État des lieux MoDaL (Multi-Scale Data Links)

Aurélien Cornet, Christian Barillot, Olivier Dameron, Alban Gaignard, Camille Maumet, Richard Redon, Anne Siegel

MoDaL (Multi-Scale Data Links)
État des lieux

Aurélien Cornet*, Christian Barillot*, Olivier Dameron*, Alban Gaignard*, Camille Maumet*, Richard Redon*, Anne Siegel*

*Auteurs classés par ordre alphabétique à l'exception du premier.

28 avril 2020

Table des matières

1 Introduction

MoDaL, pour « *Multiscale Data Links* », est un projet fédérateur groupant la région Bretagne et la région Pays de la Loire, financé par Biogenouest, et porté par Christian Barillot d'IRISA Rennes ainsi que Richard Redon de l'institut du Thorax à Nantes. Le projet est également piloté par Olivier Dameron, Camille Maumet et Anne Siegel d'IRISA Rennes ainsi qu'Alban Gaignard de l'institut du Thorax à Nantes.

Les recherches explorant le domaine du vivant sont confrontées de manière croissante au besoin de devoir lier des données hétérogènes multi-échelles. Ces données sont pour la plupart exploitées "en silos", c'est-à-dire sans moyen de pouvoir réaliser des analyses fines sur leurs complémentarités. Malgré les initiatives de mutualisation et de standardisation au travers de grandes infrastructures de recherche comme FLI (France Life Imaging), FBI (France BioImaging), ou encore l'IFB (Institut Français de Bioinformatique), trois réseaux ayant pour objectif de fédérer respectivement les acteurs nationaux en imagerie in-vivo, in-vitro et en bioinformatique, il est aujourd'hui très difficile d'envisager une exploitation algorithmique et statistique conjointe de ces diverses sources de données.

Les principaux objectifs du projet fédérateur MoDaL s'articulent autour de I) l'établissement d'un état des lieux des acteurs et des infrastructures disponibles à l'échelle du Grand Ouest afin d'identifier un ou plusieurs verrous communs II) la proposition de démonstrateurs technologiques mettant en jeu et adressant les problématiques identifiées, et III) la fédération d'une communauté scientifique autour de ces problématiques.

Ce document est le résultat des entretiens et du recensement des différentes initiatives menés au cours de la première année du projet fédérateur MoDaL. Il vise à répondre au premier objectif : l'état des lieux et l'identification de verrous communs. Les différentes interviews, individuelles ou groupées, sont résumées et une synthèse globale est proposée. Ce document sert de première base à la réflexion et la proposition de démonstrateurs technologiques adressant les verrous identifiés.

La section 2 décrit la méthodologie utilisée pour recueillir les informations nécessaires à la construction de ce document. La section 3 couvre le workshop MoDaL. La section 4 regroupe les différentes interviews auprès de la communauté scientifique du grand ouest. Enfin, la section 5 discute des perspectives pour la suite du projet MoDaL basées sur les résultats de l'état des lieux.

2 Méthodologie

Les informations ont été recueillies par le biais de deux méthodes :

1 Effectuer des interviews auprès de la communauté scientifique du grand ouest

Ces interviews ont été réalisées en suivant une charte afin d'assurer un degré d'homogénéité des informations. Cette charte est décrite dans la figure 1 ci-dessous.

Questions et sujets de discussion
Pouvez-vous vous présenter et me dire quels sont vos sujets de recherches actuels ?
Quels sont vos domaines de compétences ?
Que fait votre équipe/laboratoire en général ?
Pouvez-vous me parler de projets passés ou en cours auquels vous avez pris part ?
Quels types de données vous ou votre équipe gérez-vous ?
Avez-vous à croiser des données de différents types au cours de vos recherches ?
Si oui, rencontrez-vous des difficultés ?
Comment sont stockées ces données ?
Pouvez-vous me parler d'outils que vous utilisez ?
Discussion libre
Intérêt des données de nos jours
Pourquoi croiser les données ?
Les verrous actuels du domaine d'étude (de l'interviewé) ou de la biologie en général

2 Inviter des spécialistes au cours d'une journée dédiée

Le workshop MoDaL est une journée organisée pour discuter précisément des problématiques du projet. Des experts sont invités pour présenter leurs travaux et pour échanger au cours de discussions libres sur les différents points d'importances et verrous potentiels.

3 Workshop MoDaL

Le workshop est une journée, s'étant tenue à l'IRISA Rennes le 11/07/19, organisée autour des thématiques du projet MoDaL. Cette journée a deux objectifs : d'un côté permettre aux experts manipulant déjà des données de présenter leurs travaux et d'échanger sur leurs expériences et points de vues. De l'autre sensibiliser la communauté à ces problématiques et favoriser des collaborations futures. Le profil des invités visés est donc large afin de représenter idéalement le plus de domaines de compétences. Lors de cette première édition du workshop, 26 participants couvrant divers domaines de la recherche en santé, végétale et marine se sont réunis au cours de 7 présentations et de discussions libres. Les détails de cette journée ont fait l'objet d'un rapport antérieur [1].

Pour reprendre les conclusions de cette journée, les discussions ont convergé autour de deux grands thèmes liés : les données et les ontologies. Pour ce qui est des données, l'élément majeur qui ressort est que le déploiement de ressources et de solutions, comme les outils d'analyse ou les bases de données, se fait strictement dans le cadre d'un projet. La spécificité du sujet d'étude et de ses données dicte le format, la structure de la base de données et des pipelines d'analyse, ce qui rend difficile un croisement ultérieur avec d'autres sources de données, ou une réutilisation des outils dans un autre contexte. En règle général ces ressources sont très peu remobilisées en dehors du projet, ce qui implique une perte pour la communauté scientifique. Il faudrait rendre plus explicite les schémas et structures des bases de

données et privilégier l'usage de formats de fichier largement utilisés pour les outils d'analyse.

Il existe déjà de nombreuses ontologies, et ce dans tous les domaines de la biologie. Des guides existent et des catalogues répertorient les ontologies et les connecteurs permettant de faire le pont entre elles. Annoter ses données permet en outre de les croiser plus facilement inter et intra-institut. Pourtant, comme pour le premier thème, de nouvelles ontologies sont souvent créées dans le cadre d'un projet, soit parce qu'aucune ontologie déjà existante ne convient, soit par soucis de praticité et de rapidité. Le problème est que ces ontologies ne sont jamais "connectées" aux autres, ce qui rend difficile le croisement des données associées. Le partage de données annotées avec des ontologies différentes et sans connecteur peut être source d'erreurs, ou rendre le partage fastidieux car il faut normaliser manuellement. Une idée serait de développer des ontologies communes, des initiatives existent déjà mais elles le sont à l'échelle d'un domaine. C'est une autre solution en parallèle des connecteurs pour faire le lien entre les différentes ontologies.

Le workshop MoDaL a donné les premiers éléments d'un verrou commun aux domaines de la biologie : les ressources, c'est à dire les données, outils d'analyse et ontologies, ne voient pas leurs utilisations dépassées le cadre du projet auxquelles elles sont associées, elles manquent d'interopérabilité. Cette piste est étayée dans la partie suivante au travers de nombreuses interviews avec des acteurs en santé, végétal et marin.

4 Acteurs du grand ouest

Les acteurs du projet MoDaL travaillent sur des modèles d'études et des sujets de recherche variés. Malgré ces différences de domaine les méthodes et outils utilisés partagent des similarités, ce qui est également le cas pour les challenges rencontrés. Les différentes interviews sont séparés entre la santé, la biologie végétale et la biologie marine.

4.1 Sciences médicales

Empenn, équipe de l'IRISA Rennes spécialisée dans les neurosciences, rattachée au CNRS, l'Université Rennes 1 et l'INRIA

Interviews : Julien Louis, Francesca Galassi, Yao Chi, Isabelle Corouge, Elise Bannier

Empenn (qui signifie "cerveau" en breton) est une équipe pluri-disciplinaire dont les sujets de recherche s'articulent autour de l'imagerie médicale, principalement la neuroimagerie, et la neuroinformatique. L'un des objectifs principaux est le développement de biomarqueurs pour les maladies cérébrales, et la translation de ces recherches vers la pratique clinique. Cela se traduit par la proposition de nouvelles méthodes statistiques et informatiques, de modélisations de l'état structurel, fonctionnel et morphologique du cerveau, afin de mieux diagnostiquer et traiter les maladies.

Certains membres travaillent sur deux initiatives internationales de standardisation et de structuration des données en neuroimagerie : BIDS (Brain Imaging Data Structure) et NIDM (Neuroimaging Data Model) [2]. La première a pour objectif d'apporter des règles d'organisation des dossiers et fichiers de données brutes, c'est à dire les données d'acquisition (comme de l'imagerie Time of Flight ou TOF) en stockage sur un ordinateur ou sur des serveurs. Cela permettra notamment de faciliter le traitement automatique de ces fichiers (par des scripts ou des logiciels dédiés), ainsi que leur partage inter et intra-institut car l'arborescence et la convention de nommage des fichiers seront standardisées. La seconde, NIDM, est un projet plus ambitieux faisait le lien avec BIDS, et ayant pour objectif de standardiser l'organisation des fichiers et les ontologies pour la neuroimagerie. Cela inclut les fichiers bruts d'acquisitions, les fichiers intermédiaires de traitement par les différents workflows et pipelines d'analyse, les fichiers de résultat traités, et également une uniformisation des champs de métadonnées. Ces initiatives ont pour but de partager et croiser plus facilement les données quel que soit le lieu ou la méthode d'acquisition.

L'équipe travaille également en partenariat avec la plateforme Neurinfo pour développer une base de données appelée "Shanoir" conçue pour partager, archiver et visualiser des données de neuroimagerie. Elle est conçue pour offrir un support aux études cliniques, multi-site ou non, et propose différentes fonctions dont l'anonymisation des données patients. Shanoir organise les données avec l'ontologie OntoNeuroLOG, mais seulement pour une classification en interne (pas d'export des métadonnées liées aux projets stockés). Il arrive trop souvent que les langages et les conventions de nommage créés dans le cadre d'un projet le soient sans aucune optique d'utilisation ultérieure. Sans moyen de décrypter les termes utilités, les données deviennent difficile à remobiliser alors qu'elles pourraient être croisées avec celles d'autres études. Le cas échéant les données doivent être acquises à nouveau ce qui représente du temps et un coût non négligeable. L'enrichissement en métadonnées en utilisant des vocabulaires contrôlées a des avantages, mais cet enrichissement doit avoir un sens, à savoir le coût d'enrichissement ne doit pas dépasser le gain. Multiplier les champs de métadonnées demandent du temps pour les remplir et beaucoup de biologistes n'en voient pas l'intérêt par exemple. Dans le cas de Shanoir, la base de données propose un nombre important de champs de métadonnées pour la description des fichiers mais la grande majorité sont facultatifs. De plus les développeurs de la plateforme travaillent en partenariat avec les cliniciens pour concevoir des "study-cards", permettant de pré-remplir certains champs de métadonnées en fonction de la méthode et de la machine ayant servie à l'acquisition. S'il fallait rentrer les données à la main et ce pour l'intégralité des champs, il est probable que Shanoir serait beaucoup moins utilisée.

LTSI, laboratoire spécialisé dans l'imagerie médicale à Rennes, rattaché à l'INSERM et à l'Université Rennes 1

Interview : Bernard Gibaud

Les recherches menées au LTSI à Rennes (Laboratoire Traitement du Signal et de l'Image) s'articulent autour des sciences de l'information en santé. Imagerie vasculaire, chirurgie assistée par l'image sont des exemples des nombreux thèmes abordés. Les recherches s'articulent autour d'un noyau "signal-modèle-image" et ont

pour objectif une aide à l'interprétation et à la décision. Pour se faire le laboratoire dispose de plusieurs équipes pluri-disciplinaires se focalisant chacune sur un sujet d'étude précis.

L'équipe mediCIS du LTSI se concentre sur la modélisation des connaissances et procédures chirurgicales et interventionnelles pour l'aide à la décision. Dans un contexte ou la quantité de données cliniques est croissante, la charge cognitive des chirurgiens augmentent également. Le type de données à analyser, les différents outils développés et à maîtriser sont autant d'enjeux qui peuvent avoir un impact sur le processus décisionnel. L'équipe s'appuie sur 2 axes méthodologiques : une approche numérique pour la modélisation des connaissances basée sur la fusion de données. Et une modélisation conceptuelle basée sur la création d'ontologies formelles.

La tendance à la médecine de précision implique de pouvoir avoir accès aux données d'un patient, de partager ces données entre des hôpitaux ou laboratoires d'analyse, d'avoir une description claire des méthodes d'obtention de ces données. Si actuellement c'est le cas pour les données d'acquisitions (ou données brutes), il existe peu d'ontologies et de méthodologies pour annoter et enrichir les données de traitement (données transformées). Le projet européen MEDIRAD en est un exemple. Il vise à mieux comprendre l'impact sur la santé des radiations lors d'examens cliniques, et pour se faire plusieurs instituts partagent des imageries radiologiques. Ces images doivent être croisées et analysées ensemble, mais les métadonnées sur l'acquisition doivent être précises. Le partage de données cliniques sur des patients soulèvent aussi d'autres questions, dans un contexte de coopération entre instituts, comment assurer la sécurité des données tout en permettant leurs utilisations pour la recherche ? Certaines des données jugées sensibles peuvent être enlevées lors de l'anonymisation, et il arrive que ces données soient importantes pour la recherche.

SERPICO, équipe de l'IRISA Rennes spécialisée dans l'imagerie moléculaire et cellulaire, rattachée au CNRS, l'Université Rennes 1 et l'INRIA

Interviews : Charles Kervrann, Sylvain Prigent

Au sein de l'IRISA à Rennes, les recherches de l'équipe SERPICO ont pour objectif de mieux comprendre et caractériser la dynamique de la coordination et de l'organisation de complexes moléculaires à l'échelle de la cellule. Ils développent des outils et méthodes permettant de faire le pont entre l'échelle moléculaire et cellulaire, afin de pouvoir observer l'évolution dans le temps des complexes moléculaires. Pour cela l'équipe se base sur des techniques de microscopie optique multidimensionnelle et multimodale couplées à du marquage à la GFP (Green Fluorescent Protein). Les résultats de ces recherches et les outils et méthodes développés ont pour but d'être appliqués dans d'autres domaines, dont le médical.

Les données d'acquisition sont de l'imagerie 3D, voir 4D lorsque la composante temporelle est ajoutée. Ces données peuvent être extrêmement lourdes ce qui implique d'avoir une solution de stockage adéquate. Néanmoins la durée de stockage dépend de la durée du projet, ou plus sobrement tant que ces données ont une utilité pour le projet en cours. Cette durée est donc projet dépendant et peut

parfois poser problème sur de longues périodes. L'équipe est multi-discplinaire et travaille en partenariat avec d'autres instituts, ce qui pose parfois des problèmes d'accessibilité des outils d'analyses pour les biologistes. Des logiciels développés peuvent être trop compliqués à prendre en main pour des utilisateurs ne possédant pas les compétences requises nécessaires, comme par exemple Matlab, C++ ou Python, ainsi que des notions en machine learning. Un processus de discussion avec l'utilisateur cible est indispensable, et en règle général l'outil devrait pouvoir être utilisé par le plus grand nombre, Les développeurs de l'équipe veillent à ce que les solutions délivrées soient agnostiques. Pour citer quelques exemples : ATLAS et Flowscope.

IGDR, Institut Génétique et Développement de Rennes, rattaché au CNRS et à l'Université Rennes 1

Interview : Gwenael Rabut, Christophe Heligon

L'IGDR est un laboratoire s'articulant autour de 3 grands axes principaux impliquant la conception d'approches innovantes et multi-disciplinaires : (i)structure et dynamique de la molécule à la cellule, découvrir comment la structure impacte la fonction d'une molécule ou d'un complexe moléculaire. (ii) De la cellule à l'organisme, comprendre comment une cellule unique peut donner naissance à un organisme entier formé de cellules différentes interagissant entre elles. (iii) Du gène à la maladie, trouver des prédispositions génétiques et des altérations génétiques causales et comprendre l'impact sur formation et le développement de maladies génétiques et de cancer.

L'une des fonctions étudiées est l'ubiquitylation des protéines, et ses mécanismes régulateurs à l'échelle moléculaire. Cette modification post-traductionnelle des protéines, indispensable à la survie de la cellule, encode notamment des signaux moléculaires qui contrôlent la durée de vie et l'activité de très nombreux protéines. Des déséquilibres de cette fonction sont le plus souvent responsables de cancer.

Pour étudier cette fonction, une combinaison d'approches mêlant génétique, protéomique, biochimie et imagerie cellulaire est nécessaire. Ces différentes approches génèrent des données hétérogènes qu'il faut stocker, annoter, croiser et analyser ensemble. Ces questions sont parfois difficiles à appréhender et à gérer sans les connaissances nécessaires en informatique ou bioinformatique. Il existe des structures et plateformes de stockage spécialisées dans le grand ouest pour un grand nombre de domaine en biologie. Néanmoins il apparaît souvent que les personnes connaissant ces ressources travaillent déjà dans le milieu, ou ont l'habitude de traiter des données. Certains biologistes ayant des besoins en stockage pour leurs données, ou des besoins d'accès à du matériel précis peuvent ne pas être au courant de l'existence de ces plateformes, démontrant un manque de communication sur les ressources disponibles. L'utilisation de ces ressources peut aussi être difficile sans certaines connaissances en informatique, un point déjà discuté avec SERPICO, même si les plateformes offrent aujourd'hui de plus en plus un accompagnement dans l'utilisation de leurs services (comme FLI par exemple).

Institut du Thorax, spécialisé dans l'étude des maladies cardiovasculaires, métaboliques et respiratoires à Nantes, rattaché à l'INSERM, au CNRS et à l'Université de Nantes

Interview : Richard Redon, Christian Dina, Audrey Bihouée, Perrine Paul-Gilloteaux, Romain Bourcier, Florent Autrusseau

L'unité de recherche de l'Institut du Thorax compte aujourd'hui près de 160 personnes réparties en plusieurs équipes. Au sein de cette unité, l'équipe I se concentre sur les maladies cardiaques, en particulier leurs héritabilités, avec un accent sur l'arythmie et les problèmes liés aux valves cardiaques. Ces recherches impliquent la collaboration d'épidémiologistes, de généticiens et de cliniciens, avec l'appui d'outils informatiques et bioinformatiques.

L'institut du Thorax possède une plateforme de bioinformatique nommée Geno-BIRD offrant des services en génomique et transcriptomique, ainsi que des outils d'analyse et de stockage des données. Un support pour les projets ainsi que des formations sont également proposés. La plateforme n'a pas de projet à proprement parlé mais répond à des demandes des différentes équipes. Les sujets d'études sont variés et le sont également les types de données à croiser. La structuration et les ontologies utilisées sont dépendantes des projets et les données sont rarement réutilisées par la suite. En règle générale les données en biologie ont un cycle de vie court, bien que cela varie d'un domaine à l'autre et est aussi dépendant de la longueur du projet associé. En génomique les données sont le plus souvent annotées ce qui permet de pouvoir les réutiliser, mais c'est un cas rare. En général, les données non annotées sont "inutilisables" des années plus tard car les conventions de nommage ont été perdues, et demanderaient un temps considérable à décrypter.

Dans la même idée que les données et les ontologies, les outils, workflows et pipelines d'analyse gagneraient à être plus flexibles afin de s'adapter à un plus grand nombre de cas d'étude. Il faut favoriser leurs réutilisations en dehors du sujet initial. Cependant en augmentant la flexibilité, on peut perdre en précision et ainsi perdre l'utilité première de l'outil, c'est d'ailleurs l'une des principales motivations du développement de nouvelles solutions. La question se pose sur l'équilibre entre précision ou portée d'utilisation. En règle générale un outil est développé pour un besoin ou un projet, et il peut être plus aisé ou rapide de le développer spécifiquement pour répondre au problème que d'en adapter un déjà existant. Les communautés scientifiques continueront à développer de nouveaux outils quoi qu'il arrive, et forcer l'utilisation à une liste restreinte n'est pas envisageable. Une solution serait donc de favoriser leur interopérabilité, via des formats de fichiers d'entrée et de sortie similaires par exemple. Ces questions d'interopérabilité font déjà l'objet d'études, notamment les principes FAIR (Findability, Accessibility, Interoperability, Reusability).

Dans la recherche, les données d'acquisitions (ou données brutes) sont essentielles à toute étude, ce qui peut rendre difficile leurs accès. Certaines structures productrices de données sont conscientes de ce "pouvoir" et peuvent en limiter l'accès. Il y a un aspect politique non négligeable avec les partenariats en biologie. Ce point est particulièrement spécifique au domaine clinique, certaines données étant sensibles il peut être difficile d'obtenir les accréditations pour y avoir accès.

Santé
Sécurité des données patients
Difficulté d'accès aux données d'acquisitions (données brutes)
Perte d'informations clés pour la recherche lors de l'anonymisation des données
Ontologies spécifiques non connectées, restent dans le cadre du projet
Coût versus gain de l'enrichissement en métadonnées
Manque de communication sur les outils proposées à la communauté
Outils développés manquent de flexibilité, non agnostiques, utilisation restreinte au projet
Structuration des base de données et métadonnéesnon explicite en dehors du projet

4.2 Biologie végétale

IRHS, institut de recherche en agronomie à Angers, rattaché à l'INRA et à l'Université d'Angers

Interview : Claudine Landes, Sylvain Gaillard, Tanguy Lallemand, Sandra Pelletier, Fabrice Dupuis, Marie-charlotte Guillou, Martial Briand, Armelle Darrasse, Hervé Autret, Sophie Paillard, Julie Bourbeillon, Jean-Pierre Renou

L'IRHS appuie ses recherches sur les questions fondamentales et stratégiques de la biologie des produits horticoles et de la production des semences. La structure est conçue pour mettre en oeuvre des approches de biologie intégrative en coordonnant les expertises de généticiens, sélectionneurs, phytopathologistes, physiologistes, biochimistes, bioinformaticiens et statisticiens. L'organisme réunit 13 équipes de recherches séparées selon leurs domaines de recherche et sujet d'étude. L'équipe BIDefI regroupe mathématique, informatique et biologie et a pour mission d'apporter des solutions aux problématiques nouvelles en biologie, la mise en place et le développement de logiciels pour l'institut, et l'accompagnement des utilisateurs de ces outils dans une démarche de formation. Ces recherches se font en collaboration avec d'autres équipes de l'IRHS. Les sujets principaux d'étude sont les rosacées, en particulier la rose et le pommier. Cela se traduit par l'annotation fonctionnelle et relationnelle des gènes impliqués dans les réponses au stress, l'intégration de données, l'extraction et le croisement de connaissances à partir de la littérature.

L'équipe a pour mission la conservation et l'archivage des espèces biologiques étudiés au sein de la structure. Un projet de base de données centralisée a été monté depuis 2006 et est toujours en cours. La tache est compliquée, car l'objectif est de pouvoir stocker des espèces différentes, de la donnée brute à la donnée traitée. Les défis sont multiples : une quantité importante données, l'hétérogénéité des données qui peut poser problème lors de la conception de la structuration, les ontologies utilisées pour décrire les données et les métadonnées peuvent ne pas être compatibles d'une espèce à l'autre. La question de la centralisation contre la décentralisation se pose, et l'option de lier des bases de données plus petites ne gérant qu'une espèce, avec une ontologie propre mais en pouvant faire le lien avec les

autres a été évoqué. Cela implique de favoriser l'interopérabilité des outils, des formats de fichier et des ontologies. Des guides d'utilisation, ou une liste de format et d'ontologies les plus utilisés pourraient grandement bénéficier à la communauté.

Contrairement à la santé, les données dans le végétal sont moins contraintes et n'ont pas le problème de l'anonymisation, le partage est bien plus fréquent même si cela reste dans le cadre de collaborations bien précises. De manière générale les données sont rendues publiques seulement après la publication de l'article pour éviter la compétition. Les données sont également plus facile à acquérir et dans une plus grande quantité, comme les cohortes de graines par exemple, ce qui donne plus de poids statistique aux études.

Il existe beaucoup d'ontologies dans le végétal, certaines sont à l'échelle d'une espèce et d'autres comme Plant Ontology vise à décrire l'entièreté du végétal, les deux ont leurs points faibles et points forts. L'ontologie généraliste permet de faire le lien entre différentes espèces de manière aisée, mais est extrêmement lourde à l'utilisation, le temps d'annotation pouvant être très long si l'on est pas habitué. De plus ce genre d'ontologie peut manquer de précision pour décrire des processus très particulier. Pour cette raison, lors d'un nouveau projet, les équipes ont tendances à créer une nouvelle ontologie qui correspond parfaitement au besoin. Cela peut poser problème si le vocabulaire utilisé ne dispose pas de connecteurs avec les autres ontologies. Sans équivalent il peut être difficile de partager ou croiser ces données avec celles d'une autre étude. L'idée est donc de favoriser les ontologies spécifiques pour annoter précisément les données, mais d'utiliser des connecteurs pour pouvoir faire le lien avec d'autres ontologies.

Végétal
Les mêmes champs de métadonnées peuvent pointer vers des mécanismes différents en fonction de l'espèce
Outils développés manquent de flexibilité, non agnostiques, utilisation restreinte au projet
Structuration des base de données et métadonnées non explicite en dehors du projet
Ontologies spécifiques non connectées, reste dans le cadre du projet

4.3 Biologie marine

Station biologique de Roscoff, institut de recherche sur les écosystèmes marins à Roscoff, rattaché au CNRS et à l'Université de Sorbonne

Interview : Erwan Corre, Mark Hoebeke, Gabriel Markov, Sébastien Collin, Nicolas Henri, Loraine Guegen, Mark Cock, Gildas Le Corguillé, Olivier Godfroy

La Station biologique de Roscoff s'intéresse principalement à la biologie fondamentale et à l'étude de la biodiversité et des écosystèmes marins. Ces recherches font appel aux méthodes les plus modernes de la biologie moléculaire et cellulaire, en particulier la génomique, ainsi qu'aux interfaces avec les sciences de l'environnement, la chimie et les mathématiques. Elles ont pour but de mieux comprendre l'évolution de la vie, ainsi que le fonctionnement des écosystèmes et l'adaptation des organismes marins face au changement global. Les principaux modèles expéri-

mentaux incluent des bactéries, des végétaux comme les algues rouges et brunes, des invertébrés comme l'oursin et des vertébrés comme la lamproie.

La station dispose également d'une plateforme de bioinformatique nommée ABiMS, et offre une infrastructure de calcul et de stockage. Au sein de cette infrastructure les utilisateurs sont libres de l'organisation de leurs données, ce qui veut aussi dire que la structure de ces données et les ontologies sont projets dépendants. Les ingénieurs de la plateforme sont plus ou moins impliqués dans l'organisation des données d'utilisateurs individuels (cela varie d'un cas à l'autre en fonction de la demande), auxquels cas ils apportent des conseils d'utilisation sur la structuration des bases de données ou l'utilisation des outils. Ils interviennent également à différents niveaux dans des projets en collaboration avec la plateforme pour réaliser différentes analyses OMICS. Pour le moment, la majorité des données de la plateforme sont des données de génomiques, mais cela évolue rapidement. Les problématiques de structuration et de croisement de données hétérogènes ne se posent pas. Le développement de bases de données couplé à système de requête sur les données d'observation est en cours, ces données sont de plus en plus mises à disposition du public via des portails et cela implique d'utiliser une structuration des données et du vocabulaire contrôlés. Les données sont également mises à disposition pour une communauté plus vaste dans des entrepôts internationaux dont Pangaea.

La plateforme n'est pas encore confrontée à la gestion de différents types de données. Dans le projet TARA Oceans, les données sont stockées dans plusieurs endroits, mais une partie des données brutes est stockée à la station (en grande partie des fichiers FASTQ). Ces données ne sont pas organisées de manière structurées. L'un des leviers qui amène à la structuration des données est la publication de l'article, les données doivent être structurées pour être soumises (au NCBI par exemple). Plus généralement lorsque les données doivent être partagées, ou rendues publiques, il faut structurer ces données.

Des outils ont été développés dans le cadre du projet TARA Oceans pour visualiser les données traitées des catalogues de gènes (bactériens et eucaryotes). La problématique a ensuite migré sur les données de metabarcoding avec les projets OGA (Ocean Gene Atlas) et OBA (Ocean Barcode Atlas). L'idée de cette application est de pouvoir accéder publiquement à ces données via des requêtes (par exemple sur les noms ou sur la similarité des séquences) et voir la distribution sur le globe.

Le partage des données se fait surtout en fonction des collaborations sur les projets, ou lors de la publication d'un article. Les données et toutes les informations nécessaires pour refaire les expériences sont mises à disposition sur Pangea, un service d'hébergement de données. Les biologistes ont tendances à garder les données s'ils ne sont pas contraints de les partager, jusqu'à la publication de l'article. Les financements sont durs à trouver, et ce sont souvent des séries temporelles de plusieurs années, donc beaucoup d'efforts et la tâche est ingrate. De plus si les données sont parfaitement annotées, des connaissances en biostatistique sont suffisantes pour faire ses propres corrélations. La pression de la publication ou les données représentent tout le travail ne favorise ni le partage, ni une vision au delà de l'article, ce qui explique en partie la faible réutilisation des données.

Le besoin de croiser des données et de développer des outils et des ontologies pour structurer ces données est dans la grande majorité des cas liés à un projet. Il est donc difficile d'appréhender l'utilité de ces données passé ce projet. Néanmoins le point à adresser est le suivant : lorsque l'on rentre dans ce cas, comment faire pour assurer que les données pourront être ré-utilisées, croisées avec d'autres études ultérieurement ? La structuration des données et les ontologies utilisées semblent être dépendant du projet ou de l'espèce étudiée et cela rend difficile le croisement dans un contexte plus large. Une partie du problème est humain, mais un début de solution pourrait être d'avoir une liste des formats et des ontologies (BridgeDB) favorisant l'interopérabilité.

L'enrichissement de ces données en métadonnées a également un coût, il faut qu'il y ait un réel besoin, un gain même, à rendre ses données structurées et interopérables. L'un des points majeurs aujourd'hui est justement l'obligation de publier ses données avec l'article. Selon la complexité des ontologies utilisées, il faut parfois engager un expert des données, et tout le monde n'en voit pas l'utilité. Cela dépend des domaines, il y a une obligation de publication des données génomiques mais par pour les données de biodiversité (par exemple données de comptage), il y a donc un décalage au niveau des domaines.

Il peut également être difficile d'enrichir ses données en métadonnées, il arrive parfois qu'aucune ontologie existante ne permette de décrire les données. Dans ce cas le seul moyen est de créer sa propre ontologie spécifique, et c'est dans ce cas justement qu'il serait intéressant d'avoir un "lien" permettant d'aligner les ontologies, un point déjà relevé plus haut.

Marin
Les données de biodiversité n'ont pas l'obligation d'être publiées
Coût versus gain de l'enrichissement en métadonnées
Outils développés manquent de flexibilité, non agnostiques, utilisation restreinte au projet
Ontologies spécifiques non connectées, reste dans le cadre du projet
Structuration des base de données et métadonnées non explicite en dehors du projet

4.4 Synthèse des interviews

Malgré la diversité des échelles, des modèles d'études et des sujets de recherche, des éléments se recoupent et des verrous apparaissent. La plupart de ces verrous sont spécifiques d'un ou deux domaines mais certains sont communs à l'ensemble des domaines et pointent vers la réutilisation des ressources au delà du projet associé.

Ontologies	Données	Outils
Les ontologies sont conçues dans le cadre du projet mais n'ont pas de connecteurs	Les structures des bases de données et des métadonnées utilisées ne sont pas assez explicites	Les outils d'analyse utilisent des formats bien spécifiques et répondent à des problèmes précis du sujet d'étude en cours
Difficulté de croiser les données annotées avec celles d'autres projets	Une fois le projet terminé, il est difficile de remobiliser les données sans les informations pour les décrypter	Manque de flexibilité, difficile à utiliser sur d'autres cas

5 Conclusion et perspectives

Cette première année du projet MoDaL avait pour objectif principal d'effectuer un état des lieux des verrous communs en biologie à l'échelle du grand ouest, afin de servir de base pour la proposition et le développement d'un démonstrateur technologique. Le postulat est le suivant, les différents domaines de la biologie sont confrontés au besoin croissant de croiser des données hétérogènes issues de sources multiples. Pour affronter cette problématique, il faut mutualiser les efforts en identifiant un ou plusieurs verrous communs sur lesquels se concentrer. Par le biais des nombreuses interviews et du workshop, plusieurs éléments ont pointé vers un tel verrou : il existe un manque d'intéropérabilité entre les ressources déployées dans les différents domaines de la biologie, ces ressources étant notamment des données, des outils d'analyses, ou des ontologies. De part la nature de la recherche, des financements et de la publication, les ressources sont déployées majoritairement dans le cadre d'un projet précis. Cela signifie souvent une structuration des données et des ontologies utilisées propres au dit projet. Outre leurs réutilisations ultérieures au projet pouvant devenir difficile sans les clés pour comprendre leurs structurations, cela devient presque impossible si ces ressources doivent être croisées avec celles d'un autre projet présentant lui même une structuration et ontologie spécifique. La recherche profiterait pourtant grandement d'une meilleure intéropérabilité de ces ressources. Il est certain que les données récoltées lors de projets s'étalant sur plusieurs années pourraient servir dans d'autres contextes, et que des outils d'analyses développés pourraient, modulo quelques ajustements, être utilisés sur d'autres jeux de données issus de sujets différents. Ces données perdues représentent un temps et un coût non négligeable gaspillés, sans parler de leurs valeurs potentielles sur des enjeux importants en biologie. Néanmoins la situation avance, lentement mais sûrement, les mentalités évoluent autour de l'intérêt de ces ressources et de plus en plus d'initiatives voient le jour, que ce soit les différentes plateformes et noeuds nationaux, ou les projets internationaux.

Références

[1] Aurélien Cornet, Christian Barillot, Olivier Dameron, Alban Gaignard, Camille Maumet, Richard Redon, and Anne Siegel. WorkShop MoDaL (Multi-Scale Data Links). Research report, IRISA, Inria Rennes, March 2020.

[2] Camille Maumet, Tibor Auer, Alexander Bowring, Gang Chen, Samir Das, Guillaume Flandin, Satrajit Ghosh, Tristan Glatard, Krzysztof J Gorgolewski, Karl G

Helmer, Mark Jenkinson, David B Keator, B Nolan Nichols, Jean-Baptiste Poline, Richard Reynolds, Vanessa Sochat, Jessica Turner, and Thomas E Nichols. Sharing brain mapping statistical results with the neuroimaging data model. *Scientific data*, 3 :160102, 2016.