Dispositifs de Collecte sur la Nature et les Paysages du SINP, Réseaux d'acteurs de la FRB - Fondation pour la Recherche sur la Biodiversité). En 2017, IndexMed devient IndexMEED

Annexe 6 : indexMEED

IndexMed est devenu IndexMEED

IndexMed a prospéré et s'est agrandi tant en nombre de participants qu'en nombre de disciplines impliquées. Les implications de ses membres dans des groupes de travail internationaux (T.D.W.G, R.D.A., LifeWatch notamment) ont mis en évidence le besoin d'internationaliser le consortium, dans le but, entre autres, et aussi de participer à des réponses à appel à projets à un niveau international. Depuis le séminaire de juin 2016, IndexMed est devenu IndexMEED (Indexing for Mining Ecological and Environmental Data), un nouveau consortium ayant pour objectif d'indexer les données sur la biodiversité et l'environnement en général, issues de bases de données hétérogènes et distantes, en utilisant les graphes et en les analysant. Le consortium regroupe aujourd'hui des chercheurs d'autres champs disciplinaires que la biodiversité (Archéologie, Astronomie, Anthropologie…)

GRAMINÉES [132], une action labellisée et soutenue par le GDR MaDICS en 2017 et 2018

La mise en place d'une dynamique d'échange entre des experts en écologie / biodiversité / archéologie et des experts du domaine des S.T.I.C. est devenu une priorité dans le cadre du consortium IndexMEED. Proposée à différents financeurs sous la forme d'une action, celle-ci regroupe des représentants des deux champs disciplinaires et permettra de formaliser des besoins en terme d'analyses de données hétérogènes de la part de la communauté écologie / biodiversité / archéologie / anthropologie et de stimuler la recherche en S.T.I.C. afin de proposer des solutions plus adéquates pour l'analyse et la gestion des données écologiques dans le contexte du Big Data (prise en compte de la dimension temporelle et spatiale et des données multi-échelles et hétérogènes).

En ce qui concerne les techniques et approches S.T.I.C. étudiées et développées, la recherche sera en 2018 dirigée vers des techniques de gestion et d'analyse de graphes qui puissent prendre en compte la complexité des données hétérogènes et notamment passer à l'échelle sur des jeux des données volumineux sans détériorer la qualité des résultats obtenus. Sous l'égide d'IndexMEED et dans le cadre de GRAMINÉES, il est prévu de réaliser une première carte des compétences de laboratoires en informatique qui pourront apporter

[132] https://www.madics.fr/actions/actions-en-cours/graminees/

des outils méthodologiques ou des techniques algorithmiques adéquates pour l'analyse des données issue de l'écologie et des Sciences humaines et sociales.

Ce qu'il est envisageable pour la suite

Les cas d'étude en cours (8 en cours de réalisation dont un pour le jeu de données issu du programme CIGESMED et un pour le jeu de données issu du programme DEVOTES), une fois les graphes élaborés et les résultats présentés, seront autant d'exemples de l'utilisation de la théorie des graphes en écologie, et feront l'objet d'une diffusion dans les communautés de recherche concernées.

Lors de la prochaine année, la recherche et les approches S.T.I.C. étudiées et développées seront dirigées vers des méthodes de gestion et d'analyse de graphes qui puissent prendre en compte la complexité des données hétérogènes et notamment passer à l'échelle sur des jeux des données volumineux sans détériorer la qualité des résultats obtenus.

Il est prévu l'organisation d'ateliers sur des compétences complémentaires s'appuyant sur les cas d'études développés dans le cadre des ateliers de 2017. Les thèmes proposés par les participants concernent « la sémantique et les ontologies interdisciplinaires dans le domaine environnemental » et « Algorithmes et calculs sur les graphes basés sur des données en sciences de l'environnement ».

Ce travail doit servir de base au développement d'un réseau au niveau européen, qui permettra de continuer à rapprocher les communautés des écologues et des S.T.I.C. autour des challenges proposés par la fouille de données en écologie et environnement basée sur la théorie des graphes.

La finalisation de la carte de compétences ébauchée en 2017 permettra de détecter et d'impliquer un ensemble de laboratoires (et des personnels associés) pour envisager des actions au plan international comme, par exemple, une proposition de type M.R.S.E.I. (ANR) Cette MRESI permet de préparer une participation à des appels d'offres européens (BiodivERsA, SeasEra, H2020…) sur la thématique du Big Data, liés aux données de l'écologie et de la biodiversité.

Une étape importante sera la mise en place d'une démonstration de fouilles de graphes basés sur des données environnementales utilisant la grille de calcul (soutien France Grilles).

A plus long terme

Ces graphes sont paramétrables pour fouiller et visualiser ces données pluridisciplinaires en mettant sur le même plan des données de types écologiques, physico-chimiques, fonctionnelles (relations trophiques, traits fonctionnels), et socio-écologiques, économiques... Les questionnements scientifiques possibles concernent l'écologie des systèmes observés : bon état écologique, correspondance de patrons de contextes et de

données concernant les abondances relatives d'espèces ou des systèmes d'observation (détection des biais dans la formation des observateurs, expertise partielle dans les jeux de données, définition de la puissance de l'échantillonnage nécessaire, gestion des coûts associés).

Le projet GRAINE, une application concrète des travaux du consortium IndexMEED appliquée aux problématiques « Homme-milieux »

L'objectif général de ce projet, retenu par le LabEx Dispositif de Recherche Interdisciplinaire sur les Interactions Hommes-Milieux (DRIIHM), est de construire une méthode de visualisation de données hétérogènes basée sur les graphes dans le cadre de trois « cas d'études » issus d'Observatoires Hommes-Milieux (OHM), en utilisant le prototype développé par le consortium IndexMEED. Pour chaque OHM, cette méthode de visualisation sera appliquée i) au système d'observation, c'est-à-dire à l'organisation des métadonnées et données et à la manière dont celles-ci diffèrent (type, qualité, contenu…), et ii) au système observé ou à une partie de celui-ci (ce résultat sera fonction de la consistance des données dans chacun des OHM). Une pré-étude de faisabilité sera réalisée sur trois OHM (Bassin minier de Provence, Vallée du Rhône et Littoral méditerranéen). L'ambition est de mettre en évidence les particularités et points communs, points forts et axes d'amélioration de chaque système observé en considérant tout type de données (biodiversité, socio-écologie, économie, économétrie de l'environnement…).

Appliquer cette approche aux données produites dans le cadre des OHM permettra d'élaborer une méthode d'intégration puis de fouille de données dans le cas de données très hétérogènes liant environnement et société, ce qui constitue un des objectifs majeurs du DRIIHM. Ce travail permettra d'évaluer le potentiel des représentations par les graphes de données hétérogènes issues ou intéressant les OHM, en réalisant des ateliers de curation et d'amélioration de la qualité des données et des systèmes d'informations, de manière à produire ces premières représentations.

L'objectif fixé pour la première année est d'obtenir une carte des données généralisable à tous les OHM, permettant de mettre en exergue les descripteurs de données soit équivalents, soit traitant des mêmes objets mais avec un vocabulaire différent (ce qui arrive fréquemment dans les projets interdisciplinaires). Il permettra aussi de mettre en évidence les descripteurs ou valeurs de descripteurs polysémiques, qui induisent une ambiguïté dans l'analyse intégrée de plusieurs lots de données. La méthode mise en place et testée sur les trois OHM ciblés sera documentée pour être transposable aux autres OHM.

Objectifs opérationnels de GRAINE Des OHMs

Les objectifs opérationnels de ce projet « GRAINE Des OHMs » ébauché lors de la fin de ma thèse sont de construire une méthode de visualisation de données hétérogènes basée sur les graphes dans le cadre de trois OHM « cas d'études », en utilisant le prototype développé par le consortium IndexMed. Pour chaque OHM, cette méthode de visualisation sera appliquée

- Au système d'observation, c'est-à-dire à l'organisation des métadonnées et données et à la manière dont celles-ci diffèrent (type, qualité, contenu…), et
- Au système observé ou à une partie de celui-ci (ce résultat sera fonction de la consistance des données dans chacun des OHM).

Une visualisation intégrant les trois OHM permettra de mettre en évidence les particularités et points communs, points forts et axes d'amélioration de chaque système observé.

Appliquer cette approche aux données produites dans le cadre des OHM permettra d'élaborer une méthode d'intégration puis de fouille de données dans le cas de données très hétérogènes liant environnement et société, ce qui constitue un des objectifs majeurs du Dispositif de Recherche Interdisciplinaire sur les Interactions Hommes-Milieux.

Le travail ébauché consiste donc à évaluer le potentiel des représentations par les graphes de données hétérogènes issus ou intéressant les OHMs, en réalisant des ateliers de curation et d'amélioration de la qualité des données et des systèmes d'informations, de manière à produire ces premières représentations.

Le travail (en cours) concerne la recherche de descripteurs communs/équivalents ou de patrons de valeurs de descripteurs contenus dans les données libres d'accès et métadonnées de l'IMBE, des OHM « volontaires » et des observatoires qui en dépendent. Pour aboutir à ces représentations, des réunions similaires à celles qui ont été organisés pour les cas d'études en écologie se tiendront (curation de données, puis visualisation) sur les données des OHM impliqués et d'y solliciter les équipes techniques et les experts S.T.I.C. du consortium IndexMed.

Résultats attendus

Le résultat attendu est une carte des données généralisable à tous les OHM, permettant de mettre en exergue les descripteurs de données soit équivalents, soit traitant des mêmes objets mais avec un vocabulaire différent (ce qui arrive fréquemment dans les projets interdisciplinaires). Il permettra aussi de mettre en évidence les descripteurs ou valeurs de descripteurs polysémiques, qui induisent une ambiguïté dans l'analyse intégrée de plusieurs lots de données.

Le fait de représenter tous les objets des bases sur le même graphe permettra d'étudier les systèmes d'information en eux-mêmes. L'opérateur peut visualiser l'importance de chaque

base dans la sélection géographique correspondant à l'OHM ou à une zone plus restreinte si une typologie géographique lui est attribuée. Il est possible aussi de sélectionner une période ou d'autres critères descripteurs des bases ou des sites eux même. Cette représentation devra aussi permettre de mettre rapidement en évidence des particularités ou des erreurs dans les jeux de données.

Pour chaque OHM, il est prévu dans le cadre de ce projet

- Un premier audit/bilan des systèmes d'observation et des systèmes d'informations existants ou en développement
- Une première représentation des jeux de données des systèmes d'information sous forme d'un graphe
- Une première représentation du système observé sous forme de graphe, qui permettra d'appréhender l'importance relative de l'information accessible pour chaque champ disciplinaire dans le périmètre de l'OHM.

Pour les trois OHM volontaires, une représentation commune des données dans les systèmes d'information à partir des champs de métadonnées permettra une visualisation transversale ; Celle-ci permettra d'effectuer une première liste des améliorations possibles ordonnées par priorité/sensibilité.

Annexe 7 : Activité scientifique en support ou en complément du travail de thèse

Articles dans une revue

Romain David, Maria C. Uyarra, Susana Carvalho, Holger Anlauf, Angel Borja, Abigail E. Cahill, Roberto Danovaro, Laura Carugati, Aurélien De Jode, Jean-Pierre Féral, Dorian Guillemain, Marco Lo Martire, Laure Thierry de Ville D 'Avray, John K. Pearman, Anne Chenuil, 2018 *submitted* "Photo analyses of Autonomous Reef Monitoring Structures, as monitoring tools to detect spatial, environmental and anthropic pressure effects", ***Marine Pollution Bulletin.***

Romain David, Anna Cohen Nabeiro, Jean-Pierre Féral, Aurélie Delavaud, Anne-Sophie Archambeau, Fanny Arnaud, David Auber, Nicolas Bailly, Loup Bernard, Cyrille Blanpain, Romain Bourqui, Vincent Breton, Denis Couvet, Alrick Dias, Sophie Gachet, Robin Goffaux, Karina Gibert, Manuel Herrera, Dino Ienco, Romain Julliard, Julien Lecubin, Yannick Legre, Michelle Leydet, Grégoire Loïs, Victor Méndez Muñoz, Jean-Charles Meunier, Isabelle Mougenot, Sophie Pamerlon, Jean-Claude Raynal, Geneviève Romier, Dad Roux-Michollet, Alison Specht, Christian Surace, Thierry Tatoni 2018 *accepted*, Bilan des journées du GRAAL 2016 et avenir de l'utilisation des graphes en écologie, ***Nature, Science et Société***.

Aurélien De Jode, **Romain David**, Anne Haguenauer, Abigail E. Cahill, Zinovia Erga, Dorian Guillemain, Stéphane Sartoretto, Caroline Rocher, Marjorie Selva, Line Legall, Jean-Pierre Féral, Anne Chenuil, (2018, submitted), Multiple cryptic species, spatial and ecological differentiation in a major builder of coralligenous habitats, ***Molecular Ecology***.

Laure Thierry de Ville D 'Avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral. Application of the Ecosystem Service Concept to a Local-Scale: The Cases of Coralligenous Habitats in the North-Western Mediterranean Sea. 2017 *accepted*. ***Marine Pollution Bulletin.*** <halshs-01624589>

Abigail E. Cahill, John Pearman, Angel Borja, Laura Carugati, Susana Carvalho, Roberto Danovaro, Sarah Dashfield, **Romain David**, Jean-Pierre Féral, Sergej Olenin, Andrius Siaulys, Paul Somerfield, Antoaneta Trayanova, Maria Uyarra, Anne Chenuil, "A

comparative analysis of metabarcoding and morphology-based identification of benthic communities across different regional seas" 2018 *submitted* ***Ecology and Evolution***.

Abigail E. Cahill, Aurelien De Jode, Sophie Dubois, Zoheir Bouzaza, Didier Aurelle, Emilie Boissin, Olivier Chabrol, **Romain David**, Emilie Egea, Jean-Baptiste Ledoux, Bastien Merigot, Alexandra Anh-Thu Weber, Anne Chenuil, (online pas encore pagine 2017) A multispecies approach reveals hot-spots and cold-spots of diversity and connectivity in species with contrasting dispersal modes. ***Molecular Ecology***, Wiley, 2017, 26 (23), pp.6563-6577. 〈10.1111/mec.14389〉. 〈hal-01681650〉

Romain David, Loup Bernard, Cyrille Blanpain, Alrick Dias, Jean-Pierre Féral, Sophie Gachet, Julien Lecubin, Christian Surace, Thierry Tatoni. Visualisation de données sous forme de graphes en archéologie. Rencontre opérationnelle des archéologues d'ArkeoGIS et des écologuesd'IndexMed . ***Digital Archaeology***, iste open science, 2017, 17-1 (1), 〈 https://www.openscience.fr/Data-visualisation-in-archaeology-based-on-graph-approach-Operational-meeting〉. 〈hal-01617580〉

Stéphane Sartoretto, Thomas Schohn, Carlo Bianchi, Carla Morri, Joaquim Garrabou, Enric Ballesteros, Sandrine Ruitton, Marc Verlaque, Boris Daniel, Eric Charbonnel, Sylvain Blouet, **Romain David**, Jean-Pierre Féral, Giulia Gatti, An integrated method to evaluate and monitor the conservation state of coralligenous habitats: The INDEX-COR approach. ***Marine Pollution Bulletin***, Elsevier, 2017, 〈10.1016/j.marpolbul.2017.05.020〉. 〈hal-01541141〉

Vasilis Gerovasileiou, Thanos Dailianis, Emmanouela Panteri, Nikitas Michalakis, Giulia Gatti, Maria Sini, Charalampos Dimitriadis, Yiannis Issaris, Maria Salomidi, Irene Filiopoulou, Alper Doğan, Laure Thierry de Ville D 'Avray, **Romain David**, Ertan Çinar, Drosos Koutsoubas, Jean-Pierre Féral, Christos Arvanitidis. CIGESMED for divers: Establishing a citizen science initiative for the mapping and monitoring of coralligenous assemblages in the Mediterranean Sea. ***Biodiversity Data Journal***, Pensoft, 2016, 54 (e8692), 〈 http://bdj.pensoft.net/〉. 〈10.3897/BDJ.4.e8692〉. 〈hal-01392025〉

Roberto Danovaro, Laura Carugati, Berzano Marco, Abigail E. Cahill, Susana De Carvalho Spinola, Anne Chenuil, Cinzia Corinaldesi, Cristina Sonia, **Romain David**, Antonio Dell'Anno, Nina Dzhembekova, Esther Garces, Joseph Gasol, Goela Priscila, Jean-Pierre Féral, Isabel Ferrera, Rodney Forster, Andrey Kurekin, Eugenio Rastelli, Veselka Marinova, Peter I. Miller, Snejana Moncheva, Alice Newton, John Pearman, Sophie Pitois, Albert Reñé, Naiara

Rodriguez-Ezpeleta, Vincenzo Saggiomo, Stefan Simis, Kremena Stefanova, Christian Wilson, Marco Lo Martire, Silvestro Greco, Sabine Cochrane, Olga Mangoni, Angel Borja. Implementing and Innovating Marine Monitoring Approaches for Assessing Marine Environmental Status.. ***Frontiers in Marine Science***, Frontiers Media, 2016, 3, pp.213. 〈10.3389/fmars.2016.00213〉. 〈hal-01448726〉

Proceedings de communications à un congrès

Víctor Méndez Muñoz, Anna Cohen-Nabeiro, **Romain David**, Vicente Ivars Camáñez, Alfons Nonell-Canals, Miquel Senar, Denis Couvet, Jean-Pierre Féral, Aurélie Delavaud, Thierry Tatoni. Analysis on the Graph Techniques for Data-mining and Visualization of Heterogeneous Biodiversity Data Sets. Complexis 2017, Apr 2017, Porto, Portugal. pp.144 - 151, 2017, ***Proceedings of the 2nd International Conference on Complexity, Future Information Systems and Risk (COMPLEXIS 2017).*** 〈http://www.complexis.org/?y=2017. 〈10.5220/0006379701440151〉. 〈hal-01541140〉

Bénédicte Madon, **Romain David**, René Garello, Linwood Pendleton, Ronan Fablet. Strike-Alert: Towards Real-time, High Resolution Navigational Software for Whale Avoidance. SUSTECH 2017 : ***5th annual IEEE Conference on Technologies for Sustainability***, Nov 2017, Phoenix, United States. 2017, 〈http://sites.ieee.org/sustech/〉. 〈hal-01623903〉

Romain David, Jean-Pierre Féral, Thierry Tatoni. Architecture, concepts et services d'un système d'indexation de données distribuées pour l'observation à large échelle en écologie marine. ***32ème Conférence sur la Gestion de Données - Principes, Technologies et Applications***, Nov 2016, POITIERS, France. BDA 2016 2016, 〈https://bda2016.ensma.fr/index.html. 〈hal-01426497〉

Romain David, Jean-Pierre Féral, Anne Archambeau, Nicolas Bailly, Cyrille Blanpain, Vincent Breton, Aurélie De Jode, Aurélie Delavaud, Alrick Dias, Sophie Gachet, Dorian Guillemain, Julien Lecubin, Geneviève Romier, Christian Surace, Thierry Thierry de Ville D 'Avray, Christos Arvanitidis, Anne Chenuil, Melih Ertan inar, Drosos Koutsoubas, Stéphane Sartoretto, Thierry Tatoni. IndexMed projects: new tools using the CIGESMED DataBase on Coralligenous for indexing, visualizing and data mining based on graphs. Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.). 8th International Congress on Environmental Modelling and Software, Jul 2016, Toulouse, France. Brigham Young University BYU Scholars Archive, International Environmental Modelling and Software Society (iEMSs) 8th International Congress on Environmental Modelling and Software

Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.) http://www.iemss.org/society/index.php/iemss-2016-proceedings, 8th, pp.32, 2016, ***International Congress on Environmental Modelling and Soft ware***. 〈 http://scholarsarchive.byu.edu/iemssconference/2016/?utm_source=scholarsarchive.byu.edu%2Fiemssconference%2F2016%2FStream-C%2F32&utm_medium=PDF&utm_campaign=PDFCoverPages〉. 〈hal-01425559〉

Romain David, Jean-Pierre Féral, Sophie Gachet, Alrick Dias, Cyrille Blanpain, Julien Lecubin, Cristinel Diaconu, Christian Surace, Gibert Karina. A first prototype for indexing, visualizing and mining heterogeneous data in Mediterranean ecology: within the IndexMed consortium interdisciplinary framework. 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Nov 2015, Bangkok, Thailand. IEEE Explore, pp. 232-239, 2015, ***11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)***. 〈http://ieeexplore.ieee.org/document/7400571/〉. 〈 10.1109/SITIS.2015.119〉. 〈hal-01433600〉

Katerina Sevastou, Nadia Papadopoulou, Chris Smith, Heliana Teixeira H., Chiara Pirodd, Stelios Katsanevakis, Jean-Pierre Féral, Anne Chenuil, **Romain David**, Niki Kiriakopoulou, Sabine Cochrane, (2015) - Defining keystone species in European regional seas: what are the candidates for the Mediterranean? In: ***11th Panhellenic Symposium on Oceanography & Fisheries «Aquatic Horizons: Challenges & Perspectives***. Mytilene, Lesvos Island, Greece, 13-17 May 2015. Athens: H.C.M.R., 553-556.

Jean-Pierre Féral, Christos Arvanitidis, Anne Chenuil, Melih Ertan Çinar, **Romain David**, Emilie Egea, Stéphane Sartoretto, (2014) - “CIGESMED: Coralligenous based indicators to evaluate and monitor the "good environmental statut" of the Mediterranean coastal waters, a SeasEra project. 2nd Mediterranean Symposium on the conservation of Coralligenous & other Calcareous Bio-Concretions, Oct 2014, Portorož, Slovenia. ***Proceedings of the 2nd Mediterranean Symposium on the conservation of Coralligenous & other Calcareous Bio-Concretions.*** 〈hal-01620607〉

Stéphane Sartoretto, **Romain David**, Didier Aurelle, Anne Chenuil-Maurel, Dorian Guillemain, Laure Thierry de Ville D 'Avray, Jean-Pierre Féral, Melih Ertan Çinar, Silvija Kipson, Christos Arvanitidis, Thomas Schohn, Boris Daniel, Selmane Sakher, Joaquim Garrabou, Giulia Gatti, Enric Ballesteros, (2014) - An integrated approach to evaluate and monitor the conservation state of coralligenous bottoms: the INDEX-COR method.

Proceeding of the 2nd Mediterranean Symposium on the conservation of Coralligenous & other Calcareous Bio-Concretions, Oct 2014, Portorož, Slovenia. 2014, 〈10.13140/2.1.3180.6405〉. 〈hal-01620618〉

Romain David, Christos Arvanitidis, Melih Ertan Çinar, Stéphane Sartoretto, A Doan, Sophie Dubois, Zinovia Erga, Dorian Guillemain, Laure Thierry de Ville d''Avray, Frederic Zuberer, Anne Chenuil-Maurel, Jean-Pierre Féral, (2014) - CIGESMED habitat's characterization: a simple and reusable typology at the Mediterranean scale. 2nd Mediterranean Symposium on the conservation of Coralligenous & other Calcareous Bio-Concretions, Oct 2014, Portorož, Slovenia. ***Proceedings of the 2nd Mediterranean Symposium on the conservation of Coralligenous & other Calcareous Bio-Concretions***. 〈hal-01620541〉

Romain David, Christos Arvanitidis, Melih Ertan Çinar, Stéphane Sartoretto, Sophie Dubois, Zinovia Erga, Dorian Guillemain, Laure Thierry de Ville D 'Avray, Frederic Zuberer, Anne Chenuil-Maurel, Jean-Pierre Féral, (2014) - CIGESMED Protocols : how to implement a multidisciplinary approach on a large scale for coralligenous habitats surveys. RAC/SPA 2nd Mediterranean Symp. on the Conservation of coralligenous and other calcareous bio-concretions, Portorož, Slovenia, 29-30/10/2014, pp. 66-71, 〈10.13140/2.1.1895.0086〉. 〈hal-01620550〉

Romain David, Sophie Dubois, Zinovia Erga, Dorian Guillemain, Laure Thierry de Ville D 'Avray, Christos Arvanitidis, Melih Ertan Çinar, Stéphane Sartoretto,Frederic Zuberer, Anne Chenuil-Maurel, Jean-Pierre Féral, (2014) - CIGESMED.'s protocol and network (Coralligenous basEd. Indicators to evaluate and monitor the "Good Environmental Status" of Mediterranean coastal waters). ***Proceedings of the 5th International Symposium of Monitoring of Mediterranean coastal areas: problems and measurement techniques***, Livorno (Italy) 17-18-19 June 2014, F. Benincasa (Ed.), pp. 828-843 ; CNR-IBIMET : Florence (IT), ISBN 978-88-95597-19-5

Chapitres d'ouvrage

Romain David, Jean-Pierre Féral, Cyrille Blanpain, (2015), - Ecological Data Preservation in the context of IndexMed, In : Cristinel Diaconu [Ed] PREDON 2015 , pp.

Romain David, Sophie Dubois, Zinovia Erga, Dorian Guillemain, Laure Thierry de Ville d'Avray, Christos Arvanitidis, Melih Çinar, Stéphane Sartoretto, Frédéric Zuberer, Anne Chenuil, Jean-Pierre Féral with other contributors (2015) - "CIGESMED*'s protocol and

network (Coralligenous based Indicators to Evaluate and Monitor the « good ecological status » of the MEDiterranean coastal waters)", CIGESMED Project. Pages 828-843 ; CNR-IBIMET Florence (Italy), December 2014, ISBN : 978-88-95597-19-5 (Fifth Symposium Monitoring of Mediterranean coastal areas : problems and measurement techniques Livorno (Italy) 17-19 june 2014)

Jean-Pierre Féral, **Romain David**. Zones côtières et développement durable : une équation à résoudre. CNRS Editions, Le développement durable à découvert, pp.96-97, 2013, 978-2-271-07896-4. 〈http://www.cnrseditions.fr/sociologie/6777-le-developpement-durable-a-decouvert-sous-la-direction-d-agathe-euzen-laurence-eymard-francoise-gaill.html〉. 〈hal-01620564〉

Communications

Présentations orales (séminaires internationaux)

- 23 mars 2018 : **Romain David**, Anna Cohen Nabeiro, Aurélie Delavaud, Alison Specht, *remotely* "Developing crediting/rewarding mechanisms to foster resources (data and materials) sharing in Research: towards recommendations - Case study : the biodiversity community*" **IG Sharing Rewards and Credit (SHARC) - RDA 11th Plenary meeting***, Berlin, Allemagne https://rd-alliance.org/ig-sharing-rewards-and-credit-sharc-rda-11th-plenary-meeting
- 4 Octobre 2017 : (accepted but cancelled) **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, David Auber, Nicolas Bailly, Loup Bernard, Laure Berti-Equille, Cyrille Blanpain, Vincent Breton, Anne Chenuil-Maurel, Anna Cohen-Nabeiro, Alrick Dias, Aurélie Delavaud, Robin Goffaux, Sophie Gachet, Karina Gibert, Manuel Herrera Fernandez, Luc Hogie, Dino Ienco, Romain Julliard, Yvan Le Bras, Julien Lecubin, Yannick Legre, Michelle Leydet, Grégoire Lois, Bénédicte Madon, François Marchal, Víctor Méndez Muñoz, Jean-Charles Meunier, Jean-Baptiste Mihoub, Isabelle Mougenot, Sophie Pamerlon, Eric Peletier, Geneviève Romier, Dad Roux-Michollet, Alison Specht, Christian Surace, Jean-Claude Raynal, Thierry Tatoni, "IndexMEED cases studies using Omics data with graph theory". ***Biodiversity Information Science and Standards,*** 1, e20740. 〈10.3897/tdwgproceedings.1.20740〉. 〈hal-01761535〉
- 13 Juillet 2017 : **Romain David** and the IndexMEED community "IndexMEED consortium for data mining in ecology: How to build graphs and mine heterogeneous data for environmental research?" ***LifeWatch and EUDAT workshop : Ontology &***

Semantic Web for Research, 11 - 14 Juillet 2017, Lecce, Italy http://www.servicecentrelifewatch.eu/ontology-semantic-web-for-biodiversity-ecosystem-research-programme

- 9 – 11 mai 2017 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, David Auber, Nicolas Bailly, Loup Bernard, Laure Berti-Equille, Cyrille Blanpain, Vincent Breton, Anne Chenuil-Maurel, Anna Cohen-Nabeiro, Alrick Dias, Aurélie Delavaud, Robin Goffaux, Sophie Gachet, Karina Gibert, Manuel Herrera Fernandez, Luc Hogie, Dino Ienco, Romain Julliard, Yvan Le Bras, Julien Lecubin, Yannick Legre, Michelle Leydet, Grégoire Lois, Bénédicte Madon, François Marchal, Víctor Méndez Muñoz, Jean-Charles Meunier, Jean-Baptiste Mihoub, Isabelle Mougenot, Sophie Pamerlon, Eric Peletier, Geneviève Romier, Dad Roux-Michollet, Alison Specht, Christian Surace, Jean-Claude Raynal, Thierry Tatoni, “Prérequis, principes et bonnes pratiques concernant l'organisation des Systèmes d'Information pour une utilisation efficace des données hétérogènes et nécessairement multi-sources et distribuées en écologie, environnement et société”, ***driihm2017 : Séminaire annuel du LabEx DRIIHM 2017***, Aveiro, Portugal.
- 5 – 9 Décembre 2016 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, David Auber, Nicolas Bailly, Cyrille Blanpain, Vincent Breton, Alrick Dias, Anna Cohen-Nabeiro, Aurélie Delavaud, Sophie Gachet, Robin Goffaux, Karina Gibert, Manuel Herrera, Dino Ienco, Romain Julliard, Julien Lecubin, Yannick Legre, Grégoire Loïs, Victor Méndez Muñoz, Jean-Charles Meunier, Isabelle Mougenot, Sophie Pamerlon, Geneviève Romier, Alison Specht, Christian Surace, Thierry Tatoni, “Results of IndexMed GRAIL Days 2016: How to use standards to build GRAphs and mIne data for environmental research” ***TDWG 2016 annual conference***, Santa Clara de San Carlos, Costa Rica http://www.tdwg.org/fileadmin/2016conference/documents/TDWG_Conference_program-en_US.pdf
- 25 octobre 2016 : **Romain. David**, Invited speaker, “Learn about Open Access, open data, data papers, data sensibility and property and the debate around it and why”, ***Congrès de la Société Française d’Ecologie***, Marseille, France, http://sfecologie2016.sciencesconf.org/
- 24 – 28 octobre 2016 : **Romain David**, Jean-Pierre Féral, anne-sophie archambeau, Nicolas Bailly, cyrille Blanpain, Vincent Breton, Denis Couvet, A. Aurélie Delavaud, Alrick Dias, Sophie Gachet, Isabelle Mougenot, Julien Lecubin, Michelle Leydet, Jean-Claude Raynal, Samuel Robert, Geneviève Romier, Alison Specht, Christian Surace, Thierry Tatoni, “Graph approach of heterogeneous data, the new possibilities developed by the IndexMed consortium for data mining in Mediterranean ecology”,

Congrès de la Société Française d'Ecologie, Marseille, France, http://sfecologie2016.sciencesconf.org/

- 17 – 19 Octobre 2016 : **Romain David**, Maria C. Uyarra, Susana Carvalho, Holger Anlauf, Angel Borja, Abigail E. Cahill, Roberto Danovaro, Laura Carugati, Aurélien De Jode, Jean-Pierre Féral, Dorian Guillemain, Marco Lo Martire, John K. Pearman, Anne Chenuil, "Application of Autonomous Reef Monitoring Structures (ARMS) photo analysis to European Regional Seas and the Red Sea" ***Final Meeting DEVOTES***, Brussels, Belgium http://www.devotes-project.eu/devotes-final-conference-presentations/
- 26 – 30 September 2016 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, Nicolas Bailly, Cyrille Blanpain, Alrick Dias, Julien Lecubin, Geneviève Romier, Christian Surace, and the IndexMed community, "IndexMed: Original solutions to manage the heterogeneity of marine ecology data in the Mediterranean Sea" ***EMBS, 51st European Marine Biology Symposium***, Rhodes, Greece http://www.embs51.org/programme/
- 10 – 14 juillet 2016 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, Nicolas Bailly, Cyrille Blanpain, Vincent Breton, Aurélien De Jode, Aurélie Delavaud, Alrick Dias, Sophie Gachet, Dorian Guillemain, Julien Lecubin, Geneviève Romier, Christian Surace, Laure Thierry de Ville d'Avray, Christos Arvanitidis, Anne Chenuil, Melih Ertan Çinar, Drosos Koutsoubas, Stéphane Sartoretto, Thierry Tatoni, « IndexMed projects: new tools using the CIGESMED DataBase on Coralligenous for indexing, visualizing and data mining based on graphs », ***iEMSs 2016 : 8th International Congress on Environmental Modelling and Software***, C1 : VI Data Mining for Environmental Sciences Session in Toulouse, France, on July 10-14, 2016. http://www.iemss.org/sites/iemss2016/
- 5 avril 2016 : **Romain David**, Jean-Pierre Féral, Thierry Tatoni «IndexMed, a consortium responsible for the task of indexing Mediterranean biodiversity data: a new way for data mining in ecology», - ***E-infrastructure EGI Workshop***, Amsterdam. https://indico.egi.eu/indico/event/2895/
- 1 – 3 Décembre 2015 : **Romain David**, Abigail E. Cahill, Dorian Guillemain, Jean-Pierre Féral and Anne Chenuil, « Analysing relatives abundance of taxa on Artificial Reef Monitoring System with photoquad», ***Devotes annual meeting***, Lisbonne (Portugal). DOI: 10.13140/RG.2.1.4222.0566
- 23 – 27 november 2015 : **Romain David**, Jean-Pierre Féral, Cyrille Blanpain, Cristinel Diaconu, Alrick Dias, Sophie Gachet, Karina Gibert, Julien Lecubin, Christian Surace, "A first prototype for indexing, visualizing and mining heterogeneous data in Mediterranean ecology within the interdisciplinary framework of IndexMed

consortium", at ***11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS 2015)***. Bangkok: IEEE Conference proceeding in IEEE Xplore Digital Library – Volume 119: pp. 232-239 Thailand; 23 – 27/11/2015 DOI: 10.1109/SITIS.2015.119

- 28 septembre – 2 octobre 2015, **Romain David**, Jean-Pierre Féral, IndexMed, a consortium charged with the task of indexing Mediterranean biodiversity data, a new way for data mining in ecology, ***TDWG 2015 annual conférence***, Nairobi, Kenya. DOI: 10.13140/RG.2.1.4385.6083 https://mbgserv18.mobot.org/ocs/index.php/tdwg/2015/paper/view/793
- 24 – 25 juin 2015 : **Romain David**, « Problématique de préservation des données hétérogènes dans le domaine de la biodiversité : Enjeux et écueils concernant l'indexation, l'identification pérenne des données et leur cycle de vie », ***Assemblée générale du GDR MADICS (Masses de Données, Informations et Connaissances en Sciences)***, atelier préservation et reproductibilité des données à Lyon
- 19 – 22 Mai 2015 : **Romain David**, Christos Arvanitidis, Melih Ertan Çinar, Stéphane Sartoretto, A Doan, Sophie Dubois, Zinovia Erga, Dorian Guillemain, Laure Thierry de Ville d'Avray, Frederic Zuberer, Anne Chenuil-Maurel, Jean-Pierre Féral, CIGESMED habitat's characterization: a simple and reusable typology at the Mediterranean scale, ***CIGESMED second general assembly*** (GA2), Mytilène, Grèce.
- 19 – 22 Mai 2015 : **Romain David**, Sophie Dubois, Zinovia Erga, Dorian Guillemain, Laure Thierry de Ville D 'Avray, Christos Arvanitidis, Melih Ertan Çinar, Stéphane Sartoretto, Frederic Zuberer, Anne Chenuil-Maurel, Jean-Pierre Féral, CIGESMED's protocol and network (Coralligenous basEd. Indicators to evaluate and monitor the "Good Environmental Status" of MEDiterranean coastal waters) ***CIGESMED second general assembly*** (GA2), Mytilène, Grèce.
- 22 May 2015 : **Romain David,** « Graph and data visualisations, interpretative and analysing tools », ***General Assembly of CIGESMED***, Mytilène, Grèce
- 20 May 2015 : **Romain David,** « Where are the gaps to finalize the protocole ?», ***General Assembly of CIGESMED***, Mytilène, Grèce
- 19 May 2015 : **Romain David,** « WP6 new results and perspective », ***General Assembly of CIGESMED***, CEA Day, Mytilène, Grèce
- 29 october 2014 : **Romain David,** CIGESMED protocols: how to implement a multidisciplinary approach on a large scale for coralligenous habitats surveys, ***The 2nd Mediterranean Symposium on Coralligenous and other calcareous bio-concretions,*** 29-30 octobre 2014, Portoroz, Slovenie.

- 17 june 2014 : **Romain David,** CIGESMED field methods to improve datasets on coralligenous habitats for the effective evaluation of the "Good Environmental Status" of the Mediterranean Sea ? ***Fifth Symposium Monitoring of Mediterranean coastal areas: problems and measurement techniques,*** Livorno, Italy.
- 25 October 2013 : **Romain David,** Ensuring coherent networks of MPAs : WS5A3 instrument and actions in the European water : « Currently discussed MPA improvement and new solutions for the Mediterranean Sea : information management with the example of coralligenous habitat (***IMPAC3 Congrès Mondial des Aires Marines Protégées***, auditorium du Pharo), Marseille, France.
- 18 avril 2013 : **Romain David,** « Quality of informations systems : how to build it ? », ***Kick-off meeting CIGESMED***, Heraklion, Crète.

Co-authoring de présentations orales internationales

- 24 – 28 octobre 2016 : Jean-Pierre Féral, Thomas Saucède, Elie Poulin, **Romain David**, Christian Marschal, Gilles Marty, Jean-Claude Roca, Sébastien Motreuil, Jean-Pierre Beurier, "PROTEKER: Setting up of an underwater observatory at the Kerguelen Islands (Austral Ocean)" , ***Congrès de la Société Française d'Ecologie***, Marseille, France, http://sfecologie2016.sciencesconf.org/
- 24 – 28 octobre 2016 : Giulia Gatti, Luigi Piazzi, Thomas Schon, **Romain David**, Monica Montefalcone, Jean-Pierre Féral, Stéphane Sartoretto « Comparison of Coralligenous Indices », ***Congrès de la Société Française d'Ecologie***, Marseille, France, http://sfecologie2016.sciencesconf.org/
- 24 – 28 octobre 2016 : Giulia Gatti, Charalampos Dimitriadis, Vasilis Gerovasileiou, Thanos Dailianis, Emmanouella Panteri, Yiannis Issaris, Maria Sini, Maria Salomidi, Nikitas Michalakis, Alper Doğan, Laure Thierry de Ville d'Avray, **Romain David**, Melih Ertan Çinar, Drosos Koutsoubas Christos Arvanitidis, Jean-Pierre Féral, " Citizen Science for CIGESMED: involving divers in marine biological monitoring ", ***Congrès de la Société Française d'Ecologie***, Marseille, France, http://sfecologie2016.sciencesconf.org/
- 24 – 28 octobre 2016 : Laure Thierry de Ville d'Avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral, « Ecosystem services provided by coralligenous habitats. A step toward economic valuation. International conference on ecological sciences », ***Congrès de la Société Française d'Ecologie***, Marseille, France, http://sfecologie2016.sciencesconf.org/
- 24 – 28 octobre 2016 : Aurélien De Jode, **Romain David**, Dorian Guillemain, Jacky Dubar, Jean-Pierre Féral and Anne Chenuil, "Community ecology of the coralligenous

assemblages using a metabarcoding approach." ***Congrès de la Société Française d'Ecologie***, Marseille, France, http://sfecologie2016.sciencesconf.org/

- 26 – 30 September 2016 : Giulia Gatti, Luigi Piazzi, Thomas Schon, **Romain David**, Monica Montefalcone, Jean-Pierre Féral, Stéphane Sartoretto « Comparison of Coralligenous Indices », ***51st European Marine Biology Symposium***, Rhodes, Greece http://www.embs51.org/
- 26 – 30 September 2016 : Giulia Gatti, Charalampos Dimitriadis, Vasilis Gerovasileiou, Thanos Dailianis, Emmanouella Panteri, Yiannis Issaris, Maria Sini, Maria Salomidi, Nikitas Michalakis, Alper Doğan, Laure Thierry de Ville d'Avray, **Romain David**, Melih Ertan Çinar, Drosos Koutsoubas Christos Arvanitidis, Jean-Pierre Féral, " Citizen Science for CIGESMED: involving divers in marine biological monitoring " , ***51st European Marine Biology Symposium***, Rhodes, Greece http://www.embs51.org/
- 26 – 30 September 2016 : Jean-Pierre Féral, Melih Ertan Çinar, Stephane Sartoretto, Anne Chenuil, Christos Arvanitidis, **Romain David**, Drosos Koutsoubas and the CIGESMED community, CIGESMED project: an attempt for challenging management of coralligenous habitats in the Mediterranean Sea, ***51st European Marine Biology Symposium***, Rhodes, Greece http://www.embs51.org/
- 12 – 16 September 2016 : Abigail E. Cahill, **Romain David** , Jean-Pierre Féral, Anne Chenuil, "Community composition of macro invertebrates in the Mediterranean Sea", ***41th CIESM Congress***, Kiel, Germany, http://ciesm.org/marine/congresses/Kiel.htm
- 12 – 16 September 2016 : Aurélien De Jode, **Romain David**, Dorian Guillemain , Jacky Dubar, Jean-Pierre Féral and Anne Chenuil, "Community ecology of the coralligenous assemblages using a metabarcoding approach." ***41th CIESM Congress***, Kiel, Germany, http://ciesm.org/marine/congresses/Kiel.htm
- 19 – 21 May 2016 : Giulia Gatti, Charalampos Dimitriadis, Vasilis Gerovasileiou, Thanos Dailianis, Emmanouela Panteri, Yiannis Issaris, Maria Sini, Maria Salomidi, Nikitas Michalakis, Alper Doğan, Laure Thierry de Ville d'Avray, **Romain David**, Melih Ertan Çinar, Drosos Koutsoubas Christos Arvanitidis, Jean-Pierre Féral "Citizen Science for CIGESMED, or how to engage divers in marine ecological monitoring: first steps of a new project, ***1st international ECSA Conference "Citizen Science - Innovation in Open Science, Society and Policy"*** in Berlin. http://ecsa.citizen-science.net
- 10 Mai 2016 : J.-P. Féral, **Romain David**, C. Marschal, D. Guillemain, F. Zuberer, A. Cahill, A. De Jode, A. Chenuil , « Dive monitoring of sensitive coastal habitats : the use of ARMSs boxes and ASUs in the north west Mediterranean », ***2nd European Conference on Scientific Diving***, 9-11 May at the Sven Lovén Centre for Marine

Sciences - Kristineberg, University of Gothenburg, Sweden http://loven.gu.se/english/Study+and+work/research/ecsd2016

- 3 – 7 janvier 2016 : Abigail E. Cahill, Didier Aurelle, Emilie Boissin, Zoheir Bouzaza, **Romain David**, Sophie Dubois, Aurélien DeJode, Emilie Egea, Zinovia Erga, Jean-Baptiste Ledoux, Bastien Mérigot, Alexandra Weber, and Anne Chenuil, « Determinants of connectivity in the marine environment: a multispecies approach », congress of ***Society of Integrative and Comparative Biology***, à Portland (Etats-Unis). http://www.sicb.org/meetings/2016/index.php
- 7 – 11 octobre 2015 : Thanos Dailianis, Vasilis Gerovasileiou, Yiannis Issaris, Maria Salomidi, Maria Sini, Vasilis Gerakaris, Eleftheria-Niki Mantzani, Melina Nalmpanti, Christos Arvanitidis, **Romain David**, Jean-Pierre Féral, Coralligenous habitats assessment in SW Greece: the case of Korinthiakos Gulf, Ionian Sea, ***ICZEGAR congress*** (http://13iczegar.nhmc.uoc.gr/) (Crête) DOI: 10.13140/RG.2.1.4622.2966
- 7 – 11 octobre 2015 : Vasilis Gerovasileiou, Thanos Dailianis, Emmanouela Panteri, Giulia Gatti, Yiannis Issaris, Maria Sini, Maria Salomidi, Charalampos Dimitriadis, Nikitas Michalakis, Alper Doğan, Laure Thierry de Ville d'Avray, **Romain David**, Melih Ertan Çinar, Drosos Koutsoubas, Christos Arvanitidis, Jean-Pierre Féral, "Establishing a citizen science initiative for the mapping and monitoring of coralligenous assemblages in the Mediterranean Sea", ***ICZEGAR congress*** (http://13iczegar.nhmc.uoc.gr/) (Crête)
- 19 – 22 Mai 2015 : Jean-Pierre Féral, Christos Arvanitidis, Anne Chenuil-Maurel, Melih Ertan Çinar, **Romain David**, Emilie Egea, Stéphane Sartoretto, CIGESMED: Coralligenous based indicators to evaluate and monitor the "good environmental statut" of the Mediterranean coastal waters, a SeasEra project. ***CIGESMED second general assembly*** (GA2), Mytilène, Grèce.
- 19 – 22 Mai 2015 : Laure Thierry de Ville d'Avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral, Ecosystem services provided by coralligenous habitats - Measurement & valuation. ***CIGESMED second general assembly*** (GA2), Mytilène, Grèce.
- 27 – 31 october 2014 : Emmanouela Panteri, Sarah Faulwetter, Christos Arvanitidis, Thanos Dailianis, George Chatzigeorgiou, Eva Chatzinikolaou, Evangelos Pafilis, Christina Pavloudi, Thomas Uher, Simon D Rycroft, Alexander Kroupa, Vincent Smith, Lyubomir Penev, Edward Baker, Jean-Pierre Féral, **Romain David**, "crowd sourcing initiatives in the mediterranean bassin" Presentation at the ***Biodiversity Information Standards (TDWD) Conference 2014***, Jönköping Sweden https://mbgserv18.mobot.org/ocs/index.php/tdwg/2014/paper/view/684

Présentations orales (séminaires d'audience locale et nationale)

- 4 – 7 Juillet 2017 : **Romain David**, Luc Hogie, Anna Cohen Nabeiro T2.GT02 : Outils, compétences et savoir faire nécessaire à la fouille de graphes : quelques cas d'utilisation et les défis qui les accompagnent, ***JDEV 2017***, http://devlog.cnrs.fr/jdev2017/t2.gt02
- 26 avril 2017 : **Romain David,** "Consortium IndexMEED : les graphes, depuis l'indexation jusqu'à l'exploration des données et l'émergence de nouvelles hypothèses jusqu'à l'aide à la décision", ***Forum des utilisateurs d'Ecoscope*** - Paris, France. http://www.fondationbiodiversite.fr/images/documents/ECOSCOPE/2017_Programme_Forum_Utilisateurs_et_Journee_Portail_MD.pdf
- 28 – 29 mars 2017, **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, Fanny Arnaud, David Auber, Nicolas Bailly, Loup Bernard, Cyrille Blanpain, Vincent Breton, Anna Cohen-Nabeiro, Denis Couvet, Aurélie Delavaud, Cristinel Diaconu, Alrick Dias, Sophie Gachet, Robin Goffaux, Manuel Herrera Fernandez, Karina Gibert, Dino Ienco, Romain Julliard , Julien Lecubin, Yannick Legre, Michelle Leydet, Grégoire Loïs , Victor Mendez Munoz, Jean-charles Meunier, Isabelle Mougenot, Sophie Pamerlon, Jean-Claude Raynal, Samuel Robert, Geneviève Romier, Dad Roux-Michollet, Alison Specht, Christian Surace, Thierry Tatoni, Régine Vignes Lebbe, et la communauté IndexM(e)ed, "Indexing and Mining Ecological and Environmental Data : Architecture, concepts et services d'un système d'indexation de données distribuées pour l'observation et l'inter-opération de données à large échelle en écologie / environnement"" ***Séminaire INSIDE "le numérique au service de l'eau, de la biodiversité et du milieu marin"***, Paris, France. http://www.pole-inside.fr/fr/page/presentations-seminaire
- 26 Février 2017 : **Romain. David**, IndexMEED Consortium : Indexing and Mining Ecological and Environmental Data, ***Comité d'Appuis Scientifique et Technique (CAST) de la F.R.B.***, Paris, France. http://ecoscope.fondationbiodiversite.fr/en/home-en-gb/10-ecoscope
- 15 – 18 Novembre, 2016 : **Romain David**, Jean-Pierre Féral, Thierry Tatoni. Architecture, concepts et services d'un système d'indexation de données distribuées pour l'observation à large échelle en écologie marine. ***32ème Conférence sur la Gestion de Données - Principes, Technologies et Applications***, Poitiers, France. https://bda2016.ensma.fr/index.html.
- 11 – 13 octobre 2016 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, Nicolas Bailly, Cyrille Blanpain, Vincent Breton, Aurélien De Jode, Aurélie Delavaud,

Alrick Dias, Sophie Gachet, Dorian Guillemain, Julien Lecubin, Geneviève Romier, Christian Surace, Laure Thierry de Ville d'Avray, Christos Arvanitidis, Anne Chenuil, Melih Ertan Çinar, Drosos Koutsoubas, Stéphane Sartoretto, Thierry Tatoni : "IndexMed projects: new tools using the CIGESMED DataBase on Coralligenous for indexing, visualizing and data mining based on graphs", ***ACO 2016 A connected ocean: new approaches, new technologies, new challenges for knowledge of ocean processes*** organized by IEEE Oceanic Engineering Society, http://aconnectedocean.sciencesconf.org/

- 30 juin 2016 : Laure Thierry de Ville d'Avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral, Quelle est la valeur des services écosystémiques fournis par les habitats coralligènes ? ***Journée des doctorants de l'IMBE***, Marseille, France.
- 29 juin 2016 : Laure Thierry de Ville d'Avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral, What can be the contribution of coralligenous in the Ecosystem Service research field ? ***Cigesmed final meeting***, Marseille, France.
- 8 – 9 Juin 2016 : **Romain David**, Loup Bernard, avec Thierry Tatoni, Jean-Pierre Féral, Cyrille Blanpain, Alrick Dias, Julien Lecubin, Michelle Leydet, Christian Surace, « Aller au-delà des données agrégées dans ArkeoGIS : premiers essais de graphes au sein d'IndexMed » ***Journées Informatique et Archéologie de Paris (JIAP)*** à l'Institut d'Art et d'Archéologie, Paris. http://jiap2016.sciencesconf.org/
- 6 juin 2016 : **Romain David**, Jean-Pierre Féral, Thierry Tatoni, "Project IndexMed: original answers to the heterogeneity of ecological data for the Mediterranean Sea based on graph approaches" ***GRAIL Days 2016 GRAphs and datamIning for environmentaL research - Data, Research questions and New hypotheses - IndexMed consortium*** - Faculty of Saint Charles - Marseille university - 6 - 8 of june 2016 - https://indexmed2016.sciencesconf.org/resource/page/id/10
- 23 Mars 2016 : **Romain David**, Jean-Pierre Féral, Thierry Tatoni : Limites et freins à l'interopérabilité : quelques cas d'école et quelques clefs de réussite, ***assemblée générale de l'EMBRC (Centre National de Ressources Biologiques Marines)***, Villefranche, http://www.embrc-france.fr/fr
- 24 – 25 Février 2016 : **Romain David**, Jean-Pierre Féral, Thierry Tatoni : « Projet VIGI-GEEK : VIsualisation of Graph Intransdisciplinary Global Ecology, Economy and Sociology data-Kernel », ***Journées du défi Imagin du CNRS***, Paris, France http://www.cnrs.fr/mi/spip.php?article889
- 10 février 2016 : **Romain David**, « Contexte, mise en œuvre de moyens et développement d'outils pour appréhender l'interdisciplinarité dans le cadre d'IndexMed », ***2ème journée de restitution du défi ENVIROMICS du CNRS « promouvoir l'émergence d'approches pluridisciplinaires pour l'étude des***

objets et questions relevant des omiques environnementales » porté par la mission pour l'interdisciplinarité, Paris, France. http://www.cnrs.fr/mi/spip.php?article881

- 4 février 2016 : Laure Thierry de Ville d'Avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral, Les services écosystémiques rendus par le coralligène de Méditerranée. ***Atelier de réflexion du réseau Efèse-Mer***, Marseille, France.
- 21 Janvier 2016 : **Romain David**, « Utilisation des graphes en écologie, un nouveau moyen de développer l'interdisciplinarité dans le cadre d'IndexMed », ***Séminaire VegFrance***, Marseille, France
- 14 – 15 janvier 2016 : Gatti Giulia, Gerovasileiou Vasilis, Dailianis Thanos, Panteri Emmanouella, Issaris Yiannis, Sini Maria, Salomidi Maria, Dimitriadis Charalampos, Nikitas Michalakis, Doğan Alper, Thierry de Ville d'Avray Laure, **David Romain**, Çinar Melih Ertan, Koutsoubas Drosos, Arvanitidis Christos, Féral Jean-Pierre, Citizen science for CIGESMED : pour une cartographie et un suivi des habitats coralligènes à l'échelle Méditerranéenne, ***SÉMINAIRE Liteau "Observation et recherche en appui aux politiques du littoral et de la mer"***, à Brest http://www1.liteau.net/index.php/agenda/colloque-liteau-janvier-2016-a-brest
- 7 janvier 2016 : Laure Thierry de Ville d'Avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral, Les habitats coralligènes, trésors cachés de la Méditerranée…Qu'ont-ils à nous offrir ? ***Causerie de la commision biologie de la FFESSM***, Marseille, France.
- 10 – 11 décembre 2015 : **Romain David**, Jean-Pierre Féral, Thierry Tatoni : « Projet VIGI-GEEK : VIsualisation of Graph Intransdisciplinary Global Ecology, Economy and Sociology data-Kernel » ***Journées MASTODON2***, Paris DOI: 10.13140/RG.2.1.3792.9043
- 15 Octobre 2015 : **Romain David**, Cyrille Blanpain, Alrick Dias, Jean-Pierre Féral, Julien Lecubin, Christian Surace - "Prototype d'indexation, de visualisation et de fouille de données hétérogènes en écologie dans le cadre du consortium IndexMed : appel à participations", ***Journée thématique du PR2I Big Data Aix Marseille Université sur le Traitement de données massives***, Marseille
- 14 Octobre 2015 : **Romain David**, Cyrille Blanpain, Alrick Dias, Jean-Pierre Féral, Julien Lecubin, Christian Surace - « utilisation des graphes pour faire émerger de nouvelles questions de recherche sur la biodiversité » / démonstration d'un prototype », ***Séminaire IndexMed "Méthodes et outils pour la fouille de données hétérogènes et multi-sources en écologie"***, Marseille, France http://www.indexmed.eu/Programme-au-08-10-2015-Program-0n.html

- 29 septembre – 2 octobre : Christian Surace, **Romain David**, Jean-Pierre Féral, Cyrille Blanpain, “Project IndexMed, Interoperability in social-ecological datasets”, Biodivmex ***communication***”, Observatoire de Haute Provence 04870 Saint Michel l’Observatoire, France
- 24 – 25 septembre 2015, **Romain David**, « Limites et freins à l’interopérabilité : quelques cas d'école et quelques clefs de réussite » ***sist15 : Séries Interopérables et Systèmes de Traitement***, Marseille (France) 50 personnes
- 16 – 18 septembre 2015 : Abigail E. Cahill, Didier Aurelle, Emilie Boissin, Zoheir Bouzaza, **Romain David**, Sophie Dubois, Aurélien DeJode, Emilie Egea, Zinovia Erga, Jean-Baptiste Ledoux, Bastien Mérigot, Alexandra Weber, and Anne Chenuil, « Determinants of connectivity in the marine environment: a multispecies approach », ***Conference Marine Connectivity GDR of CNRS*** meeting Montpellier, France 80 personnes.
- 1er Juillet 2015 : Laure Thierry de Ville d'Avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral, “Les habitats coralligènes, trésors cachés de la Méditerranée…Qu’ont-ils à nous offrir ?” ***Journée des doctorants de l'IMBE***, Marseille, France.
- 21 – 22 avril 2015 : **Romain David**, « De l’inter-calibration de dispositifs de collecte et de collecteurs à la mise en place d’un réseau de suivi des habitats coralligènes en Méditerranée : le défi d’un protocole à large échelle (Programme CIGESMED)» ***Congrès annuel des doctorants organisé par l’Ecole Doctorale des Sciences de l’Environnement (ED251)***,
- 20 avril 2015 : **Romain David**, « Le big data en écologie ? Pas encore disent certains… Pas si sûr ! Avec IndexMed, Relevons ce challenge ! » ***Labex OTMED***, Aix-en-Provence.
- 30 – 31 mars 2015 : Laure Thierry de Ville d'Avray, Dominique Ami, Frédéric Aprahamian, Anne Chenuil, **Romain David**, Jean-Pierre Féral, Evaluation des services écosystémiques fournis par les habitats Coralligènes. Journée de l'OHM, Marseille, France.
- 25 Novembre 2014 : **Romain David**, « le big Big Data en écologie n’est pas pour maintenant ? Pas si sûr ! » ***Séminaire AMU « Big Data »*** organisé par la PR2I de l’AMU (Aix Marseille Université), Marseille, France.
- 24 Novembre 2014 : Zinovia Erga, **Romain David**, Aurélien De Jode, Marc Verlaque, Sophie Dubois, Line Le Gall, Dorian Guillemain, Frédéric Zuberer, Jean-Pierre Féral, Anne Chenuil “Strong genetic structure in Lithophyllum cabiochae/stictaeforme around Marseilles. Local adaptation, niche differentiation or just lack of connectivity

?" ***Colloque annuel de la Société Phycologique de France***, Aquarium de la Porte dorée, Paris

- 4 – 6 novembre : **Romain David**, IndexMed, vers l'interopérabilité des bases de données en écologie – ***Séminaire Predon***, Paris https://indico.cern.ch/event/338461/
- 5 juin 2014 : **Romain David**, Séminaire IndexMed : interopérabilité des bases de données en écologie, ***Etat des lieux sur le statut des données environnementales en France (séminaire Allenvi / relevé de conclusions)***, Marseille Saint Charles, 65 participants (http://www.indexmed.eu/-Premier-seminaire-Interoperabilite-.html)
- 25 mars 2014 : **Romain David**, « ideas about information management for multi-disciplinary network observatory » ***Rencontre annuelle de l'OHM Littoral Méditerranéen***, Marseille
- 2 septembre 2013 : **Romain David**, « intégration des données biophysique, écologiques et socio-économiques (information management) », ***Action Nationale de Formation CNRS : "structuration scientifique des zones ateliers autour du schéma conceptuel des systèmes socio-écologiques des LTER : méthodes et actions",*** Banyuls, France, http://www.za-inee.org/fr/node/511
- 2 – 4 juin 2013 : **Romain David**, « Qualification of ecological data », ***dive-training CIGESMED***, Marseille, France.
- 8 février 2013 : **Romain David**, "Capitalisation et gestion des données en écologie : Quelles responsabilités ? Quels objectifs ? Quelle stratégie ? Quelle organisation ?" ***Conseil scientifique de l'INEE***, Paris, France.

Conférence grand public

- 18 mars 2014 : **Romain David**, « Ecologie et peuplements des substrats marins : contexte de la plongée scientifique et approches de terrain. » ***14èmes Rencontres scientifiques enseignants-chercheurs sur les « Enjeux, méthodes et moyens de la recherche scientifique en milieu marin côtier »***, 22 enseignants participants.
- Laure Thierry de Ville d'Avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral, "Evaluation des services écosystémiques fournis par les habitats Coralligènes". ***Rencontres enseignants chercheurs***, Marseille, France, 14 avril 2014.

Posters

- 28 – 30 Juin 2017 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, David Auber, Nicolas Bailly, Cyrille Blanpain, Vincent Breton, Alrick Dias, Anna Cohen-Nabeiro, Aurélie Delavaud, Sophie Gachet, Robin Goffaux, Karina Gibert, Manuel Herrera, Dino Ienco, Romain Julliard, Julien Lecubin, Yannick Legre,

Grégoire Loïs, Victor Méndez Muñoz, Jean-Charles Meunier, Isabelle Mougenot, Sophie Pamerlon, Geneviève Romier, Alison Specht, Christian Surace, Thierry Tatoni, "Results of IndexMed GRAIL Days 2016: How to use standards to build GRAphs and mIne data for environmentaL research?" ***Séminaire GBIF EUBON IDIV "Symposium on the Mobilization of Structured Biodiversity Data"***, Leipzig, Allemagne. https://geobon.org/downloads/PDF/DraftAgenda_Light_DataMobilizationSymposium.pdf

- 11 – 14 Juillet 2017 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, David Auber, Nicolas Bailly, Cyrille Blanpain, Vincent Breton, Alrick Dias, Anna Cohen-Nabeiro, Aurélie Delavaud, Sophie Gachet, Robin Goffaux, Karina Gibert, Manuel Herrera, Dino Ienco, Romain Julliard, Julien Lecubin, Yannick Legre, Grégoire Loïs, Victor Méndez Muñoz, Jean-Charles Meunier, Isabelle Mougenot, Sophie Pamerlon, Geneviève Romier, Alison Specht, Christian Surace, Thierry Tatoni, "Results of IndexMed GRAIL Days 2016: How to use standards to build GRAphs and mIne data for environmentaL research?" ***LifeWatch and EUDAT workshop : Ontology & Semantic Web for Research,*** Lecce, Italy. http://www.servicecentrelifewatch.eu/ontology-semantic-web-for-biodiversity-ecosystem-research-programmee
- 5 – 7 April 2017 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, David Auber, Nicolas Bailly, Cyrille Blanpain, Vincent Breton, Alrick Dias, Anna Cohen-Nabeiro, Aurélie Delavaud, Sophie Gachet, Robin Goffaux, Karina Gibert, Manuel Herrera, Dino Ienco, Romain Julliard, Julien Lecubin, Yannick Legre, Grégoire Loïs, Victor Méndez Muñoz, Jean-Charles Meunier, Isabelle Mougenot, Sophie Pamerlon, Geneviève Romier, Alison Specht, Christian Surace, Thierry Tatoni, "Results of IndexMed GRAIL Days 2016: How to use standards to build GRAphs and mIne data for environmentaL research?" ***RDA 9th Plenary meeting poster session***, 5-7 April 2017, Barcelona, Spain. https://www.rd-alliance.org/plenaries/rda-ninth-plenary-meeting-barcelona/rda-9th-plenary-poster-session
- 9 – 10 févr. 2017 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, David Auber, Nicolas Bailly, Cyrille Blanpain, Vincent Breton, Alrick Dias, Anna Cohen-Nabeiro, Aurélie Delavaud, Sophie Gachet, Robin Goffaux, Karina Gibert, Manuel Herrera, Dino Ienco, Romain Julliard, Julien Lecubin, Yannick Legre, Grégoire Loïs, Victor Méndez Muñoz, Jean-Charles Meunier, Isabelle Mougenot, Sophie Pamerlon, Geneviève Romier, Alison Specht, Christian Surace, Thierry Tatoni, "Results of IndexMed GRAIL Days 2016: How to use standards to build

GRAphs and mIne data for environmentaL research?" ***Journées MASTODONS (grandes MASses DE DONnées Scientifiques)*** Paris, France. http://www.cnrs.fr/mi/IMG/pdf/colloque_de_restitution_du_defi_mastodons.pdf

- 23 - 24 Novembre 2016 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, Nicolas Bailly, cyrille Blanpain, Vincent Breton, Anna Cohen-nabeiro, Aurélie Delavaud, Alrick Dias, sophie gachet, Robin Goffaux, Dino Ienco, Romain Julliard, Julien Lecubin, Sophie Pamerlon, Genevieve Romier, Christian Surace, Thierry Tatoni "Architecture et services pour la fouille de données hétérogènes en écologie, dans le cadre du consortium IndexMEED", ***Journées SUCCES - Rencontres Scientifiques des Utilisateurs de Calcul intensif, de Cloud et de Stockage***, Paris, France. https://succes2016.sciencesconf.org/resource/page/id/7
- 23 juin 2016 : Julie Rostan, **Romain David**, Dorian Guillemain, Jacky Dubar, Anne Chenuil et Aurélien De Jode, Étude de la composition des communautés du coralligène en vue d'une analyse par metabarcoding, Poster de stage de Master 1 Sciences de l'Univers Écologie Environnement - Spécialité Océanographie et Environnements Marins, Université Pierre et Marie Curie
- 10 Juin 2016 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, Nicolas Bailly, Cyrille Blanpain, Vincent Breton, Aurélie Delavaud, Alrick Dias, Sophie Gachet, Julien Lecubin, Michèle Leydet, Geneviève Romier, Christian Surace, Thierry Tatoni and the IndexMed community*, « IndexMed heterogeneous data workflows for improving the quality of GBIF France services » at ***10 ans du GBIF***, Paris, France.
- 1 – 3 Juin 2016 : Michelle Barbier-Leydet, Armand Rotereau, Emmanuel Gandouin, **Romain David**, "Improvement of the EPD Tool", ***Meeting 2016 de l'European Pollen Database (EPD)***, Aix-en-Provence, France.
- 1 – 3 Juin 2016 : R. Romain David, Jean-Pierre Féral, Anne-Sophie Archambeau, Loup Bernard, Cyrille Blanpain, Alrick Dias, Sophie Gachet, Karina Gibert, Julien Lecubin, Michèle Leydet, Isabelle Mougenot, Christian Surace, Thierry Tatoni et la communauté IndexMed, "IndexMed : new tools using European Pollen DataBase for indexing, visualizing and data mining based on graphs" ***Meeting 2016 de l'European Pollen Database (EPD)*** à Aix-en-Provence, France.
- 14 – 15 janvier 2016 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, Loup Bernard, Cyrille Blanpain, Alrick Dias, Sophie Gachet, Karina Gibert, Julien Lecubin, Michèle Leydet, Isabelle Mougenot, Christian Surace, Thierry Tatoni et la communauté IndexMed, « Projet IndexMed : des réponses originales à l'hétérogénéité des données écologiques en Méditerranée » ***Séminaire LITEAU Observation et recherche en appui aux politiques du littoral et de la mer***, Brest, France. DOI : 10.13140/RG.2.1.2925.9280

- 14 – 15 janvier 2016 : **Romain David**, Dorian Guillemain, Abigail Cahill, Aurélien Dejode, Jean-Pierre Féral, Anne Chenuil, « Un an de recrutement benthique en domaine côtier méditerranéen sur des dispositifs ARMS et ASUs, premiers résultats. », ***Séminaire LITEAU Observation et recherche en appui aux politiques du littoral et de la mer***, Brest, France. DOI : 10.13140/RG.2.1.3720.9367
- 14 – 15 janvier 2016 : Dorian Guillemain, **Romain David**, Denise Bellan-Santini, Jean-Pierre Féral, Dorothée Meyer, ZNIEFF Mer PACA « Mise à jour de l'inventaire des Zones Naturelles d'Intérêt Ecologique Faunistique et Floristique Marines de la région PACA, perspectives d'améliorations ». ***Séminaire LITEAU Observation et recherche en appui aux politiques du littoral et de la mer***, Brest, France. DOI : 10.13140/RG.2.1.3768.8401
- 10 novembre 2015 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, Loup Bernard, Cyrille Blanpain, Alrick Dias, Sophie Gachet, Karina Gibert, Julien Lecubin, Michèle Leydet, Isabelle Mougenot, Christian Surace, Thierry Tatoni et la communauté IndexMed : "Improving quality of heterogeneous data with IndexMed consortium, a challenge for data mining in Mediterranean ecology, JOIN the CONSORTIUM, add value to your data!!" ***Conseil Consultatif de VegFrance***, Paris, France.
- 5 novembre 2015 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, Loup Bernard, Cyrille Blanpain, Alrick Dias, Sophie Gachet, Karina Gibert, Julien Lecubin, Michèle Leydet, Isabelle Mougenot, Christian Surace, Thierry Tatoni et la communauté IndexMed : "Fouille de données hétérogènes en écologie : les besoins émergeants en infrastructures de calcul du consortium IndexMed", ***journées SUCCES - Rencontres Scientifiques des Utilisateurs de Calcul intensif, de Cloud et de Stockage***, 5 et 6 novembre 2015 à Paris - DOI: 10.13140/RG.2.1.1498.0564
  https://succes2015.sciencesconf.org/resource/page/id/8
- 20 Octobre 2015 : **Romain David**, Jean-Pierre Féral, Anne-Sophie Archambeau, Loup Bernard, Cyrille Blanpain, Alrick Dias, Sophie Gachet, Karina Gibert, Julien Lecubin, Michèle Leydet, Isabelle Mougenot, Christian Surace, Thierry Tatoni et la communauté IndexMed : , "Improving quality of heterogeneous data with IndexMed consortium, a challenge for data mining in Mediterranean ecology that can supply EMODnet.", ***EMODnet Open Conference (European Marine Observation and Data Network): Consolidating the Foundations, Building the Future***, Ostende, Belgium - DOI: 10.13140/RG.2.1.4896.0721

- 20 octobre 2015 : Laure Thierry De Ville D'avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral, “Ecosystem services of coralligenous habitats under climate change.”, ***Conférence international MISTRALS***, Marseille, France.
- 15 Octobre 2015 : **Romain David**, Cyrille Blanpain, Jean-Pierre Féral, Christian Surace, “Prototype d’indexation, de visualisation et de fouille de données hétérogènes en écologie dans le cadre du consortium IndexMed : appel à participations”, ***Journée thématique du PR2I Big Data sur le Traitement de données massives*** DOI : 10.13140/RG.2.1.4018.8880
- 21 – 25 septembre 2015 : **Romain David**, Christos Arvanitidis, Melih Ertan Çinar, Drossos Koutsoubas, Stéphane Sartoretto, Anne Chenuil, Giulia Gatti, Jean-Pierre Féral, « Défis et coopérations scientifiques concernant les habitats marins méditerranéens « coralligènes » dans le cadre du programme européen CIGESMED », ***Ecology at the Interface: Science–Based Solutions for Human well Being***, Rome, Italie.
- 10 – 14 August 2015 : Abigail Cahill, Didier Aurelle, Emilie Boissin, Aurélien De Jode, Emilie Egea, Zinovia Erga, **Romain David**, Sophie Dubois, Jean-Baptiste Ledoux, Bastien Mérigot, Alexandra Weber, Anne Chenuil, “Determinants of connectivity in the marine environment: a multispecies approach”, ***XV congress of the European Society for Evolutionary Biology***, Lausanne, Suisse.
- 6 – 9 September 2015 : Chris Smith, Nadia Papadopoulou, Katerina Sevastou, Anita Franco, Heliana Teixeira, Chiara Piroddi, Stelios Katsanevakis, Karin Fürhaupter, Olivier Beauchard, Sabine Cochrane, Silje Ramsvatn, Jean-Pierre Féral, Anne Chenuil, **Romain David**, Niki Kiriakopoulou, Anastasija Zaiko, Snejana Moncheva, Kremena Stefanova, Tanya Churilova, Olga Kryvenko. “Keystone species across European seas: a species catalogue” in ***Unbounded boundaries and shifting baselines: Estuaries and coastal seas in a rapidly changing world***, London, UK. http://www.estuarinecoastalconference.com/special-session-everyone.html
- 7 – 10 July 2015 : Laure Thierry De Ville D'avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral : Ecosystem services of coralligenous habitats under climate change, IMBE - GREQAM, Marseille, France : ***colloque «Our Common Future Under Climate Change », International Scientific Conference of COP21***, Paris, France DOI: 10.13140/RG.2.1.4239.2089
- 28 juin – 2 juillet 2015 : **Romain David**, Cyrille Blanpain, Jean-Pierre Féral, Christian Surace, with contributors : The challenge of indexing Mediterranean ecological data, call for scientific cooperation under the consortium IndexMed, Formation ***FOCOLISE, école thématique CNRS/EGC/Unistras***, Strasbourg, France DOI: 10.13140/RG.2.1.5012.5923

- Aurélien De Jode, Sophie Dubois, Zinovia Erga, Romain David, Dorian Guillemain, Jean-Pierre Féral, Frédéric Zuberer, Anne Chenuil, with contributors « Understanding ecological functioning of coralligenous habitats, and building New Indicators based on genetic tools to assess their GES (good environmental status) » colloque ***"Aix-Marseille et la Méditerranée : défis et coopérations scientifiques"***, Marseille, France.
- **Romain David**, Cyrille Blanpain, Jean-Pierre Féral, Christian Surace, with contributors « Le défis de l'indexation des données en écologie méditerranéenne : Appel à coopérations scientifiques dans le cadre du consortium IndexMed », colloque ***"Aix-Marseille et la Méditerranée : défis et coopérations scientifiques"***, Marseille, France.
- **Romain David**, Christos Arvanitidis, Melih Ertan Çinar, Drossos Koutsoubas, Stéphane Sartoretto, Anne Chenuil, Jean-Pierre Féral, « Défis et coopérations scientifiques concernant les habitats marins méditerranéens « coralligènes » dans le cadre du programme européen CIGESMED », colloque ***"Aix-Marseille et la Méditerranée : défis et coopérations scientifiques"***, Marseille, France.
- Katerina Sevastou, Nadia Papadopoulou, Chris Smith, Heliana Teixeira, Chiara Pirodd, Stelios Katsanevakis, Jean-Pierre Féral, Anne Chenuil, **Romain David**, Niki Kiriakopoulou, Sabine Cochrane, (2015) - Defining keystone species in European regional seas: what are the candidates for the Mediterranean? In: ***11th Panhellenic Symposium on Oceanography & Fisheries «Aquatic Horizons: Challenges & Perspectives»***. Mytilene, Lesvos Island, Greece, 13-17 May 2015. Athens: H.C.M.R., 553-556.
- 21 – 22 avril 2015, Laure Thierry De Ville D'avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral, "Evaluation des services écosystémiques produits par les habitats coralligènes - Revue des méthodes économiques et suivis écologiques pertinents." ***Congrès de l'école doctorale des sciences de l'environnement***, Cassis, France,
- 2-5 December 2014 : Chris Smith, Nadia Papadopoulou, Katerina Sevastou, Anita Franco, Heliana Teixeira, Chiara Piroddi, Stelios Katsanevakis, Karin Fürhaupter, Olivier Beauchard, Sabine Cochrane, Silje Ramsvatn, Jean-Pierre Feral, Anne Chenuil, **Romain David**, Niki Kiriakopoulou, Anastasija Zaiko, Snejana Moncheva, Kremena Stefanova, Tanya Churilova, Olga Kryvenko., (2014) - Report on identification of keystone species and processes, ***Devotes annual meeting***, Ancona, Italie.
- 24 novembre 2014 : Zinovia Erga, **Romain David**, Aurélien De Jode, Marc Verlaque, Sophie Dubois, Line Le Gall, Dorian Guillemain, Frédéric Zuberer, Jean-Pierre Féral,

Anne Chenuil, 2014. Strong genetic structure in Lithophyllum *cabiochiae/stictaeforme* around Marseilles. Local adaptation, niche differentiation or just lack of connectivity? ***Annual conference of French Phycology Society***, Paris, France.

- 29 – 30 October 2014, Laure Thierry De Ville D'avray, Dominique Ami, Anne Chenuil, **Romain David**, Jean-Pierre Féral : with contributors, How food security can be linked to the ecological and economic status of the coralligenous habitats of the Mediterranean Sea? ***Biodiversity and Food Security – From Trade-offs to Synergies,*** Aix-en-Provence, France, DOI10.13140/2.1.1657.1205
- 29 – 30 octobre 2014 : **Romain David**, Christos Arvanitidis, Melih Ertan Çinar, Stéphane Sartoretto, A Doan, Sophie Dubois, Zinovia Erga, Dorian Guillemain, Laure Thierry de Ville d'Avray, Frederic Zuberer, Anne Chenuil-Maurel, Jean-Pierre Féral, « CIGESMED habitat's characterization : a simple and reusable typology at the Mediterranean scale » ***2ème Symposium CAR ASP : The 2nd Mediterranean symposium on Coralligenous and other calcareous bio-concretions,*** Portoroz, Slovenie. DOI10.13140/2.1.1607.7763
- 29 – 30 octobre 2014 : Melih Ertan Çinar, Jean-Pierre Féral, Christos Arvanitidis, **Romain David**, Ergun Taskin, Thanos Dailianis, Alper Dogan, Vasilis Gerovasileiou, Ertan Dagli, Veysel Aysel, Yiannis Issaris, Kerem Bakir, Maria Salomidi, Maria Sini, Sermin Açik Cinar, Alper Evcen, Charalampos Dimitriadis, Drossos Koutsoubas, Stéphane Sartoretto, Senem Önen, (2014) - "Preliminary assessment of macrobenthic coralligenous assemblages across the Mediterranean sea" ***2ème Symposium CAR ASP : The 2nd Mediterranean symposium on Coralligenous and other calcareous bio-concretions,*** Portoroz, Slovenie. DOI10.13140/2.1.3246.1768
- 09 – 12 octobre 2014 : Ζηνοβία Εργά, **Romain David**, Dorian Guillemain, Frédéric Zuberer, Θάνος Ντάϊλιάνης, Βασίλης Γεροβασιλείου, Μαρία Σίνη, Δρόσος Κουτσούμπας, Marc Verlaque, Anne Chenuil, Jean-Pierre Féral, (2014) - Study of the genetic diversity of the rhodophyte Lithophyllum stictaeforme/cabiochiae in the North Ouest Mediterranean. ***7th Helecos congress***, Grèce.
- 17 – 19 june 2014 : **Romain David**, Sophie Dubois, Zinovia Erga, Dorian Guillemain, Laure Thierry de Ville D 'Avray, Christos Arvanitidis, Melih Ertan Çinar, Stéphane Sartoretto,Frederic Zuberer, Anne Chenuil-Maurel, Jean-Pierre Féral, "CIGESMED* or how to integrate combined scientific and amateur large sets of heterogeneous data on coralligenous habitats for the effective evaluation of the "Good Environmental Status" of the Mediterranean Sea?", ***Fifth Symposium Monitoring of Mediterranean coastal areas: problems and measurement techniques,*** Livorno, Italy.

- 28 – 30 April 2014 : Sophie Dubois, **Romain David**, Alexander Ereskovsky, Zinovia Erga, Anne Chenuil, Jean-Pierre Féral, Genetic structure of populations of the bryozoan *Myriapora truncata*, a bio-builder and component of the Coralligenous habitat. ***International Mediterranean Conservation Sciences Conference for the young scientists***, La Tour du Valat, France. https://doi.org/10.13140/rg.2.1.4894.5683
- 28 – 30 April 2014 : Zinovia Erga, **Romain David**, Anne Chenuil, Jean-Pierre Féral, CIGESMED : Cryptic species and genetic structure population of the corallinale complex of species « Lithophylum cabiochae / stichtaeforme », builder and component of the coralligenous, ***International Mediterranean Conservation Sciences Conference for the young scientists***, La Tour du Valat, France.
- 28 – 30 April 2014 : Guillemain D., **Romain David**, De Ville d'Avray L., Féral J.P., CIGESMED : Implementation of coralligenous habitats characterization protocols for the effective evaluation of the good environmental status of the Mediterranean Sea. Cartography of coralligenous habitats by profiles surveys, ***International Mediterranean Conservation Sciences Conference for the young scientists***, La Tour du Valat, France.
- 28 – 30 April 2014 : Laure De Ville d'Avray, **Romain David**, Dorian Guillemain, Christos Arvanitidis, Anne Chenuil, Melih Ertan Çinar, Stéphane Sartoretto, Jean-Pierre Féral, CIGESMED : Coralligenous populations study based on spatial variability of habitat profiles, by the mean of photo-quadrats and image processing software (PhotoQuad), ***International Mediterranean Conservation Sciences Conference for the young scientists***, La Tour du Valat, France.
- 2013 : **Romain David**, CIGESMED* intercalibration of data and observations on coralligenous habitats for the effective evaluation of the "Good Environmental Status" of the Mediterranean Sea? - ***Séminaire interne UMR CNRS-IRD, AMU-UAPV Institut Méditerranéen de Biodiversité et d'Écologie Marine et Continentale*** - Ingénierie Écologique).
- 28 October – 1 November 2013 : Sandrine Ruiton, Charles Boudouresque, **Romain David**, Florent Renaud, (2013) - An ecosystem-based approach to evaluate the status of Mediterranean ecosystems habitats (extended to WFD, MSFD and other marine environmental policy), ***40th CIESM Congress***, Marseille, France, 28 October - 1 November 2013
- 21 – 27 octobre 2013 : **Romain David**, Christos Arvanitidis, Anne Chenuil, Melih Ertan Çinar, Anthony Fremaux, Stéphane Sartoretto, Jean-Pierre Féral, CIGESMED* intercalibration of data and observations on coralligenous habitats for the effective evaluation of the "Good Environmental Status" of the Mediterranean Sea? ***IMPAC3***

3ème congrès international des aires marines protégées, Marseille, France, DOI10.13140/2.1.2768.4802

- 27 – 28 mars 2012 : **Romain David**, Alain Pibot, The French Nature and Landscape Information System (SINP) : Strategy for research, ***Ateliers scientifiques de façade de l'Agence des Aires Marines Protégées (Cartham - cartographie des Habitats Marins)***, Marseille, France.
- 18 – 21 octobre 2011 : Agnès Pouliquen, **Romain David**, Alain Pibot, The French Nature and Landscape Information System (SINP) : a collaborative tool for biodiversity data, information and knowledge sharing, ***International symposium "effect of climate changes"***, Biarritz, France, DOI : 10.13140/RG.2.1.4632.4242

Organisation de séminaires et d'ateliers

2017

- 04-07 juillet : JDEV Journées nationales du DÉVeloppement logiciel, Marseille Animation du groupe de travail « Outils, compétences et savoir-faire nécessaires à la fouille de graphes : quelques cas d'utilisation et les défis qui les accompagnent »
- 22-23 juin : Journées « sciences des données » du GDR MaDICS et action GRAMINÉES : recueil des attentes en compétences et en disponibilité en compétences, 140 participants
- 23 juin : "Visualisation de données" sous forme de graphes dans le cadre du consortium IndexMEED, Paris, 15 participants http://www.madics.fr/event/journees-madics-2017-23-juin/
- 15 mai, 29 mai et 26 juin : Ateliers Curation de données organisé au CESAB et à la à l'ISCC dans le cadre du consortium IndexMEED, Aix en provence, Paris, 12, 4 et 15 participants
  https://docs.google.com/document/d/1lEm8U-MQ2nSTHQnL2hbSpSZF5BoOVun0DQAjz7L9UUM/edit

2016

- Last Meeting Working Group 2 CIGESMED Work Shop Datapapers and IPT what and what for? Concepts: accessibility / norms / metadata / Open data and which data for the Data papers Implementation, Station Marine d'Endoume, Marseille, 12 participants
- Last Meeting Working Group 1 CIGESMED Work Shop Thesaurus (presentation of definitions by domain) and introduction about ontologies, Villa Station Marine d'Endoume, Marseille, 12 participants.

- Working Group IndexMed, Fouille de données à l'aide de graphes, questionnements scientifiques autour de vos données - CESAB, Aix en Provence, 33 participants.
- Working Group IndexMed, Curation de données dans l'objectif de construire des graphes : prérequis, bonnes pratiques, standards et verrous scientifiques - CESAB, Aix en Provence, 33 participants.

2015

- Working Group 2 CIGESMED, « Pertinence/Relevance of indicators : what is a successful indicator ?» - Villa Valmer, Marseille, 18 participants.
- Working Group 1 CIGESMED, « WG Graph » et « WG Thesaurus » - Station Marine d'Endoume, Marseille, 20 participants.
- Atelier prospectif consortium IndexMed, « Méthode et outils de fouille de données hétérogènes en écologie » - co animation avec Nicolas Bailly (HCMR), Marseille, 50 participants.
- Forum ouvert région PACA, Sciences participatives, biodiversité et changement climatique : vers un protocole à la carte, Marseille Région PACA, 20 participants.
- 22 mai : « CIGESMED Coralligenous thesaurus », 3éme Assemblée Générale CIGESMED à Mytilène, Grèce, 25 participants
- 21 mai : « Working Group on coralligenous new definition », 3éme Assemblée Générale CIGESMED à Mytilene, Grèce, 25 participants

2014

- 5 juin : animation de la table ronde Indexmed : interopérabilité des bases de données en écologie, Marseille Saint Charles, 65 participants
- 6 mai : « Working Group on Knowledge Trees (KT) and other mapping methods », 2éme Assemblée Générale CIGESMED à Izmir, Turquie, 25 participants
- 5 mai : « Working Group on Protocols », 2éme Assemblée Générale CIGESMED à Izmir, Turquie, 25 participants

Rapports liés au travail de thèse

2017

- **Romain David**, Anna Cohen Nabeiro, Aurélie Delavaud, Loup Bernard, Guillaume Body, Yvan Le Bras, Michelle Leydet, Consortium IndexMEED, 2018. Visualisation sous forme de graphes de données en écologie et environnement : retour sur les ateliers 2017, mission 2017.

2016

- Jean-Pierre Féral, Christos Arvanitidis, Anne Chenuil, Melih Ertan Çinar, **Romain David**, Emilie Egea, Stephane Sartoretto, (2016) - CIGESMED : Coralligenous based Indicators to evaluate and monitor the « Good Environmental Status » of the Mediterranean coastal waters. SeasEra project Towards Integrated Marine Research Strategy and Programmes. Final report, 179 pp. DOI : 10.13140/RG.2.2.35960.03848 〈hal-01448881〉
- Chris Smith, Nadia Papadopoulou, Katerina Sevastou, Anita Franco, Heliana Teixeira, Chiara Piroddi, Stelios Katsanevakis, Karin Fürhaupter, Olivier Beauchard, Sabine Cochrane, Silje Ramsvatn, Jean-Pierre Feral, Anne Chenuil, **Romain David**, Niki Kiriakopoulou, Anastasija Zaiko, Snejana Moncheva, Kremena Stefanova, Tanya Churilova, Olga Kryvenko. (2014) - Report on identification of keystone species and processes, July 2014 : Report number: Deliverable 6.1, DEVOTES Project. 105 pp + 1 Annex, DOI10.13140/2.1.3227.7763

Stages co-encadrés avec J.P. Féral pendant la thèse

2015

- Jacky Dubar, (2015) - Mise en oeuvre d'un protocole pour l'étude de la composition taxonomique des habitats coralligènes en fonction de la luminosité. Rapport de stage, Technicien Supérieur de la Mer option Génie de l'Environnement Marin, CNAM-Intechmer (Cherbourg), 34 pp.

2014

- Wallid El Guerrabi, (2014) - Qualification de données de CIGESMED : Un système de prospection de données pour l'aide à la gestion des habitats coralligènes. Rapport de stage d'école d'ingénieur. 47 pp. + annexes.
- Dorian Guillemain, (2014) - Etude de la variabilité de la structure des peuplements coralligènes sur des îlots marseillais à profondeur constante dans le cadre du programme européen SeasEra CIGESMED. Rapport de stage de Master 2. 23 pp. + annexes.
- Laure Thierry De Ville d'Avray, (2014) - Etude de la variabilité de la composition observée d'habitats coralligènes liée à l'application d'un protocole d'observation au moyen de quadra-photos. Mémoire de fin d'étude, master d'océanographie, spécialité biologie et écologie marine, Université d'Aix-Marseille, Marseille. 34 pp.

Annexe 8. Résultat des ateliers : document “how to do” sur la curation

La curation de données environnementales étape par étape

en vue d’une analyse reposant sur la théorie des graphes

Ce document est issu des échanges ayant eu lieu au cours des ateliers “Curation de données pour la visualisation et l’analyse de données s’appuyant sur les méthodes de graphes”, organisés par le consortium IndexMEED en 2017. Cette première version n’est pas un document complet, et a pour vocation de sensibiliser sur les différents aspects de la curation (Il ne décrit pas en l’état précisément les processus). Il sera amendé lors des prochains ateliers.

Pour citer ce document : Romain David, Anna Cohen-Nabeiro, Aurélie Delavaud, Michelle Leydet-Barbier, Sophie Pamerlon, Anne Quesnel-Barbet, Loup Bernard, 2018, “La curation de données environnementales étape par étape en vue d’une analyse reposant sur la théorie des graphes”,

Pour plus d’informations, vous pouvez :

- Consulter le site internet https://indexmeed2017.sciencesconf.org/
- Envoyer un mail à romain.david@imbe.fr et à ecoscope@fondationbiodiversite.fr

Objectifs du document : Décrire, étape par étape, les actions préconisées sur un jeu de données pour préparer puis importer des données dans un logiciel de graphes, en mettant en avant les points de difficultés, les erreurs à ne pas commettre et les bonnes pratiques.

Sommaire

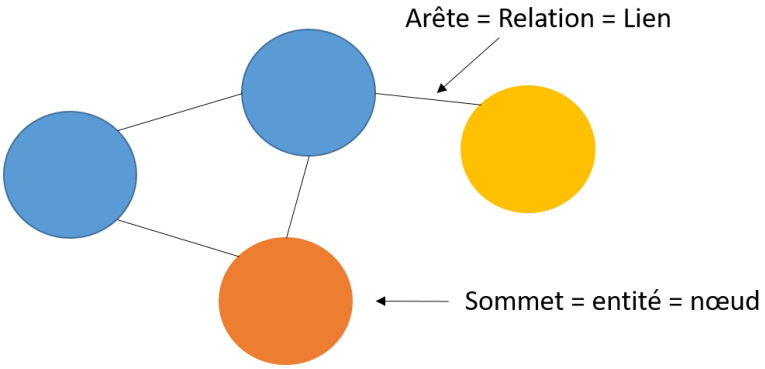
1) Enjeux et principes de la curation de données
2) Avant : la mise en place du protocole de curation de données
3) Pendant : l'application du protocole de curation de données en vue d'une représentation sous forme de graphes
4) Après : le contrôle de la qualité de la curation de données

Enjeux et principes de la curation de données

Dans le domaine de l'environnement, la fréquence des campagnes de collecte de données (missions de terrain, capteurs optiques ou radar, suivi de la qualité des eaux, recensement automatique ou semi-automatique des taxons, etc.) permettent d'acquérir un volume considérable de données. A cause de leur hétérogénéité, celles-ci sont néanmoins difficiles à agréger pour avoir une vue d'ensemble (on parle parfois d'empilement de bases de données). Améliorer l'interopérabilité nécessite de s'appuyer sur la curation de ces données avec une sémantique commune.

Pour curer un jeu de données, il faut aller plus loin que le "nettoyage" de ses imperfections. La curation de données, du latin *curare* qui signifie "prendre soin", est essentielle avant tout processus d'analyse ; elle consiste à améliorer la capacité des données à décrire un système de manière **univoque et explicite**. Elle est essentielle pour préparer un jeu de données pour chaque nouveau type d'analyse ou agréger différents jeux de données d'origines, de structures et de formats différents. La curation permet notamment de construire des graphes, qui sont pertinents pour modéliser plus efficacement les facteurs "composantes" des interactions biologiques malgré leur hétérogénéité.

Rappel : *Un graphe est un ensemble de points que l'on appelle des* ***nœuds*** *(sommets en mathématique ou objets en informatique) reliés par des traits (segments) ou flèches nommées arêtes (ou bien* ***liens****, arcs ou* ***attributs****). L'ensemble des arêtes entre les nœuds forme une figure similaire à un réseau. La représentation de données sous forme de graphes permet de relier des objets (champs/****entités*** *de la base de données ou valeurs de ces champs) ayant des natures différentes (valeurs quantitatives, qualitatives ordonnées ou non ordonnées).*



Dans l'optique de créer des graphes, il faut choisir dans la base de données les "objets" (c'est-à-dire le ou les champs de la base de données) qui vont servir de "noeuds"/"sommets" aussi appelé "entités", et les champs descripteurs (que l'on appelle attributs) des objets choisis pour en faire des liens/arêtes aussi appelés "relations".

Toute qualification de la donnée ou requalification de la donnée peut être utilisée comme nouvel objet ou comme attribut d'un objet existant, à condition de suivre un processus d'adaptation.
Un processus de nettoyage des erreurs doit d'abord être mis en place. Ensuite, un processus de requalification des données peut être mis en place. Les principales transformations que vous allez effectuer sur les données sont de deux ordres : celles qui permettent de découvrir la donnée (et ses qualités et potentiels) et celles qui permettent de les améliorer. Vous devrez à partir de l'étape de découverte faire des choix concernant les entités et sur le type de relations et de graphes que vous allez construire.
Voici quelques exemples d'étapes (qui doivent être adaptées au type de données) :

- Sélectionner les enregistrements valides, publiés ou non, retirer les archives, sélectionner un type (il s'agit là de choisir le périmètre de sa représentation),
- Choisir les attributs suffisamment renseignés pour l'analyse,
- Compléter les valeurs manquantes lorsque c'est possible,
- Retirer les entités sans attribution de valeurs pour le(s) descripteur(s) choisi(s) (sinon, ce sera une entité reliée à rien),
- Compter pour chaque entité le nombre de valeurs de descripteurs indiquées, les classer et estimer ainsi l'intérêt de les représenter sous forme de graphe (il faut qu'il y ait idéalement plusieurs valeurs de descripteurs pour que les noeuds soient reliés entre eux par un réseau assez dense). Il s'agit là d'un élément charnière de l'étape de découverte, qui dépendra des critères choisis pour sélectionner les entités,
- Si chaque entité n'a qu'une valeur par attribut et/ou chaque valeur d'attribut ne correspond qu'à une seule entité, il peut être souhaitable par exemple de combiner un autre descripteur (pour créer un autre type de lien) ou bien de regrouper des valeurs du descripteur en une valeur les comprenant tous les deux (exemple : "pêche artisanale" et "pêche industrielle" deviennent "pêche"). Ces choix doivent être fait en considérant autant les critères informatiques (qualité des données) que leur signification scientifique. Il s'agit là d'un élément charnière de l'étape de requalification, qui aura un impact fort sur le type de graphe représenté.

<u>Préparer la curation de données nécessite la mise en place d'un protocole</u>

Objectifs de cette étape :

- Identifier et quantifier les tâches nécessaires pour aboutir à un fichier importable contenant des données consistantes (c'est-à-dire sans erreurs et bien renseignées).
- Produire les documents qui permettront de se resservir des données transformées et bien comprendre le travail effectué (élaboration des dictionnaires de données, plan de curation, stratégies d'identification et de traçabilité de chaque version, règles d'usages).

Contenu de cette étape

- Analyse des points d'amélioration et de leurs difficultés (avec différents types d'acteurs : informaticiens et spécialistes du domaine concernant les données) correspondant à des transformations qui peuvent être de deux types : corrections et requalifications.
- Définition des moyens humains et des outils disponibles pour chaque tâche/transformation pour chaque type d'acteur.
- Définition des tâches/transformations à effectuer, priorisation en fonction des objectifs analytiques et leur rapport "coût avantage".

- Allocation des ressources aux tâches avec plusieurs scénarios (du possible au souhaitable, à court et long terme) -> ces informations sont à insérer et documenter dans le plan de gestion des données.
- Priorisation des tâches en fonction de leurs effets escomptés sur les verrous à l'analyse ou sur leurs efficacités en terme d'obtention de résultats à court et long terme.
- Regrouper ces éléments dans un document "Protocole de curation de données".

Points de difficulté

- Prévoir à long terme le maintien de ces processus de normalisation ou de suivi de l'évolution des standards (et donc les ressources nécessaires).
- Bien décrire les étapes de transformation de la donnée pour chaque tâche, et rendre le plus explicite possible cette documentation "Protocole de curation de données" (imaginer qu'elle doit pouvoir être reproduite par un opérateur qui ne connaît pas le jeu de données).
- Travailler sur la précision des définitions liées aux champs de bases de données dans le dictionnaire de données (cela doit pouvoir être compris et reproduit par un opérateur qui ne connaît pas le jeu de données), surtout lors d'un empilement de données hétérogènes et multi-sources (c'est-à-dire, issues de différents observateurs / systèmes d'observation) - lister les incohérences en cas de données multi-sources.

Conseils et bonnes pratiques

- Se référer en premier lieu aux standards existants dans le domaine interdisciplinaire (normes INSPIRE, EML, Géographiques GML, Darwin Core...) pour ce qui est des métadonnées obligatoires, les données standardisées (+ vocabulaires contrôlés, thésaurus si aucun standard n'est disponible). L'utilisation des standards garantit une meilleure interopérabilité des données si elles doivent être agrégées avec d'autres. Ces standards sont officiellement agréés par les États via une organisation nationale de standardisation (Afnor pour la France), agréés au niveau Européen (comme le CEN ou le ETSI), ou encore issu d'un traité international (comme ISO - International Organization for Standardization).
  Exemple : Les formats de la date et de l'heure doivent être standardisés suivant la norme ISO 8601 (https://fr.wikipedia.org/wiki/ISO_8601). Il y a six niveaux de granularité dans ce format, selon les applications.
  - La date ISO pour le 14 août 2017 s'écrit 2017-08-14.
  - L'heure ISO s'écrit 20:54:15Z pour 20h54m15s.
  - La date et l'heure ISO s'écrivent 2017-08-14T20:54:15Z
- Construire un dictionnaire de données (c'est-à-dire, lister les champs de la base de données, puis les définir de manière explicite et univoque en utilisant si possible des référentiels existants, et enfin décrire le format en donnant un exemple).
- Lister les contraintes d'intégrité que vous imposez lors de la saisie et de la correction de données (par exemple une donnée obligatoire : "le champ email principal doit être rempli pour valider la saisie d'un nom", celui ci doit contenir le caractère @).
- Tester (et chronométrer) la durée de certaines transformations sur un petit lot de données.
- Relever toutes les opérations nécessaires et les décrire. Exemples : lister les caractères "interdits" comme par exemple les caractères accentués qui ne sont pas codés de la même manière d'un système d'exploitation à un autre, le type de séparateurs à changer pour respecter les recommandations et usages les plus

répandus ";" proscrire dans les noms de champs les caractères spéciaux ou accentués, remplacer les blancs par underscore ("_").

Appliquer le protocole de curation de données en vue d'une représentation sous forme de graphes

Objectif de cette étape

- Rendre le jeu de données importable dans un logiciel de visualisation puis analysable par des algorithmes d'analyse de graphes.

Contenu de cette étape

- Appliquer les transformations aux données une par une, en vérifiant le résultat entre chaque opération.
- Remplir le journal des transformations en y inscrivant le numéro de version de votre jeu de données, vérifier que la transformation est suffisamment bien décrite dans le journal (et qu'elle correspond à celle inscrite dans le protocole de curation).
- Qualifier la donnée (voir le détail ci-dessous) et vérifier que le qualificatif et ses valeurs possibles sont bien décrits dans le dictionnaire de données fait pendant la préparation du plan de curation.
- Versionner le fichier une fois les transformations validées (idéalement avec une nomenclature de nommage décrite dans le plan de curation).
- Extraire et re-structurer les données en vue de les visualiser sous forme de graphe.

La qualification (ou requalification) de la donnée

- La qualification (ou requalification) de la donnée est un enrichissement de la donnée / de la base de données par de nouveaux champs nettoyés ou transformés (en regroupant des catégories en une seule par exemple).
- Qualifier une donnée, c'est aussi donner une information supplémentaire sur celle-ci : est-elle fiable ? L'instrument de mesure/l'observateur aussi ? A-t-elle été validée ?
- Requalifier une donnée c'est aussi une manière de transformer les valeurs et/ou le format d'un champ (par exemple, un âge est transformé en une catégorie d'âge).

Restructuration des données

- Pour être visualisées sous forme de graphe, les informations concernant les noeuds et les liens doivent être structurées différemment selon le type de logiciel de visualisation. Le logiciel Tulip est le plus souple car il est possible d'utiliser un seul tableur avec une ligne par entité (ou noeud) et d'aller générer la ou les table(s) de relation (ou liens).

	A	B
1	source	target
2	dataset_11	keyword_326
3	dataset_11	keyword_327
4	dataset_161	keyword_328
5	dataset_161	keyword_329

Extrait de la table de liens du jeu de données ECOSCOPE.

	A	B	C	D
1	id	label	type	categorie
2	dataset_11	suivi des peuplements	dataset	pub
3	dataset_161	sheep ex situ collection for the french national cryobank	dataset	pub
4	dataset_162	rainbow trout ex situ collection for the french national cryobank	dataset	pub
5	dataset_164	pig ex situ collection for the french national cryobank	dataset	pub

Extrait de la table de noeuds du jeu de données ECOSCOPE.

- Plusieurs points doivent attirer notre attention pour que le jeu de données soit visualisable sous forme de graphe :
    - Définir les noeuds potentiels (autrement dit, quels sont les champs que je définis comme étant des entités),
    - Définir les liens potentiels pour chaque type de noeuds choisis à l'étape précédente (autrement dit, quels sont les champs que je définis comme étant des relations entre les entités / faisant le lien entre deux noeuds),
    - Re-formater le fichier selon la structure demandée par le logiciel. Ressources selon le logiciel : Tulip, Gephi, Neo4j...

Quelques actions fréquentes

- Rechercher un à un les caractères "interdits", c'est-à-dire ceux qui peuvent poser un problème soit en informatique (accents dans les noms de champs), soit en terme d'analyse (opérateurs dans des champs numériques par exemple, homogénéiser les séparateurs décimaux comme les points ou les virgules).
- Rechercher et corriger/supprimer les doublons et fautes de frappe (avec par exemple avec les filtres d'Excel ou les facettes d'OpenRefine).
- Contrôler le respect des règles d'intégrité. (Une contrainte d'intégrité est une règle qui définit la cohérence d'une donnée ou d'un ensemble de données de la base de données. Par exemple, une adresse dans la table des adresses doit correspondre à au moins un propriétaire.)
- Quelques exemples d'actions de nettoyage dans le processus de curation : Ne pas laisser de cellules vides, définir et décrire les attributs et les valeurs qui peuvent être donnés aux objets dans un fichier à part, suivre un format standard pour les champs normés (comme les dates par exemple), ne pas mettre plusieurs informations dans la même cellule (l'unité d'une mesure par exemple).

Points de difficulté

- Anticiper la perte d'information, par exemple avec un "copier remplacer" qui transforme deux valeurs différentes en deux valeurs identiques. (Exemple : Jean DAVID et Jacques DAVID deviennent J. DAVID).
- Attention aux champs remplis par un espace ou un autre caractère invisible, par exemple un i majuscule et un L minuscule, ils peuvent dé-doubler la valeur d'un descripteur (ces doublons peuvent être mis en évidence avec OpenRefine par exemple).
- Attention aux caractères spéciaux, ils peuvent avoir selon le système d'exploitation sous lequel ils ont été saisis, des codes ascii différents, ce qui leur donne des valeurs

distinctes pour l'ordinateur, sans que ce soit visible de la part de l'opérateur (encodage des données à vérifier lors de l'import sous Excel ou OpenRefine).
- Il est parfois difficile d'avoir à chaque fois des valeurs pour tous les descripteurs / de déceler les homonymes et les termes polysémiques.
- Attention : tout noeud qui est le seul à posséder une relation sera déconnecté du graphe, et donc inutile dans l'analyse. Pour éviter cet état de fait, il peut être utile de regrouper différentes valeurs d'attribut en une seule plus globale.

Conseils et bonnes pratiques

- Garder un original non modifié des données brutes, l'entreposer au moins en double et dans deux endroits différents.
- Dupliquer le champ avant de faire un chercher/remplacer dans le tableur / la base de données.
- Faire un check sur une ou plusieurs lignes connues suite à ces modifications.
- Sauvegarder une fois que les effets de la transformation ont été "attentivement" contrôlés, si possible par un autre opérateur.
- Tenir un répertoire des erreurs rencontrées et un journal des transformations effectuées (le remplir entre chaque transformation !)
- Enregistrer une nouvelle version du document à chaque nouveau type de transformation effectuée.
- Vérifier la validité du champ transformé et renseigner selon la typologie définie dans le protocole de curation pour décrire la validité des données.
- Il est possible et parfois conseillé de créer plusieurs descripteurs à partir d'un champ. (Cela peut être fait en transformant une valeur numérique en 2 catégories dans un champ, et en 3 catégories dans un autre, en regroupant les valeurs dans des classes comme les âges en classes d'âge ou les métiers en domaines professionnels.)
- Il faut être très explicite sur les choix de valeurs particulières. Parfois, NA pour "Non Available" est utilisé, mais NA peut aussi être "Non Attribué". NR signifie Non Renseigné, mais aussi Nombre de Restes en archéologie. Remplacer les vides par NA [se mettre d'accord sur NA, N/A, NR etc...] est une pratique qui demande d'être bien d'accord sur ce que signifiait un champ vide (absence de valeur ou d'observation, ou aucun individu observé, non renseignement...), et ce qu'il va signifier une fois remplacé par NA.

Contrôle qualité de la curation de données

Objectif de cette étape : Vérifier la qualité de la mise en oeuvre du protocole de curation. Plusieurs aspects de la qualité du jeu de données modifiés doivent être contrôlés : la transformation du jeu de données (et sa reproductibilité), le respect des standards utilisés et l'efficacité du protocole de curation (qui devra certainement être ré-appliqué et autant que possible automatisé pour les nouvelles données).

Contenu de cette étape

- Après chaque type d'analyse des données (statistiques inférentielles, algorithmie basée sur les parcours de graphes…), reconsidérer chaque transformation du protocole en évaluant leur efficacité/pertinence via les résultats de l'analyse.
- Après une évolution du système d'observation ou des standards, évaluer "l'état d'obsolescence" des données et le rapport "coût / avantage" de leur mise à jour.

- Corriger le protocole pour en recréer une nouvelle version intégrant les transformations pour une mise à jour et les transformations pour un nouveau type d'analyse.
- Ré-implémenter le protocole en fonction des conclusions tirées du contrôle qualité.

Point de difficulté

- Il est parfois nécessaire de réeffectuer le travail à partir des données "brutes" pour suivre l'évolution d'un standard / protocole ou pour envisager une autre série d'analyses.
- L'évaluation est difficile si l'on n'a pas été très explicite sur les transformations effectuées et leurs significations.
- Assurer la compatibilité avec les données antérieures lors de toute évolution de la base de données et/ou de ses interfaces.
- Gestion de l'obsolescence : lorsque le système d'observation et les protocoles associés évoluent, ou lorsque les standards considérés lors de la définition du dictionnaire de données évoluent, les données curées doivent suivre ces évolutions. D'une manière générale, chaque contrôle doit vérifier que ces évolutions sont intégrées dans la dernière version du jeu de données (cela fait partie du plan de gestion de données). NB : Certains systèmes permettent d'automatiser ces évolutions, et les standards proposent généralement une rétro-compatibilité avec leurs anciennes versions.

Conseils et bonnes pratiques

- Pour vérifier la validité d'un standard, il est rentable de considérer prioritairement les "changements dans le standard" qui sont généralement cités par les organismes qui les tiennent à jour.
- Les métadonnées doivent en particulier être mises à jour lors de ce contrôle qualité : les versions des standards et des protocoles utilisés.
- Adopter une politique de versionning et s'y tenir. Par exemple, mettre à jour la version soit de manière périodique, soit à chaque modification d'un enregistrement, soit à chaque apport d'un nouveau jeu de données, soit à chaque itération du processus de curation, soit à chaque nouvel usage / nouveau type d'usage du jeu de données. Ces choix et la manière de les typer doivent être documentés dans le plan de gestion des données et dépendent du type de cycle de vie de ces données. https://www.rd-alliance.org/groups/data-versioning-wg
- Réaliser des manipulations itératives sur les données (préparation, curation, vérification de la qualité).
- Tester la compatibilité des anciennes données avec les nouvelles et retravailler ces anciennes données, mais aussi prévoir que les nouvelles données puissent être analysées et confrontées aux anciens jeux de données. Documenter ces tests dans le plan de gestion des données (cette remarque est aussi valable pour les logiciels, les environnements de travail, les processus de traitement…) https://www.rd-alliance.org/groups/reproducibility-ig.html http://www.madics.fr/actions/actions-en-cours/
- Intégrer des étapes "curation de données" et leurs descriptions dans le plan de gestion de données (idéalement, on doit retrouver tous les protocoles de curation dans le plan de gestion des données).

- Prévoir des moyens et se fixer des règles de maintien et des limites par rapport à la gestion de l'obsolescence des données.
- Les graphes peuvent être utilisés comme outil d'analyse de la qualité d'un protocole de curation. Par exemple, les noeuds isolés, ou hors de toute relation peuvent l'être à cause d'erreurs ou de champs non/mal renseignés.

En perspective

Le mille-feuille des acteurs et des bases de données en écologie est aujourd'hui un frein à l'inter-opération. Chaque organisme a adopté un système qui répond à ses propres besoins, avec ses propres syntaxes, structures et éléments sémantiques. Ces éléments relèvent souvent plus du « jargon métier » que de véritables concepts établis de manière partagée par toutes les communautés travaillant dans le domaine de la biodiversité.

Le travail itératif de curation de données en vue de produire des graphes permet d'augmenter le potentiel d'analyse de ces données et d'améliorer les possibilités d'inter-opérations des systèmes d'information communautaires en vue d'analyses plus intégratives. Ce travail n'est bien sûr possible qu'avec une donnée partagée respectant les principes FAIR (Findable, Accessible, Interoperable, and Re-usable).

Un catalyseur essentiel facilitant et augmentant l'efficacité de ces démarches de curation réside dans la participation des différents acteurs de la donnée scientifique aux communautés interdisciplinaires structurant standards, outils et principes autour du partage de la donnée et sa reconnaissance.

Deux verrous sont actuellement à lever pour i) améliorer cette participation de toutes les parties prenantes, ii) développer l'inter-opération entre bases de données et iii) évoluer vers un cycle de vie de la donnée reprenant les principes d'une démarche qualité : Ces deux verrous sont i) la disponibilité en compétences pour préparer les données et former les chercheurs (curation, data management plan, visualisation) et ii) une gouvernance établie des systèmes d'évaluation du respect des standards qui nécessite des organes pérennes et reconnus par tous.

Ressources et outils sur :

Les principaux standards et référentiels cités :

- http://www.esabii.biodic.go.jp/ap-bon/meetings/documents/20_article01_07_APBON-Symposium-GBIF.pdf

Les outils facilitant la curation :

- Galaxy-E outil Workflow, permettant de lister le process de l'utilisateur
- L'utilisation de Openrefine + Galaxy permet de traiter l'aspect reproductibilité et la portabilité des processus analytiques. Ce workflow peut ainsi être créé par un "expert" et réutilisé par la suite par des non experts.
- OpenRefine permet de compter le nombre de catégories par champ pour détecter les erreurs.
- En savoir plus sur la curation
    - https://www.reseau-canope.fr/savoirscdi/societe-de-linformation/reflexion/la-curation/la-notion-de-curation.html
    - http://www.dlib.org/dlib/january15/assante/01assante.html
    - Digital Curation Center (DCC) Curation Lifecycle Model - http://www.dcc.ac.uk/resources/curation-lifecycle-model
    - https://www.rd-alliance.org/groups/deconstructing-data-life-cycle-agile-curation.html

Les principales transformations sur les contenus

- https://fr.wikipedia.org/wiki/Curation_de_contenu

Le versionning

- Le Versioning des données / bases de données est essentiel pour s'assurer de la bonne préservation des données. Mais des interrogations demeurent : à quel niveau / granulométrie de modification doit-on créer une nouvelle version ? On identifie trois niveaux principaux : 1-à chaque modification/par période ; 2-chaque année ; 3-à chaque utilisation / téléchargement. (Des conseils concernant les choix à faire seront intégrés dans la prochaine version de ce document).

Quelques problèmes rencontrés dans les jeux de données lors des ateliers curation

- Les champs vides
- La polysémie
    - Définition polysémie / équivalence de termes,
    - Supprimer les accents peut conduire à des erreurs.
- Cas de l'encodage comme source d'erreur(s!) :
    - Les accents peuvent entraîner des erreurs. Faut-il les laisser et faire en sorte de prévenir les futurs utilisateurs/analyseurs de la donnée de la présence d'accent…. OU obliger l'utilisateur / mettre en place des routines informatiques pour les supprimer/gérer…. OU créer des dictionnaires indexés avec toutes les versions d'écritures possibles ? Les informaticiens peuvent utiliser l'UTF-8 pour avoir des données très simples. La solution proposée est la duplication des champs (un avec accent, un sans).
    - Les erreurs de frappe sont elles aussi en cause dans de nombreux cas : "O/0" ; "l/I" ; " / " ; chiffre arrondi ; retour chariot etc….
- Erreurs associées aux imports exports de données :
    - Séparateur comme les point-virgules dans le csv ; et dans le contenu d'un champ,

    - CSV : pas forcément idéal si des champs mémo/texte long sont présents dans le jeu de données, ils autorisent le retour chariot, le point-virgule et autres caractères utilisés par le csv,
    - Les lignes vides en fin de tableur sont à éviter.
- La syntaxe (partie à compléter lors d'un futur atelier)
- Validité des données : définir un niveau de fiabilité fait partie de la curation. Il est conseillé d'ajouter un champ validé/non validé pour chaque descripteur de données traité et de le renseigner pour chaque enregistrement (cela peut être réalisé par lot). Les données invalides sont parfois difficiles à détecter, et certaines ne peuvent être détectées que par les experts du système observé. Lors des processus de curation, décrire l'incertitude sur les données est primordial pour pouvoir vérifier la pertinence de l'analyse et/ou permettre la réutilisation pertinente des données dans un autre cas d'usage / par un autre utilisateur. De plus, plusieurs systèmes de validation peuvent être en "contradiction" les uns avec les autres. Par exemple, une donnée valide peut vouloir dire donnée plausible/impossible ou une donnée validée par un expert (exemple SINP-INPN). Ce champ validation doit donc être décrit de manière explicite et univoque dans le dictionnaire de données comme n'importe quel autre descripteur de données.
- Une autre erreur fréquente est l'arrondi de la machine à la n-ième décimale trouvable uniquement en visualisant le chiffre en "texte".

Les contraintes d'intégrité

- Les contraintes d'intégrité permettent d'interdire qu'une personne ne soit pas associée à une adresse mail.
- Elles doivent être écrites, et un processus de vérification de la donnée doit être implémenté.
- Elles permettent aussi de rendre cohérents deux champs (par exemple, la date de naissance et une catégorie d'âges)

Le plan de gestion des données (DMP pour Data Management Plan) décrit comment les données seront gérées tout au long de leur vie (quelles données seront collectées ou générées, comment elles seront gérées, partagées et préservées pendant et après le projet). Ces plans de gestion des données deviennent des critères d'éligibilité pour les projets nationaux, européens et internationaux.

Pour la communauté de la biodiversité, l'outil DMP OPIDoR https://dmp.opidor.fr/ peut aider les acteurs à écrire leur plan de gestion des données.
https://www.earthobservations.org/documents/geo_xii/GEO-XII_10_Data%20Management%20Principles%20Implementation%20Guidelines.pdf

---

Annexe 9. Résultats préliminaires des ateliers sur la curation de données

Visualisation sous forme de graphes de données en écologie et environnement : retour sur les ateliers 2017

Janvier 2018

Consortium IndexMEED

CONTRIBUTEURS

Merci aux experts qui ont répondu présent et qui ont encadré les ateliers : David Auber (LaBRI), Romain Bourqui (LaBRI), Alrick Dias (OSU Institut Pythéas), Baptiste Laporte (FRB), Julien Lecubin (OSU Institut Pythéas).

Merci aux participants aux deux ateliers « Curation de données », à l'« Atelier préparatoire à la visualisation » et à l'atelier « Visualisation de données » : Robert Arfi (IRD), Loup Bernard (UNISTRA), Guillaume Body (ONCFS), Pierpaolo Brena, Yvan Le Bras (MNHN), Grégoire Loïs (MNHN), Michelle Leydet (IMBE), Sophie Pamerlon (GBIF-France), Anne Quesnel-Barbet (Université Lille 2), Eloïse Trigodet (MNHN).

Animateur du consortium

Romain David (IMBE-CNRS).

Coordination de l'étude

Anna Cohen Nabeiro (FRB), Romain David (IMBE-CNRS) et Aurélie Delavaud (FRB).

Début de la mission : Janvier 2017

Fin de la mission : Novembre 2017

Financement

Ce travail a été financé par la Fondation pour la Recherche sur la Biodiversité, le GDR MaDICS, l'Université Aix-Marseille. Ont participé : merci aux CESAB, ECOSCOPE, FRB, GBIF, IMBE, LAM, Labex DRIIHM (OHM Bassin Minier de Provence, OHM Vallée du Rhône, OHM Littoral méditerranéen), Fédération ECCOREV FR 3098, OSU Pythéas, LabEx OT Med et GDR MaDICS pour leur soutien humain et financier à l'organisation des ateliers et manifestations de 2016 et 2017.

Remerciements

Merci à Denis Parade (Scénario Interactif) pour son appui technique à l'utilisation du logiciel Gephi, merci à Robin Goffaux (FRB) et Samir Hamdi-Cherif (FRB) pour leur aide à la curation de données sur Excel, merci à Romain David (IMBE) pour son aide à la réalisation des étapes de curation. Remerciements aux intervenants S.T.I.C. (D. Auber, R. Bourqui, A. Dias, J. Lecubin)

Citation

Romain David (IMBE-CNRS) Anna Cohen Nabeiro (FRB), Aurélie Delavaud (FRB), Loup Bernard (UNISTRA), Guillaume Body (ONCFS), Yvan Le Bras (MNHN), Michelle Leydet (IMBE), Consortium IndexMEED, 2018. Visualisation sous forme de graphes de données en écologie et environnement : retour sur les ateliers 2017, mission 2017.

1. Présentation de l'étude

Dans le domaine de l'écologie et de la biodiversité, l'augmentation des fréquences et volumes d'acquisition de données (observations de terrain, capteurs optiques, capteurs radar, télédétection, systèmes de suivi de la qualité des eaux, recensement automatique ou semi-automatique des taxons, séquençage, génotypage, etc..) a abouti à une accumulation considérable de données hétérogènes et dispersées qu'il est nécessaire d'organiser, documenter et trier pour les exploiter à des fins de recherche et d'appui opérationnel à l'expertise.

Ateliers 2016 à l'Université Aix-Marseille et au CESAB

Le consortium Indexmed, rebaptisé IndexMEED pour « Indexing for Mining Ecological and Environmental Data », développe des processus d'indexation et de qualification de données, hétérogènes et distantes. Ces processus permettent de construire des graphes à partir de données concernant la biodiversité et de les exploiter avec des objectifs d'indication, d'aide à la décision et de formulation de nouvelles hypothèses scientifiques. Les graphes permettent de représenter et d'explorer des interactions entre des objets d'observation ou d'expérimentation, des variables mesurées, des paramètres pris en compte.

L'année 2016, consacrée à l'indexation des données à des fins d'harmonisation des systèmes d'information en respectant les objectifs et les contraintes métiers des acteurs, a permis de mettre en évidence le potentiel des approches basées sur les graphes - déjà éprouvées dans d'autres domaines du « big data » - pour la fouille de données ainsi que les lacunes en terme de compétences et d'expérience de la communauté de recherche en écologie pour adapter et utiliser ces approches.

Ainsi, le consortium IndexMEED a réalisé, au cours de l'année 2017, une série d'ateliers thématiques pour développer les approches de graphes auprès des communautés de l'écologie et de l'environnement. Les premiers ateliers, réalisés en parallèle à Paris, à Aix-en-Provence et à Marseille, se sont focalisés sur les étapes de curation des jeux de données, c'est-à-dire leur modification pour les rendre pertinents à l'import dans des logiciels de visualisation et de traitement de graphes. Les seconds ateliers se sont déroulés à Paris, et ont permis de visualiser les données des participants sous forme de graphes grâce à l'intervention de spécialistes.

Ce document présente les premiers résultats issus de ces ateliers.

2. Résultats préliminaires

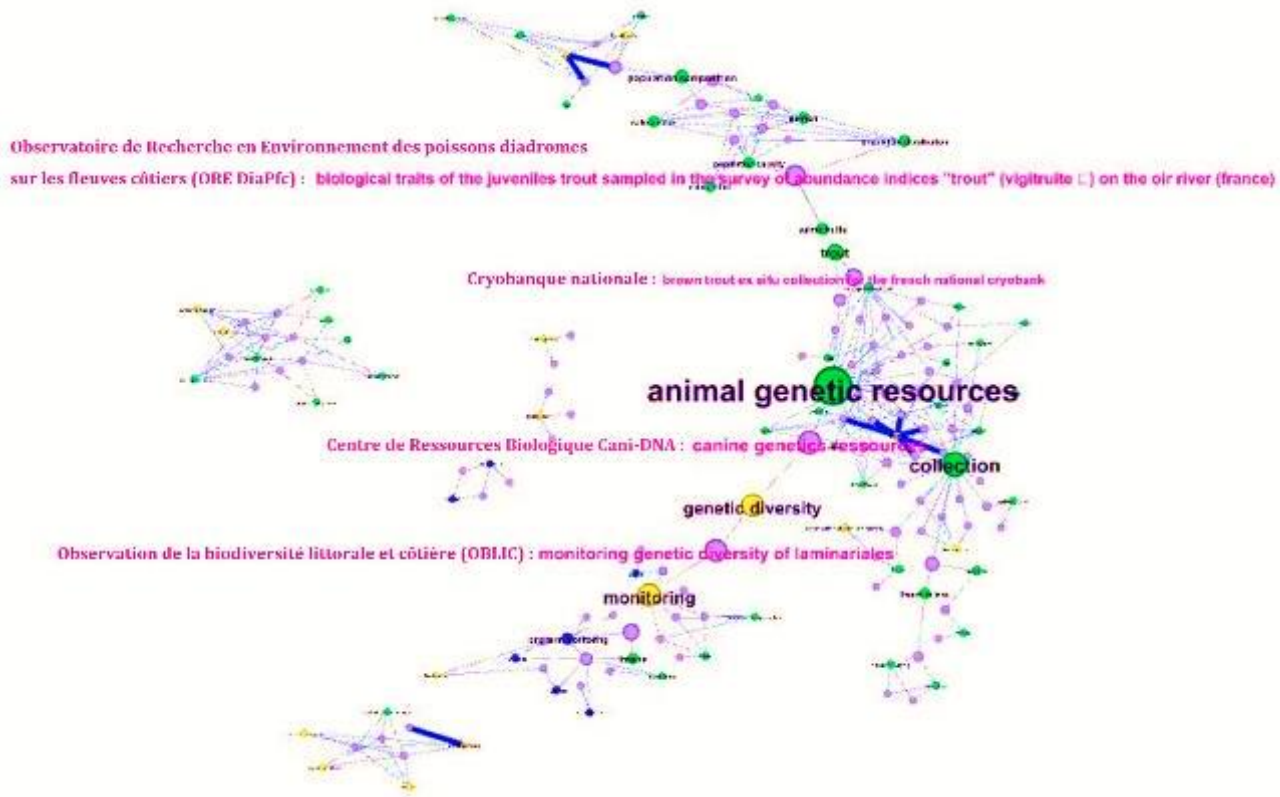
LE CATALOGUE DE MÉTADONNÉES DU PORTAIL ECOSCOPE

Description du cas d'étude

Le jeu de données utilisé est une base de métadonnées décrivant des jeux de données d'observation de la biodiversité à différents niveaux d'organisation du vivant, ainsi que les dispositifs à partir desquels les données ont été produites. Les métadonnées permettent de répondre aux questions Qui, Quoi, Où, Quand, Comment, Pourquoi : titre du dispositif et du jeu de données, un résumé, un descriptif des variables suivies, pour quelles couvertures temporelles et spatiales, quels taxons, quels matériel et méthodes, quels contacts, quelles conditions d'accès et d'utilisation des données, une description de la collection si existante, l'URL vers la base de données… Cette base existe depuis 2015, et s'appuie sur le logiciel PostGre pour le stockage des métadonnées, sur une application de saisie et d'export au format XML, et sur PostGIS pour le stockage des informations géoréférencées. Le profil de métadonnées suit le standard EML, enrichi de standards utilisés au niveau national. En tout, ce sont 50 lignes d'enregistrements pour la table dispositif et 150 lignes pour la table jeu de données. Une cinquantaine d'utilisateurs a permis l'alimentation de cette base. Les métadonnées sont en libre accès à l'adresse www.ecoscope.fondationbiodiversite.fr/ecoscope-portal/.

Problématique 1

Quels sont les liens que l'on peut observer entre jeux de données en fonction des mots-clefs les décrivant ? Quels sont les jeux de données et les mots-clefs qui permettent de connecter les différentes parties du graphe ?

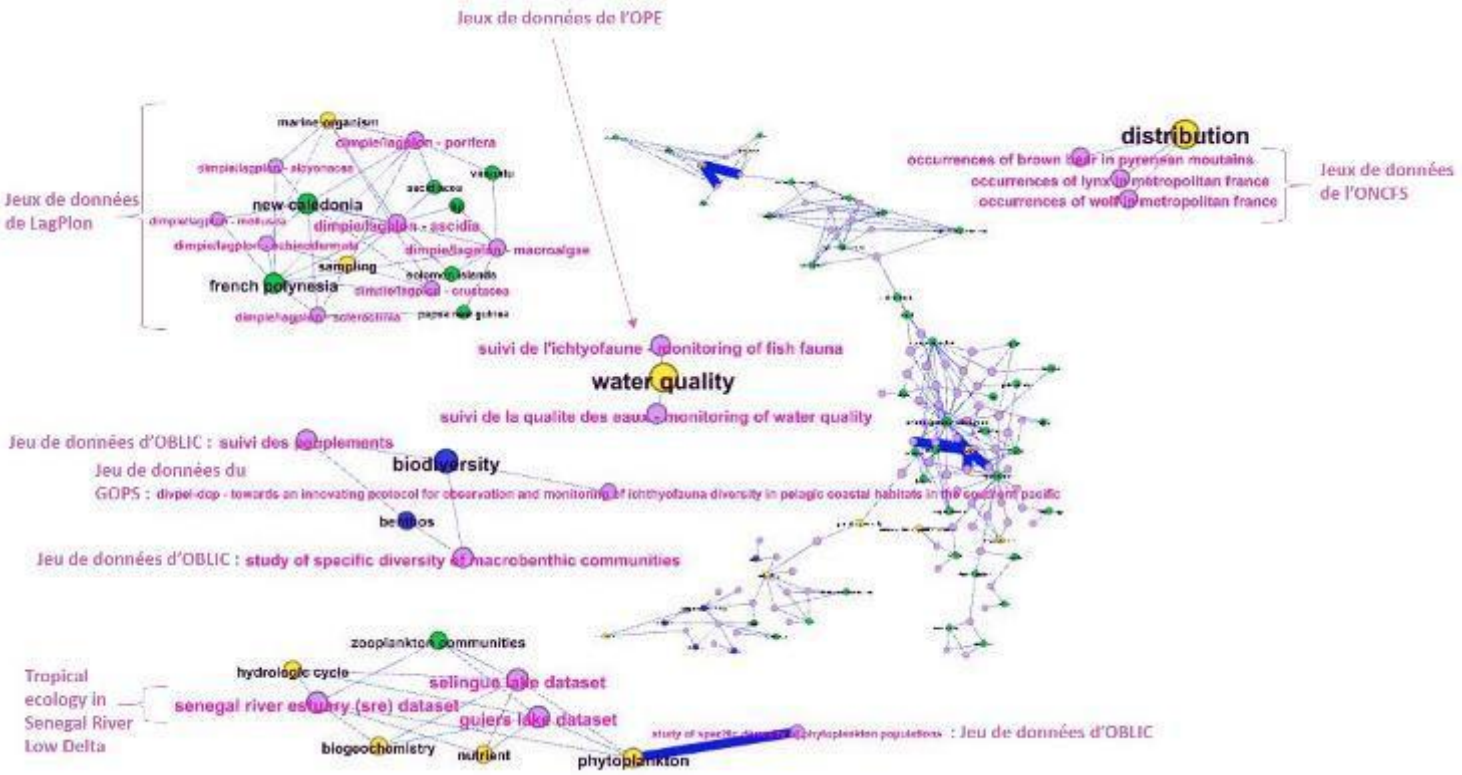


Les jeux de données (nœuds roses) sont reliés par les mots-clefs les décrivant dans le portail. Les mots-clefs peuvent être issus de trois thésaurii différents : nœuds verts pour AGROVOC, nœuds jaunes pour Gemet et nœuds bleus pour EnvThes. Tous les labels des mots-clefs sont visibles, alors que les labels des jeux de données ont été pour la plupart dissimulés afin de faciliter la lecture du graphe. Une analyse statistique a été réalisée, afin de faire apparaître en plus gros les nœuds (et leurs labels) par lesquels il faut passer le plus fréquemment pour rejoindre les autres nœuds. C'est la « betweeness centrality », ou centralité d'intermédiarité. Les labels des jeux de données ayant la plus grande betweeness centrality ont été rendus visibles.

Les mots-clefs et les jeux de données ayant la plus grande betweeness centrality sont en quelque sorte à l'interface entre différentes parties du graphe. C'est le cas d'un jeu de données de l'ORE DiaPFC, d'un jeu de données de la Cryobanque nationale, du mot-clef Animal genetic resources, d'un jeu de données de Cani-DNA, du mot-clef Collection, du mot-clef Genetic diversity, d'un jeu de données d'OBLIC et du mot-clef Monitoring.

Problématique 2

Les données utilisées pour ce second graphe sont les mêmes que pour le premier. C'est l'analyse statistique qui diffère, car il s'agit de mettre en avant les nœuds « centraux », c'est-à-dire ayant des distances équivalentes avec les autres nœuds.



Via l'analyse de la « closeness centrality », les nœuds centraux et leurs labels ont une plus grande taille que les autres nœuds. Il ne s'agit pas des nœuds ayant le plus de liens, mais de ceux ayant une distance équivalente avec les autres nœuds auxquels ils sont reliés.
Les noeuds qui ressortent de cette analyse sont les noeuds centraux de certaines parties du graphe déconnectées entre elles, qui forment des sous-graphes presque complets. Ils apparaissent comme centraux car ils sont connectés avec presque tous les autres nœuds du sous-graphe.

Perspectives

La base de métadonnées du portail ECOSCOPE pourrait être mise à profit pour découvrir de façon approfondie le paysage des observatoires de recherche sur la biodiversité. Plusieurs sujets de graphes peuvent être envisagés :
- la visualisation des liens entre les dispositifs en fonction des Essential Biodiversity Variables (EBV) et des objectifs renseignés pour voir si se dessine un « patron » scientifique cohérent ;
- la visualisation des parentés (sémantique, taxonomique…) entre jeux de données ;
- la visualisation d'éventuels clusters en fonction des pratiques d'Accès et Partage des Avantages (APA) issues du protocole de Nagoya, et mettre cela en lien avec les pratiques de gestion des ressources génétiques ;
- la visualisation d'éventuels clusters en fonction de catégories de pressions ou de services écosystémiques étudiés, qui varient selon l'échelle territoriale et les habitats ;

- la visualisation des réseaux et du financement des organisations au sein de ces réseaux.

LA BASE DE DONNÉES EUROPEAN POLLEN DATABASE

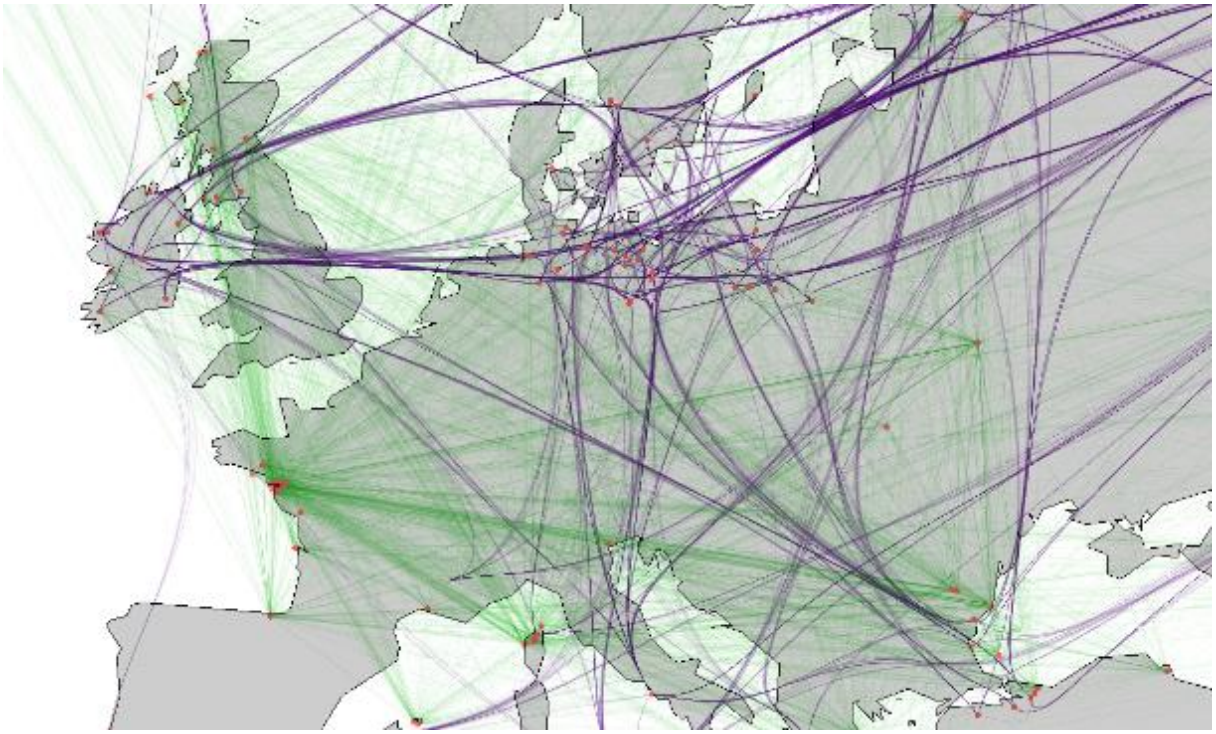
Description du cas d'étude

L'European Pollen Database (EPD) est une base de données relationnelle contenant des données et des métadonnées de pollens fossiles, essentiellement provenant d'archives naturelles (sédiments lacustres, tourbières, sédiments marins) collectées sur le continent eurasien ou à proximité. Elle a été fondée en 1989 par une équipe internationale de paléo-environnementalistes pour effectuer des études climatiques prédictives et des cartes de la végétation à l'échelle du continent. Le but de l'EPD est de développer une plate-forme ouverte pour favoriser l'étude scientifique des enregistrements palaeo-écologiques à long terme afin d'aborder divers thèmes tels que la biogéographie, l'histoire de la végétation, la conservation des écosystèmes. Plusieurs milliers de données sont ainsi enregistrées dans les tables, renseignant ainsi les sites, les pollens trouvés dans chaque site et leur datation quand elle existe, ainsi que des éléments de contexte (type de site, localisation précise, altitude).
Cette structure à but non lucratif est gérée dans l'Institut Méditerranéen de Biodiversité Marine et Terrestre et d'Ecologie par Michelle Leydet (Aix-en-Provence, France). Elle est hébergée actuellement au Cerege (Aix-en-Provence, France) et le serveur est financé par Eccorev (Joel Guiot, Aix-en-Provence, France). L'EPD a été soutenue financièrement par l'Université d'Aix-Marseille depuis 2007.

Les données polliniques sont issues directement des palynologues européens ou de contributions de personnes tiers. Dans le cadre collaboratif d'échanges inter-bases de données internationales, elles proviennent de Pangaea (base de données multy-proxies, Bremerhaven, Germany), Neotoma (base de données multi-proxie, PennState University, Pennsylvania, USA), Palycz (Czech Quaternary Palynologycal database, Charles University, Prague, Czcechoslovakia), Alpadaba (Institute of Plant Sciences, University of Bern, Bern, Switzerland). Par ailleurs, les projets internationaux utilisant l'EPD soumettent de nouvelles données et apportent des corrections sur les données existantes. L'EPD comporte plusieurs groupes de travail constitués d'experts volontaires qui améliorent la qualité des données depuis 10 ans.

Problématique

Il s'agit de visualiser sur une carte les carottes issues de sondages dans différentes zones, du nord de l'Afrique à la Sibérie.



Les sommets en rouge représentent les sondages polliniques géo localisés en Europe extraits de l'European Pollen Database. Les arêtes représentent les liens entres les sondages carottés à la même altitude et dans le même type d'environnement.
Les sondages sont issus de différents types de site (lac, tourbière, marais…) renseignant ainsi sur le contexte de dépôts sédimentaires et polliniques. Les sondages polliniques carottés situés dans un lac de nature non spécifié (Lacustrine) entre 0 et 100 m d'altitude sont liés par des arêtes violettes. Les sondages polliniques carottés dans un environnement non décrit (Unknown) situés entre 0 et 100 m d'altitude sont liés par des arêtes vertes. Ces sondages devront être renseignés à partir des publications associées à ces données ou être référés à des cartes physiques et/ou géologiques.

Perspectives

La description des sites n'est pas toujours correctement documentée voire manquante, et nécessite un champ d'investigation de recherche bibliographique plus poussé. Le graphe permet de mettre en évidence les enregistrements non décrits. Cet outil va permettre d'améliorer la qualité des jeux de données en sélectionnant des zones géographiques prioritaires insuffisamment renseignées et répondant cependant aux besoins d'étude dans le cadre d'un projet de recherche. En effet, il en ressort que 13.5% des sondages doivent être vérifiés pour compléter la description des sites polliniques. 74% de ces sites non renseignés sont publiés et peuvent être renseignés. Les 26% restant nécessitent un appui cartographique supplémentaire. Dans le cadre collaboratif inter-bases de données entre Neotoma et l'EPD, la correction de ce champ rendue plus aisé grâce à cet outil, est en cours de correction depuis décembre 2017 et se poursuit.
En outre, cet outil permettra aux palynologues d'évaluer l'évolution d'une espèce végétale au cours du temps et dans l'espace en sélectionnant un pas de temps, un champ d'altitude et/ou une aire géographique. Il sera possible dès lors de combiner des champs comme le contexte de dépôt, le type de végétation environnante, la proximité d'un site archéologique.

LA BASE DE DONNÉES VIGIE-NATURE OBSERVATOIRE DES PAPILLONS DE JARDIN

Description du cas d'étude

Dans le cadre du programme Vigie-Nature porté par le Muséum national d'Histoire naturelle, du programme « Observatoire de la Biodiversité des Jardins » de l'association Noé, et en partenariat avec la Fondation Nicolas Hulot pour la Nature et l'Homme, l'***Observatoire des Papillons des Jardins"*** a été créé. Cet observatoire est ensuite renommé "**Opération Papillons**"

http://vigienature.mnhn.fr/page/operation-papillons.

Base de données centralisée, cet observatoire grand public permet de rassembler, puis d'analyser les observations collectées dans les jardins. Cette action a reçu le soutien de la Fondation d'entreprise Veolia Environnement et de l'entreprise Gamm vert.

À terme, c'est un véritable réseau de surveillance des espèces communes de papillons de jour qui sera mis en place, permettant de suivre l'évolution des populations et de mieux comprendre les dynamiques écologiques, en lien avec les changements climatiques par exemple. L'Opération Papillons, première expérience d'observatoire grand public de la biodiversité en France, est un outil capital pour construire, dans les prochaines années, des actions adaptées à la protection des papillons et de la biodiversité en général.

Un sous-ensemble de données datant de 2016 a été utilisé, et les identifiants des jardins, des espèces observées et des coordonnées GPS associées aux jardins ont été récupérés.

Deux représentations différentes des données ont été utilisées :

- une représentation « centrée espèces » (nœuds = espèces observées ; liens = jardins communs à des observations d'espèces)

- une représentation « centrée jardins » (nœuds = jardins ; liens = espèces observées au sein de jardins différents)

Problématique 1 : Représentation centrée espèce

L'algorithme *GEM (Frick)* est tout d'abord appliqué sur l'ensemble du graphe pour permettre une représentation plus lisible des données (Figure 1). Elle permet déjà grossièrement de 1/voir que les données forment un ensemble assez « compact », lié, avec un « noyau » de nœuds très liés, et une répartition des autres nœuds liés au sein de deux « ceintures » distinctes et 2/d'identifier la présence de 19 nœuds chacun « déconnectés » de tout autre nœud.

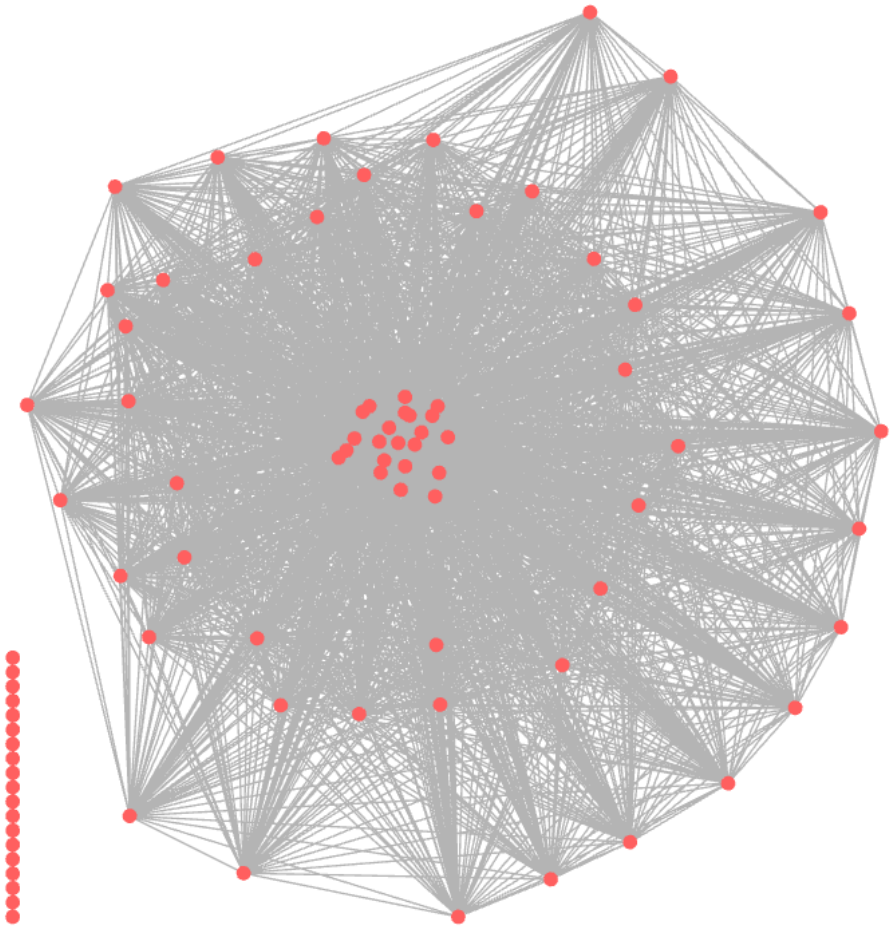


Figure 1 : Représentation graphique du sous-jeu de données Observatoire des papillons de jardins après application du layout GEM (Frick)

La consultation de la liste des 19 espèces « isolées » (Tableau 1 de l'annexe) permet de mettre en évidence une valeur aberrante (Carte Géographique) présente dans la donnée brute.

GROUPNAME
Sylvain azuré
Robert-le-Diable
Petits Nacrés
Petites Tortues
Paon du Jour
Mélitées
Marbrés
Hespéries orangées
Grande Tortue
Fluorés - Colias jaunes
Fadets
Demi-Deuils
Carte Géographique
Brun des Pélargoniums
Belle-Dame
Azuré des nerpruns
Azuré Porte-Queue
Autres papillons
Mille pattes

Tableau 1 de l'annexe *: Liste des espèces représentées par les 19 noeuds qualifiés d'"isolés"*

Le graphe obtenu ne permet pas de mettre en évidence de patterns, c'est-à-dire de tendances, dominants ; toutes les espèces semblent être présentes en association dans au moins un jardin. En utilisant l'outil « highlight node neighborhood », on observe effectivement le même type de pattern avec un lien entre toutes les espèces (Figure 2).

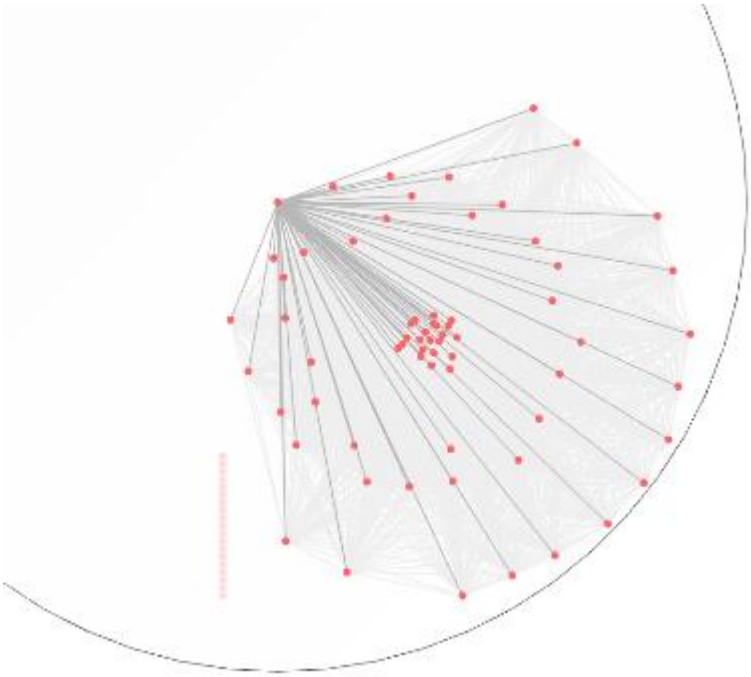


Figure 2 : Représentation graphique des données via l'utilisation de l'outil "highlight node neighborhood"

Problématique 2 : Représentation centrée jardin

Une approche sans utilisation des coordonnées GPS est d'abord choisie, et le « Layout GEM (Frick) » est appliqué, ce qui permet de représenter les données de manière adéquate et avec une notion de "distance" entre les nœuds (Figure 3).

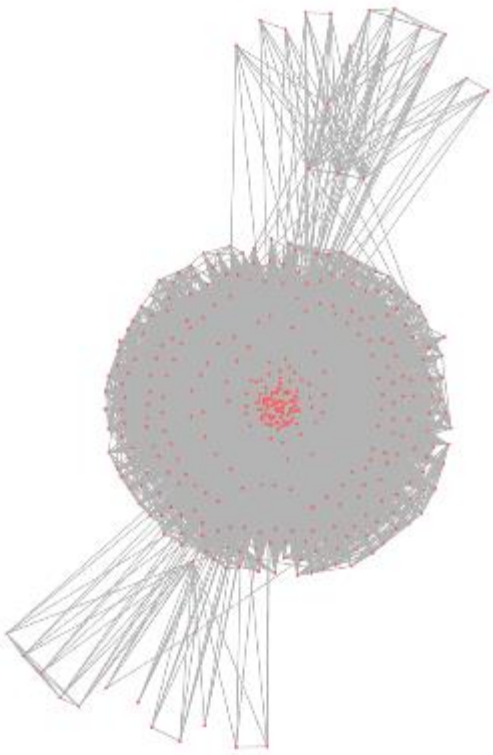


Figure 3 : Représentation des données "centrées jardin" après application du layout GEM (Frick)

Un algorithme de classification hiérarchique est ensuite appliqué, ce qui génère deux groupes de “jardins”.

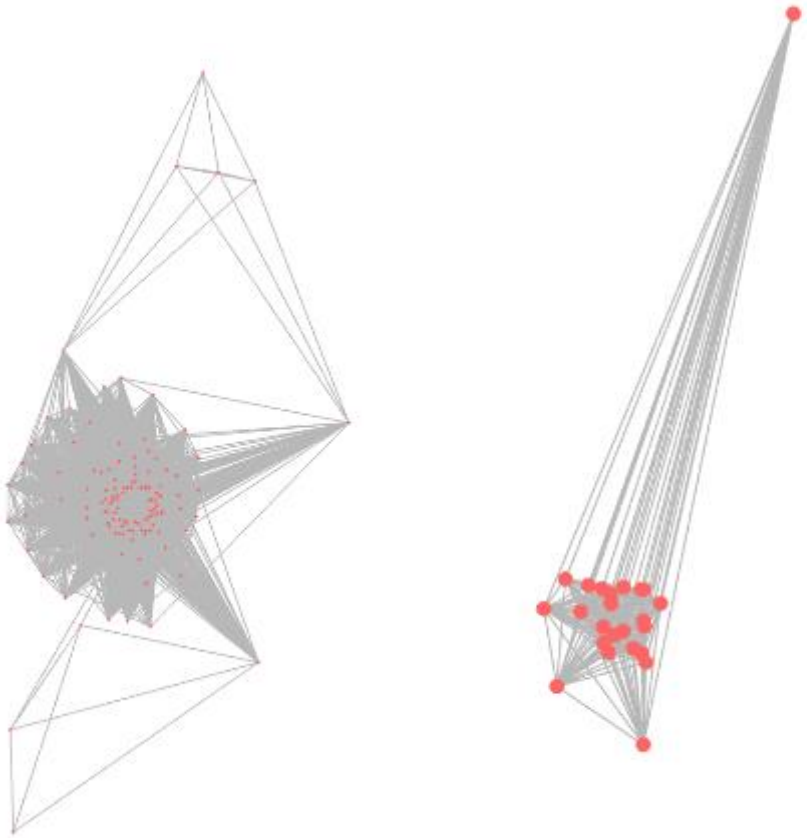
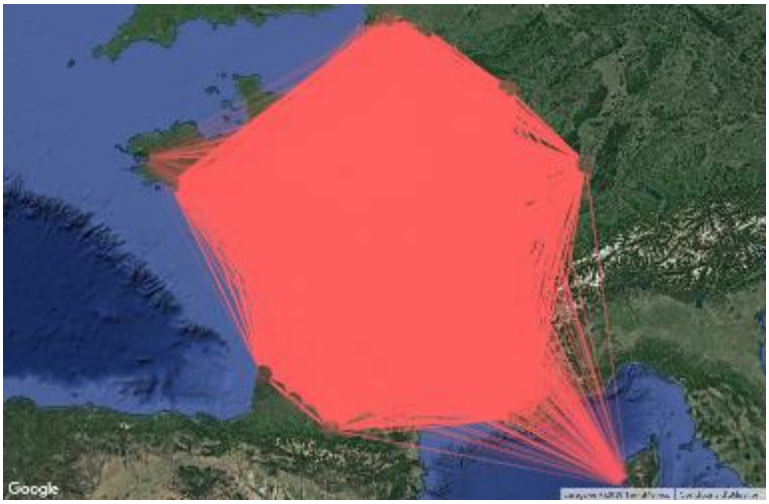


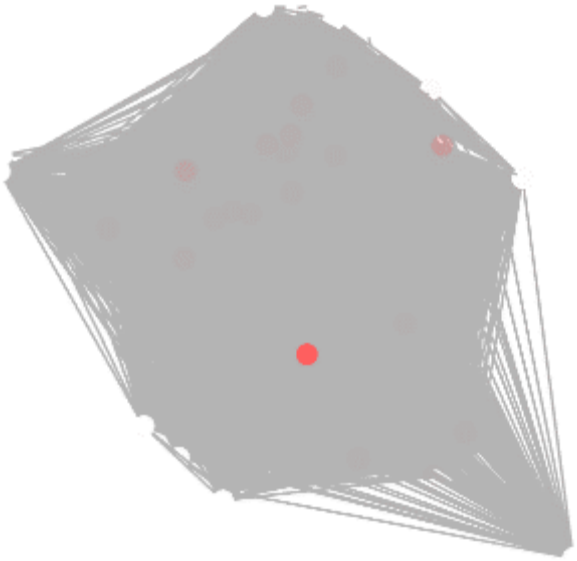
Illustration 4 : Représentation des deux groupes d'entités obtenus après application d'un algorithme de classification hiérarchique.

Les informations des deux groupes de données ont été exportées et des analyses complémentaires doivent être menées pour tenter de révéler ce qui différencie les deux groupes de jardins.

Problématique 3 : Géo-Spatialiser les données

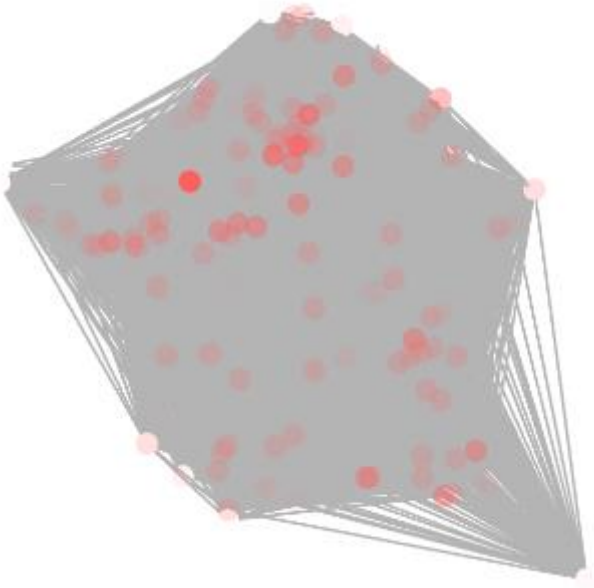
Une seconde approche avec utilisation des coordonnées GPS est ensuite choisie. Le “Layout GEM (Frick)” est appliqué pour représenter les données avec une notion de distance entre les nœuds. Un panneau de visualisation “Geographic view” est utilisé, et l’algorithme Betweenness centrality et la représentation par dégradé de couleur via la méthode d’alpha mapping sont appliqués.





Les points qui ressortent correspondent aux villes d'Aurillac, Metz et Le Mans et représentent donc des villes par lesquelles passent les maximums de “chemins”. Aurillac semblerait donc être une ville “charnière” du plus grand nombre d’espèces identifiées parmi le plus grand nombre de villes.

L’algorithme Degree, qui permet de calculer les nœuds ayant le plus de liens, et la méthode d’alpha mapping pour représenter les nœuds par dégradé de couleur sont ensuite appliqués.



Ici ressortent les villes présentant le plus de liens, donc le plus grand nombre d’observations d’espèces dans de nombreuses autres villes.

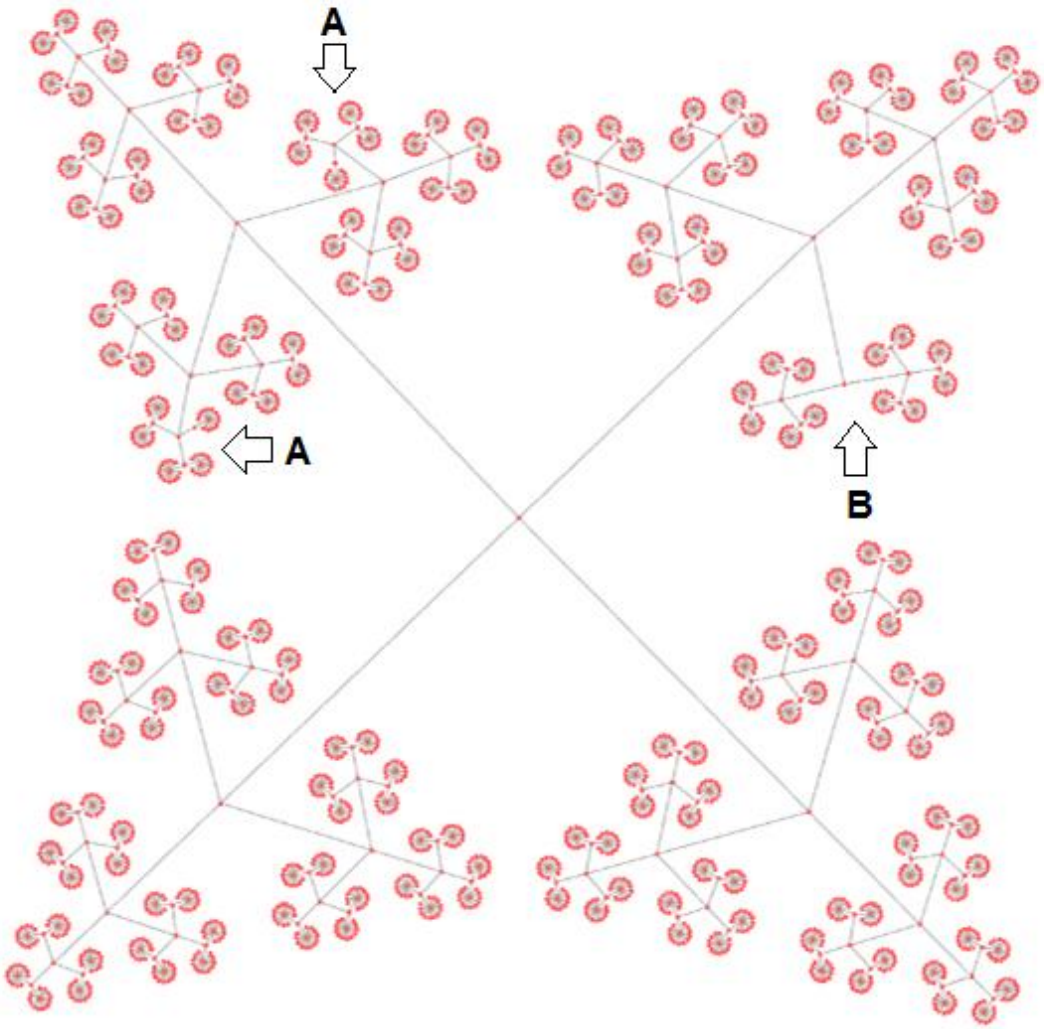
LA BASE DE DONNÉES « ARMS » (Artificial Reef Monitoring Système) Du programme Européen DEVOTES

Description du cas d'étude

De 2013 à 2017, le projet européen DEVOTES (Développement D'Outils innovants pour la compréhension de la biodiversité marine et l'évaluation du bon état environnemental) a testé le potentiel des ARMS (Artificial Reef Monitoring System) initialement créés pour des suivis dans des récifs coralliens dans les mers européennes. Ces dispositifs permettent de normaliser le suivi de communautés benthiques dans les substrats durs, et de mettre en évidence les liens entre les structures des communautés installées sur les plaques de ces petits récifs artificiels et les contextes naturels et anthropiques dans lesquels ils ont été installés. Ce programme de recherche vient en appui de la DCSMM Directive Cadre Stratégie pour le Milieu Marin, et notamment du descripteur 1 qui concerne la diversité biologique et du descripteur 4 concernant les réseaux trophiques.
Ce jeux de données contient les premiers résultats concernant les structures de communautés macroscopiques ayant colonisé des plaques contenues dans ces récifs après plus d'un an d'immersion dans différentes conditions environnementales dans trois mers d'Europe (Mer atlantique, près de la baie de Biscaye, Mer Méditerranée du nord-ouest et Mer Adriatique), ainsi que dans la Mer Rouge. Les compositions communautaires, déduites de photographies, sont décrites par régions marines, par sites et par surfaces distinctes des dispositifs ARMS. Divers facteurs et variables environnementaux, reflétant à la fois le niveau d'anthropisation et la diversité de l'habitat local, décrivent chaque site.
Les données sont accessibles en ligne dans un entrepôt de données en accès libre, dès la parution de la publication scientifique. En attendant cette mise à disposition plus soignée accompagnée d'un "*datapapers"*, elles sont téléchargeables sur le site du projet devotes http://www.devotes-project.eu/devotool/.

Problématique 1

On souhaite représenter le plan d'échantillonnage des plaques sous forme de graphe, afin de mettre en évidence des éléments manquants ou des erreurs et oublis (des sites avec moins de plaques que d'autres par exemple). Pour cela, on représente les plaques en les reliant au récif, puis en reliant le récif au site, puis le site à la mer de manière à former un arbre (qui est une sorte de graphe). Les nœuds terminaux, en rouge, représentent les faces de plaques. Un simple algorithme de recherche de dissymétrie permet de mettre en évidence soit les pertes d'un dispositif, soit les disparités d'analyse (perte d'une photo de plaques, ou oubli de détermination d'un morceau de plaque). Dans le graphe ci-dessous, les 4 branches de l'arbre représentent les 4 mers. Il manque 2 faces de plaques dans une mer (A : perte de 2 photos par l'opérateur), et une ARMS dans une autre mer (B : arrachée du substrat lors d'une tempête)



Problématique 2 Visualiser avec le logiciel Tulip les plaques et les regroupement de plaques reliées à partir des fréquences relatives de taxons recueillis sur des éléments de récifs artificiels en PVC conçus initialement par la NOAA et répartis à différents endroits du récif (et donc exposés différemment aux facteurs de contextes comme l'exposition à la lumière, au courant, ou aux différentes pressions décrites soit par des capteurs, soit à dire d'expert…)

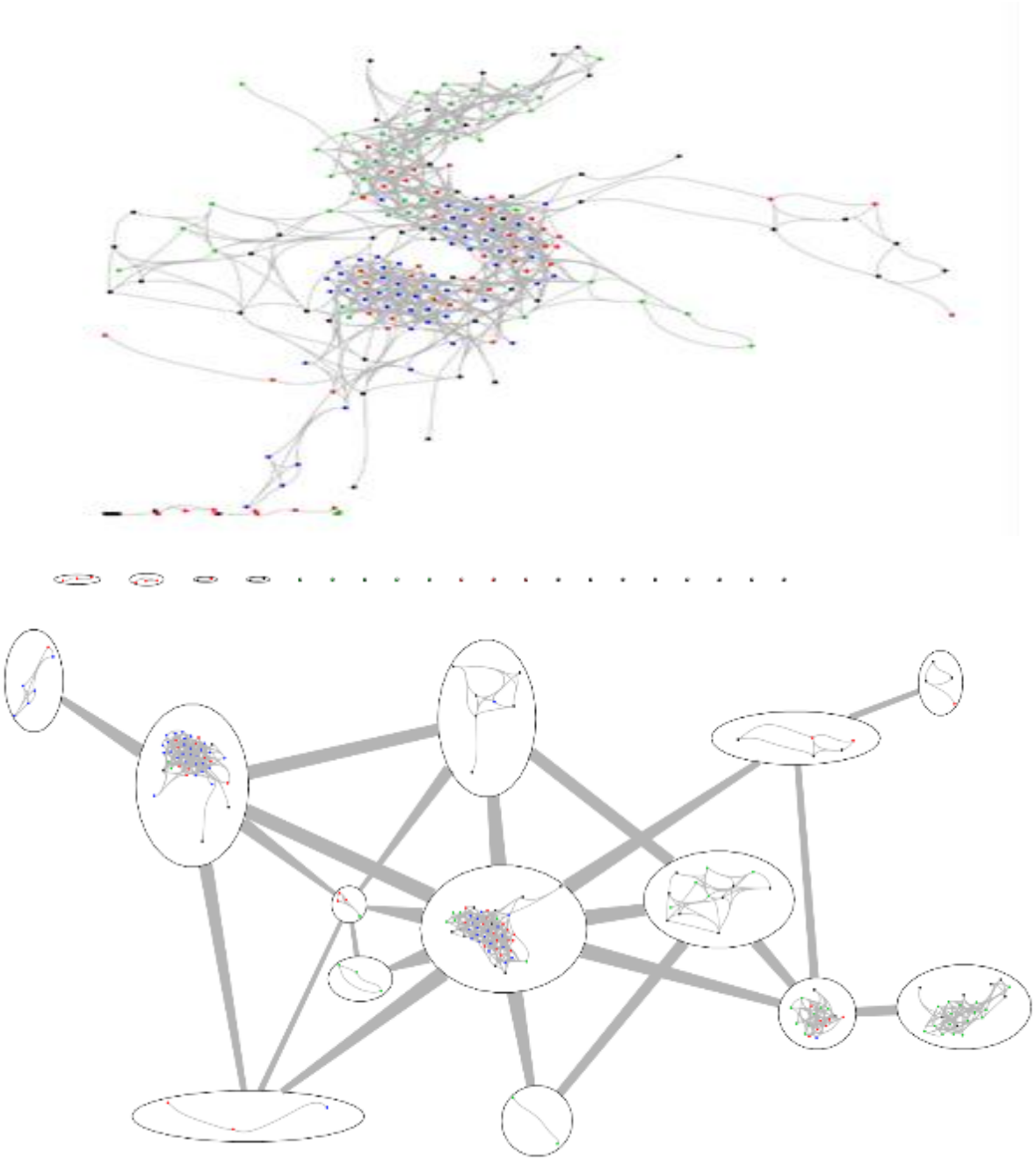


Fig 2 : Dans cette visualisation préliminaire, les fréquences relatives ont été déterminées à partir de photos. 6 facettes de plaquettes ont été analysées dans 3 récifs, pour 3 sites pour chacune des 4 mers régionales (Golfe de Gascogne, méditerranée nord occidentale, Adriatique et Mer Rouge). Les nœuds de ce graphe représentent des facettes de plaques en PVC, et les liens relient les facettes pour chaque pourcentage d'un taxon en commun avec une autre facette. Chaque couleur représente une mer régionale différente. Dans la première figure, on visualise les premiers clusters, où les plaques avec les compositions en taxons les plus similaires sont les plus proches, et les sites des différentes mers sont colorés de manières différentes avec des regroupements plus ou moins hétérogènes (Fig 2 en haut). La deuxième figure a été construite en sélectionnant uniquement des fréquences relatives en commun supérieures à 10%. Elle permet de mettre en exergue les regroupements en fonction de combinaisons de valeurs de paramètres environnementaux (en faisant varier la forme, la taille et/ou la couleur des nœuds en fonction des classes de valeurs des différents paramètres) avec un autre algorithme de

visualisation mettant en évidence les "clusters" (Fig 2 en bas). Lorsque les formes, tailles et/ou couleurs des nœuds se regroupent dans un cluster, on visualise une relation entre certains regroupements et des groupes de valeurs de variables de contextes. Il faut alors tester la significativité de ces corrélations grâce à des méthodes statistiques propres aux graphes.

Les prochaines étapes : Les combinaisons de facteurs de contextes peuvent être associées à chaque site (soit des combinaisons de descriptions de pression sous forme de booléen comme présence absence, soit plus finement avec des catégories ordonnées ou non ordonnées). Lorsque des combinaisons de facteurs de contextes se regroupent dans un cluster (une même gamme de température avec une même gamme de courant, lumière, et/ou facteurs de pressions), on peut considérer que la structure de communauté de ce cluster correspond à la combinaison en question. Les premiers tests montrent qu'un travail sur la sémantique concernant les descriptions de contexte est incontournable, surtout à large échelle, pour améliorer le potentiel de l'analyse et les qualités du suivi comme sa comparabilité dans le temps. Les prochaines étapes auraient pour objectif de les normaliser, puis de temporaliser les graphes pour analyser les effets saisonniers d'une part, et l'effet de différentes perturbations d'autre part. L'objectif est de mettre en évidence l'effet des changements de contextes sur les suivis biologiques sous forme de « graphes temporels » (et utiliser les successions biologiques comme des changements dans des fréquences relatives d'espèces comme indicateurs de changement d'état (avec une cause ou une conjonction de causes naturelles, et/ou anthropiques). Lors du recueil des données, des analyses de communautés faites via des analyses métagénomiques ont été effectuées. Ces représentations pourraient inclure les analyses métagénomiques avec le même type d'approche (des clusters de gènes en tant que liens entre plaques). Il est aussi possible de temporaliser le plan d'échantillonnage surtout lorsque celui-ci devient un dispositif permanent avec des objectifs de gestion (On visualise alors que l'arbre perd alors des éléments au fur et à mesure de la survenue des anomalies, tout en en affichant la cause si celle-ci est connue).

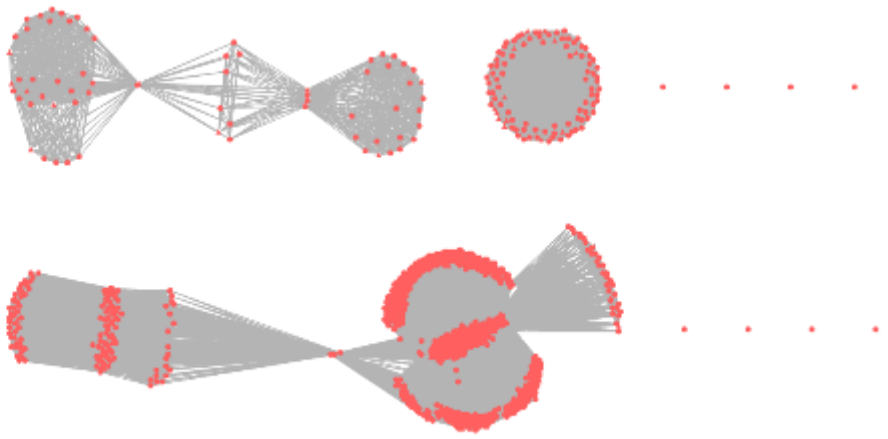
UNE OUVERTURE VERS D'AUTRES DISCIPLINES EN LIEN AVEC L'ÉCOLOGIE AVEC LA BASE DE DONNÉES ARKEOGIS

Description du cas d'étude

ArkeoGIS est une application pluridisciplinaire. Cette plateforme en ligne permet la mise en commun et l'interrogation d'une interface cartographique de données scientifiques spatialisées concernant le passé (archéologie, environnement…). Les bases de données sont issues de travaux de chercheurs institutionnels (effectués aussi bien dans le cadre de leurs recherches personnelles que dans le cadre de contrats), d'étudiants avancés, de sociétés privées, de services d'archéologie mais aussi issues de travaux de paléo-environnementalistes, d'historiens ou de géographes. L'interdisciplinarité est encouragée, tous ces travaux / bases de données étant partagés (protocole d'import), accessibles et requêtables en ligne (webSIG) par les utilisateurs d'ArkeoGIS. Chaque utilisateur dispose d'un espace projet personnalisable. Il peut interroger en ligne tout ou partie des bases disponibles, afficher ses résultats sur plusieurs fonds de carte, archiver et exporter les résultats de sa requête vers d'autres outils (export CSV).

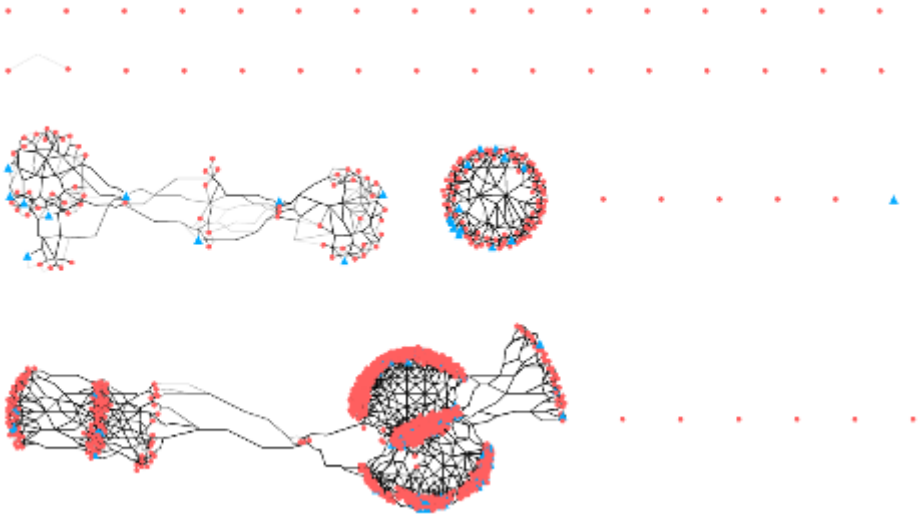
Problématique 1

La pratique de l'inhumation est-elle variable dans le temps ? Afin de répondre à cette question, un premier graphe est réalisé via le logiciel Tulip. Les tombes représentent les nœuds du graphe. Deux nœuds sont liés si les dates de fin d'exploitation du site correspondent, ces fourchettes de dates de fin (champ ENDING_PERIOD de l'export initial) ayant été coupées en deux dates numériques via la console Python. Les données prises en compte sont des données d'archéologie funéraire en vallée du Rhin supérieur en provenance de plusieurs bases de données, extraites au format CSV/UTF8/;/"" via ArkeoGIS. Le fichier a été intégré via le bouton CSV dans TULIP, qui l'a reconnu.



Le graphe permet de faire apparaître par période et par base de données les tombes à inhumation, sachant qu'à chaque période les deux pratiques (inhumation et incinération) coexistent, mais que des tendances lourdes sont identifiées. Le graphe fait ainsi apparaître plusieurs groupes de bases et de liens, dont l'étendue est plus large que celle connue dans leur définition initiale, que celle-ci soit chronologique ou spatiale.
Les sites de chronologie indéterminée, nombreux, ne sont reliés à rien et ne sont donc pas intéressants pour cette analyse. En revanche, ils permettent d'affiner le travail de curation et de mieux saisir comment les données doivent être organisées afin d'obtenir un graphe propre.

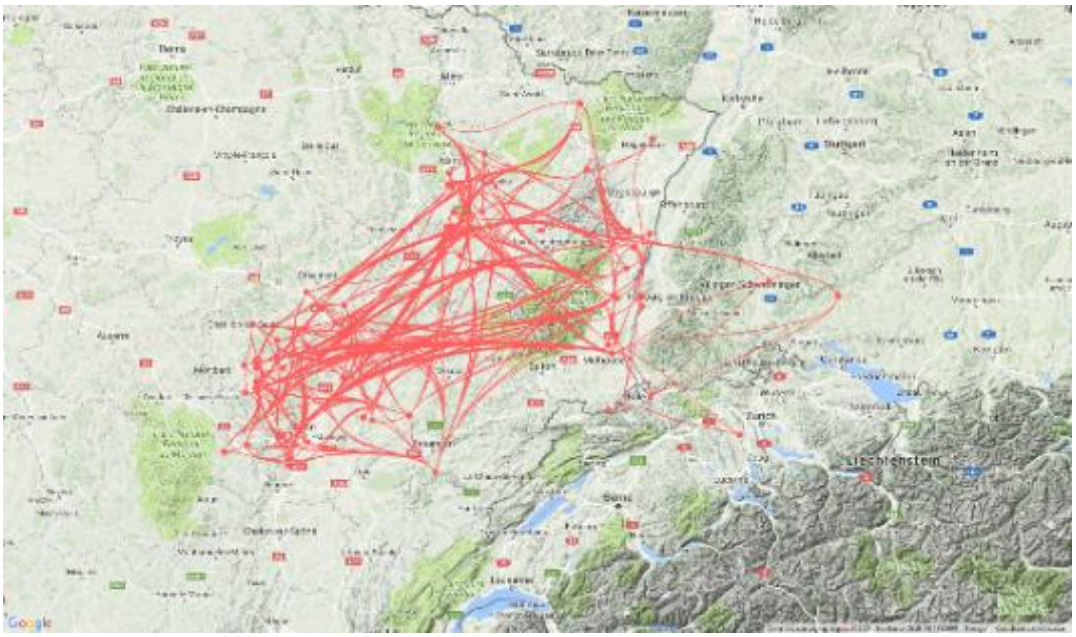
La lisibilité du graphe a été améliorée via la fonction edgebundle (lourd pour la carte graphique d'un portable i5 / 2,4Ghz). L'inhumation est maintenant représentée par des triangles bleus.



Ces deux représentations graphiques permettent de distinguer une répartition des sépultures par période (de gauche à droite). Les liens entre les groupes sont des sites mal renseignés : soit l'information « période » est manquante, soit les pratiques funéraires ne sont pas décrites. Ils sont cependant placés au bon endroit : entre les autres.
Le troisième groupe depuis la gauche paraît correspondre à une période spécifique pour laquelle l'incinération est le mode privilégié de traitement des défunts.

Problématique 2

Il est possible de représenter les données issues de plusieurs extraits de bases de données agrégées au sein d'ArkeoGIS. Il s'agit ici de représenter des épées de l'âge du Fer (1er millénaire av. J.-C.) en tant que nœuds, liés en fonction de leurs lieux de découverte et de leurs datations. ArkeoGIS a permis d'identifier les objets au sein de différentes bases sources, le fichier .CSV exporté a immédiatement été reconnu par TULIP (EPSG, longitude et latitude). La création de ce type de représentation était jusqu'alors très chronophage, l'utilisation des fonctions des graphes permet d'économiser beaucoup de temps de traitement ici.



Perspectives

Les prochaines étapes seront de nous familiariser avec l'outil, afin de créer des requêtes plus signifiantes et exploitant les possibilités de paramétrage des graphes. Concrètement, au lieu de rajouter des colonnes dans nos tableurs, il paraît envisageable de paramétrer la distance entre les points, ou d'attribuer des rangs dès le premier traitement de l'information.
Ceci prendra du temps de formation et nécessitera de nouvelles rencontres, par la suite le croisement des données avec des données environnementales (type de faune marine à proximité des amphores, type de flore à proximité de gisements de cuivre) devrait être possible et devrait ouvrir de nouvelles formes de pluridisciplinarité à partir de lots de données préexistants.
Un autre aspect intéressant des graphes concerne la curation des données : les points isolés ou trop éloignés du centre du graphe sont soit trop peu caractérisés, soit en erreur. Une représentation graphique des erreurs sur un tableur permet un grand gain de temps pour les corrections.

Résumés de vulgarisation :

Dans le domaine marin, des protocoles d'observation développés dans de nombreux cadres produisent un grand volume de données hétérogènes, difficiles à agréger et à utiliser. Ce travail propose i) des méthodes, protocoles et recommandations pour construire et/ou soutenir la mise en place de réseaux de suivis multi-usagers,) des utilisations novatrices des données.
Deux cas d'étude ont été choisis : les habitats coralligènes à l'échelle de la Méditerranée et la colonisation de récifs artificiels dans différentes mers régionales.
L'expérimentation à large échelle se base sur des méthodes de mesure les plus simples possibles, décrites très explicitement dans des termes standardisés, sur des opérateurs inter-calibrés et une méthode de traitement des données. Un mécanisme de couplage de données de différentes origines reposant sur la requalification des facteurs descriptifs hétérogènes et une méthode d'analyse et de fouille de données basé sur la théorie des graphes sont proposées.

In the marine domain, observation protocols developed in many settings produce a large volume of heterogeneous data that are difficult to aggregate and use. This work proposes to develop i) methods, protocols and recommendations to build and / or support the establishment of multi-user monitoring networks, ii) innovative uses of data.
Two case studies were chosen: coralligenous habitats at the Mediterranean scale and the colonisation of artificial reefs in different regional seas.
Large-scale experimentation is based on the simplest possible measurement methods, described very explicitly in standardised terms, on inter-calibrated operators and a method of data processing. A mechanism for coupling data from different origins based on the requalification of heterogeneous descriptive factors and a method for analysis and data mining based on graph theory is also proposed.

Résumés de couverture / cover summary

Dans le domaine de l'environnement marin, des protocoles d'observation développés dans de nombreux cadres produisent un grand volume de données hétérogènes, difficiles à agréger car centrées sur l'utilisation souvent spécifique à un métier. L'accès et le partage des données à large échelle est pourtant incontournable pour mieux cerner les enjeux de protection de la biodiversité et des ressources marines, et anticiper leur détérioration irréversible. Pour répondre à ces enjeux, il est nécessaire de renforcer l'efficacité des systèmes d'acquisition de connaissances, et d'organiser l'accès aux données pour tous les utilisateurs potentiels. Améliorer la cohérence entre systèmes d'observation et systèmes d'information est l'objectif cadre de ce travail. La mise en place efficace de systèmes d'observation à large échelle nécessite un état des lieux des connaissances, des compétences disponibles et des verrous à lever. Les réseaux de suivi pour la protection environnementale deviennent multi-usagers et doivent produire suffisamment de descripteurs fiables pour élaborer une indication et un reportage performants.
Ce travail propose i) des méthodes, protocoles et recommandations pour construire et/ou soutenir la mise en place de réseaux de suivis de la biodiversité (du gène aux espèces et aux habitats), opérationnels et pérennes à toutes les échelles, ii) des utilisations novatrices des données, de leur conservation et de l'organisation de leurs accès permanent.
Deux cas d'étude ont été choisis : les habitats coralligènes à l'échelle de la Méditerranée et la colonisation de récifs artificiels (ARMS) dans différentes mers régionales, en focalisant sur : la construction de réseaux de suivi et d'observation pérennes, le partage efficace et l'inter-opération des connaissances à long terme et les méthodes d'analyses de données exploitant les avancées dans le domaine du *Big Data*, et de l'analyse de données hétérogènes sous forme de graphes.
Le test des protocoles élaborés dans le cadre de cette thèse montre qu'une expérimentation à large échelle doit être décrite très explicitement dans des termes standardisés au-delà du champ disciplinaire de l'écologie marine et se baser sur des méthodes de mesure les plus simples possibles. Ces tests ont aussi montré l'importance de la formation des opérateurs et d'une intercalibration itérative.
Le travail sur l'architecture des systèmes d'information met en évidence l'inévitabilité d'un système de gestion modulaire, orienté "métier" et décentralisé. Un mécanisme de couplage de données de différentes origines accessibles sous forme de flux paramétrables et ouverts est proposé (observations de terrain et données décrivant les contextes). Il s'appuie sur la requalification des facteurs descriptifs hétérogènes reposant sur un arbitrage collaboratif entre spécialistes. Grâce au développement d'un prototype, une méthode d'analyse et de fouille de données basé sur la théorie des graphes a été expérimentée. Différents types de démonstrations possibles partant des données ont été construits grâce à l'organisation d'ateliers multidisciplinaires de curation et de visualisation de données sous forme de graphes.
En conclusion, nous recommandons que l'information produite soit contrôlée par itérations en temps réel et que les processus de curation de la donnée soient mis en œuvre en même temps que la conception des procédés d'observation. Les définitions de standards et les accès aux données de contextes nécessitent un travail collaboratif, interdisciplinaire, itératif produit sur le long terme sur un plan international. Ils doivent être considérés pour toutes leurs utilisations possibles.
En perspective, l'utilisation de la grille de calcul pour faire de la fouille de graphes de manière parallélisée a été préparée lors des ateliers, avec le challenge d'un passage à l'échelle avec des données distribuées et très hétérogènes formant des graphes de plus d'un milliard de nœuds et plusieurs centaines de milliards de liens.

***

In the field of marine environment, observation protocols developed in many settings produce a large volume of heterogeneous data, which are difficult to aggregate since they focus on the use that is often specific to a profession. However, access and sharing of data on a large scale is essential to better understand the challenges behind protecting biodiversity and marine resources, and to anticipate their irreversible deterioration. To address these challenges, there is a need to strengthen the effectiveness of knowledge acquisition systems, and to organise data access for all potential users. Improving the consistency between observational systems and information systems is the main objective of this work. Effective implementation of large-scale observational systems requires an inventory of knowledge, available skills and locks to be lifted. Monitoring networks for environmental protection become multi-user and must produce sufficient reliable descriptors to develop an effective reporting.
The present work proposes to develop i) methods, protocols and recommendations to build and / or support the establishment of networks for monitoring biodiversity (from gene to species to habitats), operational and sustainable at all scales, (ii) innovative usage of data, their conservation and the organisation of their permanent access.
Two case studies were chosen: coralligenous habitats at the Mediterranean scale and the colonisation of artificial reefs (ARMS) in various regional seas, focusing on: the construction of sustainable monitoring and observational networks; efficient sharing and interoperability of long-term knowledge along with data analysis methods exploiting advances in the field of *Big Data*, and the analysis of heterogeneous data in the form of graphs.
The testing of the proposed protocols developed during this thesis shows that a large-scale experiment must be described very explicitly in standardised terms beyond the disciplinary field of marine ecology and to be based on the simplest possible measurement methods. These tests also demonstrated the importance of operator's training and iterative intercalibration.
The work on the architecture of informational systems highlights the inevitability of a modular management system, business-oriented and decentralised. A mechanism for coupling data of different origins accessible in the form of parameterisable and open flows is proposed (field observations and data describing the contexts). It is based on the requalification of heterogeneous descriptive factors built on a collaborative arbitration between specialists. Thanks to the development of a prototype, a method of analysis and data mining based on graph theory has been experimented. Diverse types of possible demonstrations based on the data were constructed through the organisation of multidisciplinary curation and graphical data visualization workshops.
In conclusion, we recommend that the information produced is iteratively controlled in real time and that the curation processes of the data is implemented at the same time as the design of the observational processes. Standard definitions and access to context data require collaborative, interdisciplinary, iterative, long-term, international work. They must be considered for all their possible uses.
In perspective, the use of the computation grid to carry out graph mining in a parallel way was prepared during the workshops, with the challenge of scaling up distributed and very heterogeneous data and forming graphs with more than one billion nodes and several hundred billions links.