Les **défis méthodologiques** et leurs outils

Observation à distance

Auteurs : Thomas Houet (LETG), Yan Ropert-Coudert (CEBC), Laurent Longuevergne (Géosciences Rennes)

3 PRIORITÉS SCIENTIFIQUES À ABORDER D'ICI 2030

- **Inciter et renforcer des systèmes d'observation à long-terme en mobilisant des modalités de captation complémentaires**
- **Fédérer et structurer l'observation à distance (développement instrumental, acquisition et traitement...) des systèmes socio-environnementaux**
- **Développer une observation nécessaire et parcimonieuse : nécessité d'un compromis entre densification de la résolution, données pertinentes et explorations scientifiques**

Introduction

La captation de données dans le domaine de l'écologie et de l'environnement, mais également en sciences humaines et sociales ou en géosciences, représente un pré-requis et un enjeu de recherche considérable. En effet, un très grand nombre d'applications et de disciplines repose sur la captation de données : la climatologie, la caractérisation et le suivi de la biodiversité et des milieux terrestres ou aquatiques dans lesquels elle évolue, la bio-géochimie de l'eau, de l'air ou encore des sols par exemple. Cette captation de données, notamment à distance, repose sur une instrumentation autonome et/ou contrôlée à distance, mais ne nécessitant plus la présence régulière des scientifiques. L'accessibilité de certains terrains d'étude, et les conditions qui y règnent, peuvent être difficiles, voire dangereuses pour les scientifiques (volcans, zones inondables ou polluées, zones arctiques…).

Cette instrumentation est à la fois un objet de recherche et un moyen pour alimenter les questions scientifiques en données pertinentes. La recherche procède en effet souvent d'observations opportunistes ou dans le cadre d'observatoires, et l'instrumentation permet de multiplier notre capacité d'observation, de quantifier des phénomènes biophysiques ou sociaux, au-delà de nos cinq sens, de couvrir des domaines souvent inaccessibles à de multiples échelles spatiales et temporelles. Les appels d'offres nationaux et internationaux mobilisés au cours de la dernière décennie, au travers des programmes d'investissements d'avenir (Equipex…) ou autres projets d'innovation technologique, en témoignent.

En écologie et environnement, trois principales méthodes de captation de données sont mobilisées : le *bio-logging*, la télédétection et l'instrumentation *in situ*. Les deux premiers reposent sur une instrumentation embarquée sur des vecteurs vivants (animaux via le *bio-logging* et la bio-télémétrie) ou technologiques (nano satellite, drone aérien ou sous-marin, ballon, *rover* ou robot…), alors que le dernier concerne une instrumentation fixe, ponctuelle et autonome permettant l'acquisition à haute fréquence de variables bio-physiques, souvent incluse au sein d'un réseau de capteurs hétérogènes (station météorologique, capteurs de qualité de l'eau ou de l'air, caméra observant une portion de l'espace ou détectant le passage d'animaux…). Une quatrième source de données, non abordée durant l'atelier, peut être considérée au travers des contributions participatives de la société à la recherche (smartphone, inventaires, questionnaires…). Cette observation « grand public » permet de démultiplier les capacités d'observation, de capter, densifier, en sus de données biophysiques, un grand nombre de « points de vue » (perception sociale, subjectivité, relations sociales, jeux d'acteurs…), mais également d'articuler les connaissances indispensables pour évaluer la capacité adaptative de la société aux changements globaux.

L'atelier « Observation à distance » des prospectives de CNRS Écologie & Environnement a regroupé 36 participants. Il s'est tenu autour de trois temps forts. Le premier a consisté en un état des lieux, fondé sur une présentation rapide de chacune des contributions soumises à l'atelier d'une part, et d'un questionnaire en ligne interactif. Le second temps consistait à travailler en sous-ateliers portant chacun sur un des trois principaux modes de captation évoqués (*bio-logging*, télédétection et instrumentation *in situ*). Le dernier temps était une restitution des travaux des sous-groupes en plénière et d'échanges avec l'ensemble des participants. La méthode a consisté à ne pas parler, dans la mesure du possible, de ses propres enjeux, mais bien d'adopter une attitude critique. L'objectif était ainsi de pouvoir récolter le matériau nécessaire pour réaliser une synthèse prospective portant sur les enjeux et les verrous actuels, et sur les pistes futures concernant l'observation à distance. Cette synthèse est structurée comme suit : la première partie résume les principales contributions et la composition de l'atelier. La seconde partie synthétise les discussions concernant les verrous, enjeux et perspectives pour le *bio-logging*, la télédétection et l'instrumentation *in situ*. La troisième partie discute ces éléments dans une perspective plus systémique.

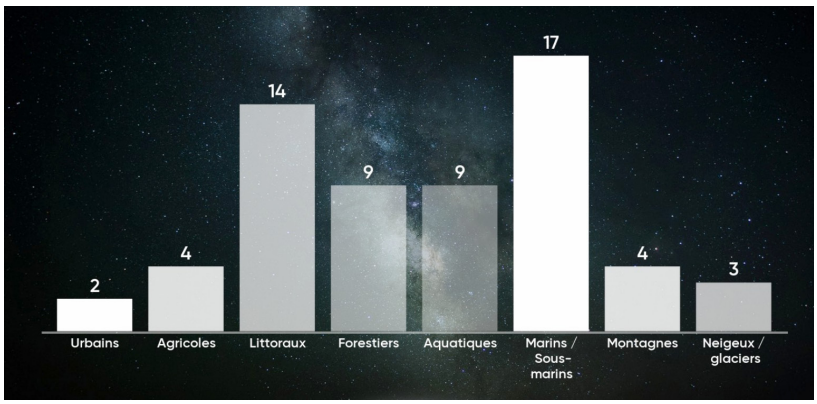


Figure 1. Répartition des biomes étudiés par les 32 répondants au questionnaire de l'atelier « Observation à distance ».

L'atelier était composé de 36 personnes, dont 32 répondants au questionnaire, parmi lesquelles 19 hommes, 12 femmes et 1 qui préfère ne pas répondre. La majorité d'entre elles sont chercheuses et chercheurs ou équivalent (18), 9 sont IT, 3 sont de jeunes chercheuses et chercheurs (Doc/post-doc), et 2 « autres ». L'ensemble des nouvelles régions étaient représentées par 1 ou plusieurs laboratoires de recherche, avec une légère surreprésentation de la partie rochelaise. L'ensemble des biomes concernés par cette observation à distance sont représentés avec une légère sous-représentation des milieux montagnards, neigeux/glaciaires, agricoles et urbains et une surreprésentation du milieu marin probablement en lien avec le grand nombre de personnels rochelais sur place (Figure 1). Les nuages de mots générés concernant les domaines d'applications et les variables suivies étant très éclatés, sans réel dénominateur commun, témoignent de l'extrême variabilité et interdisciplinarité de l'observation à distance. Enfin sur 32 participants, 16 personnes se sont déclarées comme ayant un niveau de compétence moyen à très expérimenté en instrumentation *in situ*, 12 en télédétection et 6 en *bio-logging*.

Synthèse des contributions

Au total, 13 contributions ont été reçues en amont de cet atelier (3 « *bio-logging* », 5 « télédétection », 5 « *in-situ* »). D'une manière générale, les travaux scientifiques mettent en avant la capacité des développements instrumentaux à réexplorer le fonctionnement et les trajectoires des socio-écosystèmes sous un angle nouveau, notamment les questions associées aux relations biotiques-abiotiques et aux cascades d'échelles. Les travaux de Levin (1992) ont démontré que les processus écologiques agissent à diverses échelles spatiales et temporelles, et génèrent des motifs à des échelles différentes. Dans certains cas, les motifs doivent être compris comme émergeant des comportements collectifs de grands ensembles d'unités à plus petite échelle. Dans d'autres cas, les motifs sont imposés par des contraintes à plus grande échelle. La clé de la prédiction et de la compréhension réside dans l'élucidation des mécanismes qui sous-tendent les motifs observés.

Les outils présentés permettent de raffiner les échelles de sensibilité sur des domaines explorés de plus en plus larges - dans le temps pour l'instrumentation *in situ*, ou dans l'espace pour

les observations spatialisées - et de définir la manière dont le vivant « navigue » dans ces habitats. Les outils permettent de faire émerger des structures en patchwork à petite échelle et redéfinissent le rôle fondamental de l'hétérogénéité dans le fonctionnement global des systèmes observés (ex. les habitats mobiles et éphémères associés aux fronts océaniques, ou bien les zones humides). Les capteurs permettent une auscultation quasi continue des systèmes écologiques, en économisant la présence humaine et apportant des informations critiques sur les dynamiques transitoires et l'impact des stress abiotiques sur le vivant (ex. la cognition).

Les questions associées au traitement des données sont également soulevées. Au-delà des questions de FAIRisation traitées dans un autre atelier, il s'agit de faire émerger de nouveaux outils à l'interface entre producteurs et utilisateurs, permettant d'extraire les informations pertinentes dans des quantités de plus en plus importantes de données. La numérisation des collections est un exemple significatif et pose à la fois la question de la pérennisation des spécimens numériques et de la définition d'outils pour s'immerger et identifier les informations pertinentes au sein des quantités astronomiques de données générées.

Enfin, les contributions mettent en avant l'effort de structuration transverse des communautés autour des développements de l'instrumentation environnementale *in natura*, qui émanent des chercheurs (réseau métier Drone & Cap), mais également du CNRS par la nouvelle Commission spécialisée instrumentation innovante et transverse (CSIIT) et de la Mission pour les initiatives transverses et interdisciplinaires (MITI).

Bio-logging

État des lieux et enjeux

Le *bio-logging* est une approche qui s'est énormément "démocratisée" au cours de ces deux dernières décennies. Elle a été appliquée à un nombre de taxons croissant, mais demeure essentiellement confinée aux animaux de masse corporelle suffisante pour accommoder les appareils embarqués sans que cela ne nuise à leur santé et/ou ne modifie leur comportement. Les taxons dits « invisibles » (insectes, petits poissons) et leurs interactions ne peuvent pour l'instant être appréhendés par le *bio-logging*. Dans ces conditions, la possibilité d'obtenir des jeux de données sur l'ensemble des échelons d'un écosystème, notamment les échelons intermédiaires, représente un enjeu de taille, surtout si cela est réalisé de manière dynamique (spatialement et temporellement). Il serait en effet essentiel de pouvoir étudier les interactions en temps quasi-réel entre des individus au moyen de réseaux de *bio-loggers* embarqués sur les animaux comme, par exemple, au moyen de *loggers* de proximité (Prange *et al.,* 2011). Ces réseaux pourraient intégrer en sus des vecteurs mobiles que sont les animaux, des vecteurs fixes comme les végétaux et ainsi s'intégrer dans des systèmes fixes d'instrumentation *in situ* pour obtenir une vue plus holistique des écosystèmes.

La démocratisation du *bio-logging* a aussi eu comme effet d'accroître de manière exponentielle la quantité et le type de données que les *bio-loggers* délivrent. L'enjeu à ce niveau est de trouver les sites de dépôts qui permettraient d'accueillir cette masse de données afin qu'elle soit mise à disposition en libre accès et à large échelle. Si plusieurs initiatives existent pour les données de localisation (ex. GBIF*, OBIS* ou *Movebank*), il n'y a que peu de sites qui permettent d'accueillir et de gérer les données dites « lourdes » telles que les données d'accélérométrie, d'acoustique ou encore les données issues de *video-loggers*.

* *Glossaire*

Verrous

Certaines espèces sont de taille encore trop réduite pour être équipées avec des *bio-loggers*, d'autres ont des modes de vie trop cryptiques pour être capturées plusieurs fois et les données embarquées deviennent ainsi irrécupérables. Ces difficultés conduisent à une hétérogénéisation forte de notre capacité à étudier certains taxons par rapport à d'autres selon une approche individu-centrée. Elles expliquent également la quasi-impossibilité d'appréhender le fonctionnement des écosystèmes via le *bio-logging* puisque nous ne pouvons examiner chaque niveau trophique selon la même méthodologie, et encore moins investiguer les interactions et les flux, par exemple d'énergie, entre individus sur l'ensemble (ou *a minima* les principaux) des taxons constituant les chaînes alimentaires. D'un point de vue économique, il existe encore peu de compagnies produisant des outils du *bio-logging* en France. La difficulté principale réside dans la transition entre la création d'appareils en R&D dans les laboratoires à une production en masse dans une start-up. Malgré des procédures mises en place aux niveaux institutionnels pour assister les laboratoires dans la démarche de création de start-ups, le processus a du mal à s'installer dans le cas du *bio-logging*. Le marché est-il trop spécifique ou les développeurs du *bio-logging* (souvent des biologistes ou personnes avec une fibre biologiste) sont-ils moins motivés à évoluer vers une professionnalisation de leurs métiers ?

Solutions et enjeux futurs

L'un des premiers chantiers pour pousser la démocratisation du *bio-logging* et espérer répondre aux questions susmentionnées serait d'arriver à une forme de standardisation des *bio-loggers*, c'est-à-dire à une harmonisation technologique avec la définition de format et de types de mesures qui pourraient être partagés entre différentes équipes. Cette standardisation des appareils iraient ainsi de pair avec une standardisation des données qui impliquerait nécessairement une standardisation dans leurs modes de stockage et surtout de partage. Pratiquement, il faudrait donc faciliter, ou en tout cas, mieux utiliser et catalyser les réseaux métiers existants afin que les communautés utilisatrices du *bio-logging* - mais aussi des autres systèmes d'instrumentation - échangent plus/mieux. Le rôle des réseaux métiers serait de proposer des formations à une bonne utilisation du *bio-logging* (respect de l'animal utilisé comme objet d'études ou « auxiliaires de la recherche », compréhension des données et de leurs valeurs intrinsèques, compréhension aussi des limitations inhérentes aux systèmes de mesure…).

À titre d'exemples, le groupe « *Drones & Cap* » ou la Cellule faune sauvage de CNRS Écologie & Environnement pourraient servir de lieux pour mettre en œuvre de telles formations. Alternativement, il pourrait être intéressant de créer une branche nationale de la société internationale du *bio-logging* (https://www.bio-logging.net/).

Outre une meilleure communication entre utilisateurs, de telles plateformes pourraient se voir confier le rôle de « médiateurs » entre les scientifiques et le grand public. En effet, un des enjeux majeurs autour du *bio-logging* consiste à mieux sensibiliser la société sur nos pratiques. Il faut en effet souvent démontrer et convaincre le grand public que nos activités scientifiques ne se font pas au détriment de la faune étudiée. En *bio-logging*, l'utilisation des animaux à des fins scientifiques a très souvent un aspect bénéfique pour la conservation de l'animal, même si de nombreuses études n'ont pas nécessairement cette finalité et reste des études purement fondamentales. Il faut ainsi mettre en avant les études visant à diminuer l'impact du *bio-logging* sur les animaux et éviter de colporter l'image de « cowboy du *bio-logging* qui tague tout ce qui bouge ». Une meilleure communication des publications d'impact et de solutions potentielles pour diminuer cet impact, ainsi qu'un meilleur partage des bonnes pratiques devraient conduire à une meilleure compréhension par le public de nos approches.

Si la standardisation des données et capteurs du *bio-logging* est souhaitée et souhaitable, l'exploration, la découverte, les essais de nouveaux types de *bio-loggers*, mesurant de nouveaux paramètres avec de nouvelles combinaisons - en un mot l'innovation - reste vitale. Sans innovation dans les outils de mesure, la recherche se

sclérose car, très vite, les outils ne seront plus capables d'aider les chercheurs à répondre aux nouvelles questions qui s'imposent à eux.

Le Graal est évidemment d'arriver à une génération de *bio-loggers* qui n'aient aucun ou quasiment aucun impact sur le sujet qui le porte. Il est question ici de taille, de profils de l'appareil mais aussi de systèmes de fixation sur ou dans le corps de l'individu, en gardant à l'esprit que la capture pour équipement/déséquipement entraîne également un impact et qu'il faudra donc travailler sur de nouvelles méthodologies pour que les approches de capture et les temps de contention soient optimums. Ces *bio-loggers* optimisés pourraient se concentrer sur la mesure de paramètres physiologiques « fins et complexes », c'est-à-dire des paramètres qui traduisent une réponse rapide, pertinente, et intégrative de l'organisme à des *stimuli* environnementaux qui peuvent être des paramètres physiques, chimiques, ou biologiques. La capacité à transmettre en temps réel la mesure d'un tel changement survenant au niveau de la physiologie de l'individu ouvrirait de nombreuses avenues de recherche. Imaginons ainsi une mesure sanguine du taux de corticostérone, analysée de manière dynamique, qui informerait les rangers d'un parc de l'arrivée d'un braconnier ; ou bien de capteurs de concentration en polluants chimiques pour mesurer l'exposition des individus à la pollution en temps réel. En poussant plus loin, ces *bio-loggers* pourraient même mesurer l'ADN environnemental et ainsi effectuer des échantillonnages biologiques du milieu. L'apparition de la technologie MinION (https://nanoporetech.com/products/minion) laisse présager de la réalisation de tels rêves dans un futur relativement proche. Ainsi, nous pouvons envisager être en capacité de mesurer des paramètres génotypiques et phénotypiques en lien avec les conditions environnementales et à fine résolution spatiale depuis des *bio-loggers*.

En généralisant l'utilisation des *bio-loggers*, en multipliant leurs déploiements sur différents taxons au sein d'un écosystème, il deviendrait possible de mesurer, et éventuellement gérer en temps réel, le fonctionnement des écosystèmes, surtout si les approches *bio-logging* sont couplées avec celles évoqués dans ce document (télédétection et mesures *in situ*). En parallèle de la domotique apparaît alors l'« Animalotique » !

Télédétection

État des lieux et enjeux

La télédétection bénéficie de plus de 30 ans d'avancées dans le domaine satellitaire pour proposer aujourd'hui des constellations de satellites d'observation de la Terre, une grande variété de données (dans les domaines passif et actif), ayant déjà soulevé les enjeux du *Big Data* depuis plusieurs années. Plus récemment, l'essor des drones, notamment aériens, via des offres commerciales standardisées, a permis de revisiter certains challenges liés à la détection de petites portions de surfaces terrestres grâce aux résolutions centimétriques offertes par ces vecteurs aériens. Cet essor ne s'est pas limité au domaine aérien, mais également aux domaines maritime, sous-marin, souterrain… Si l'un des principaux enjeux du satellitaire est aujourd'hui de fournir de produits finis, aux méthodes éprouvées et démocratisées, à destination de la recherche et de l'action publique, la télédétection par drone fait encore face à un certain nombre d'enjeux. Par exemple, force est de constater que la démocratisation rapide des drones a conduit à une pratique dominante reposant sur l'utilisation d'offres « clés en main », avec une confiance aveugle dans les capteurs. Aujourd'hui, de nombreuses études scientifiques en écologie et environnement, ou en géosciences, remettent en cause la qualité des mesures et appellent à une évaluation des capteurs. De plus, la montée en compétence des utilisateurs en sciences amène à une recherche de nouveaux capteurs et/ou modes de captation.

Dans le même temps, la grande flexibilité d'acquisition offerte par les drones amène à deux évolutions majeures des pratiques en télédétection. Tout d'abord, on constate un changement de paradigme passant d'une exploitation et adaptation des méthodes aux images satellitaires disponibles (date et heure fixes d'acquisition) vers une adaptation des protocoles d'acquisition (horaire, jour/nuit, répétitivité temporelle) pour détecter et caractériser les processus (socio)écologiques (usages…) ou biogéochimiques. Ensuite, cela engendre une évolution des données produites, tant dans leur dimension horizontale (résolution spatiale…) que verticale (profils atmosphériques…), permettant de passer d'un état de surface (une occupation du sol, détection d'espèces), à des paramètres biophysiques (taux de chlorophylle…) ou encore à des traits fonctionnels (hauteur, densité foliaire), qui permettent de définir des indicateurs clefs comme les EBV (*Essential Biodiversity Variables* - Pereira *et al.,* 2013).

Aujourd'hui, le CNRS est la plus grosse société de drones aériens en France, avec 160 télépilotes approuvés dont 134 opérationnels, 223 drones répertoriés dont 180 opérationnels, et près de 1110 heures de vol entre octobre 2021 et octobre 2022. L'évolution inhérente des types de données produites par drone et de leurs caractéristiques (résolutions spatiales, temporelle, spectrale…) et la volumétrie générée soulèvent de nouveaux défis de visibilisation de celles-ci (FAIRisation), mais également des ressources disponibles et mobilisables. Enfin, les drones soulèvent l'hypothèse forte, mais pas encore totalement démontrée, qu'ils constituent le trait d'union entre une mesure ponctuelle et locale et une donnée spatiale exhaustive (plus ou moins résolue/étendue), l'échelon intermédiaire indispensable pour faire le lien entre les relevés *in situ* et le satellitaire (Alvarez-Vanhard *et al.,* 2020).

Verrous

La structuration actuelle du CNRS ne permet malheureusement pas d'avoir une vision holistique des forces et moyens en présence, dont les demandes de moyens relèvent d'initiatives trop locales (redondances de matériels à l'échelle d'UMRs, entretien non pérenne, dispersion des ressources). Au-delà de l'effet de mode des drones aériens, il y a une faible prise en compte de la pérennisation des matériels et des compétences malgré un accompagnement réglementaire efficace et proactif de la DIRSU* Drone. Par ailleurs, si le domaine aérien est désormais plutôt bien circonscrit réglementairement (voire trop), ce n'est pas encore le cas pour les autres domaines (marin, sous-marin…).

Par ailleurs, à notre connaissance, il n'existe aucune solution permettant une centralisation d'une telle volumétrie de données *raster*, une mise en commun (interopérabilité) des données déjà acquises afin d'éviter la revisite de certains sites d'étude par exemple. L'enjeu ne doit pas relever d'un institut ou d'un autre et doit pouvoir s'appuyer sur des structures locales en cohérence avec des plateformes produisant et fournissant de la donnée pour des collectifs scientifiques locaux ou régionaux (type DIPEE*/OSU/MSH).

D'autres part, les chercheurs en écologie et environnement sont généralement des utilisateurs finaux des matériels proposés, des « thématiciens » qui n'ont peu ou prou d'interactions avec d'autres communautés largement impliquées dans les développements en robotique ou capteurs, ou encore dans la définition des programmes spatiaux. Si des développements semblent évidents, l'interdisciplinarité nécessaire au développement de nouvelles données ou à la mise en cohérence des données drone/satellitaire ou drone/*in situ*, n'est pas assez développée.

Perspectives et recommandations

Trois perspectives/recommandations majeures ressortent des discussions qui ont eu lieu dans ce sous-atelier en lien avec la structuration, les développements scientifiques et l'interdisciplinarité. Concernant la structuration, il semble important de s'appuyer sur des réseaux émergeants (MITI

* *Glossaire*

Drones & Cap, capteurs en environnement…) pour recenser et faire connaître les matériels, les compétences existantes localement, les domaines d'applications actuels et possibles. Cela nécessite de « sortir les drones des instituts », tout en cherchant à inciter à une structuration de pôles régionaux interdisciplinaires/inter-instituts (DIPEE/OSU, MSH). Cela devrait permettre la mise en commun de moyens/gros équipements (drone/capteurs, moyens de calcul…), en phase avec les développements nécessaires pour la mise à disposition des données et de favoriser l'innovation.

Concernant les développements scientifiques et l'interdisciplinarité, trois pistes semblent essentielles :

- inciter à la proactivité dans le domaine du développements de drones et de capteurs, notamment à travers la définition de cahiers des charges à destination de CNRS Sciences informatiques ou de CNRS Nucléaire & Particules qui aient du sens pour les applications en écologie et environnement (mise en cohérence des mesures locales, expérimentales) et qui anticipent de nouveaux besoins (continuum de mesures dans la Zone Critique - terrestre <–> marin par exemple, développement de nouveaux capteurs radar/SWIR*/hyperspectraux, suivi du carbone atmosphérique ou dans les sols…) ;
- inciter à poursuivre le changement de paradigme offert par ces données innovantes : passer de la statistique descriptive (où la donnée sert à fournir des variables explicatives vis-à-vis d'une variable écologique à expliquer) à une modélisation des flux (gènes, faune, air, eau…) grâce à la précision et structure 3D par exemple des données ;
- inciter au développement de méthodes permettant de mettre en cohérence les mesures au sol, par drone et satellitaires (IA, adaptation de domaine…) pour réussir enfin à relever le défi du changement d'échelle.

In situ

État des lieux et enjeux

Les réseaux de capteurs *in situ* s'entendent aujourd'hui comme des outils de surveillance automatisés, en continu, d'un domaine défini. Ils se sont développés pour enrichir les données acquises par des observateurs humains et systématiser des méthodologies d'observation objectivées, automatiser des tâches répétitives et/ou dangereuses, ou bien encore alerter sur des conditions spécifiques appelant à une action humaine. La météorologie a été pionnière dans ce domaine avec le déploiement des premières stations automatiques dès 1940, qui transmettaient déjà les observations par radio à un centre de données.

Tansley (1935) et Lindeman (1942) ont défini les écosystèmes comme le système indivisible de biota et de leur environnement, où les cycles organiques et inorganiques sont indissociables, s'organisant autour de flux de matières et d'énergie. Les capteurs se focalisent généralement sur la captation de ces variables abiotiques, offrant des points de vue sur les interactions complexes entre le vivant et son environnement physique, qui régissent le fonctionnement des socio-écosystèmes. Ces outils techniques permettent également d'étendre nos sens et de faire émerger des contrôles invisibles (radon, polluants) ou inaccessibles (souterrains, volcans), et peuvent également, à certains égards, favoriser la reconnexion entre l'humain et la nature (Litleskare *et al.*, 2020).

L'implémentation des réseaux de capteurs – et donc la stratégie d'observation – est étroitement liée aux objectifs scientifiques des études. Par exemple, le croisement de plusieurs types de données sur des gradients permet de documenter des habitats, et amène ainsi à des éléments sur la définition des interactions biotiques-abiotiques et aux questions d'adaptation (Melero *et al.*, 2022). Les enjeux associés aux changements globaux appellent à des suivis holistiques sur des temps longs pour suivre les trajectoires

* *Glossaire*

des socio-écosystèmes, identifier les cascades d'impact, points de bascules (Ragueneau *et al.*, 2018). Les systèmes de suivis *in situ* doivent ainsi être capables de couvrir des domaines d'étendues variables pour répondre à ces questions scientifiques : du très local (micro-météorologie), à l'échelle continentale (ex. le projet A20 https://acousticobservatory.org/home_1/), en passant par les systèmes expérimentaux (ex. mésocosmes) et les observatoires, sur une très grande diversité de variables.

La portée des dispositifs techniques de captation, lorsqu'ils sont déployés au sein des dispositifs d'observation dans les territoires (Bretagnolle *et al.*, 2018, Gaillardet *et al.*, 2018), s'élargit pour favoriser la collaboration entre disciplines, mais également pour permettre l'articulation des connaissances entre les sphères académiques et non académiques (gestionnaires, décideurs). Ainsi, les variables essentielles, qu'elles soient climatiques (https://public.wmo.int/en/programmes/global-climate-observing-system/essential-climate-variables) ou de biodiversité (https://geobon.org/ebvs/what-are-ebvs/) restent l'expression d'un consensus communautaire pour définir des preuves empiriques nécessaires pour documenter, comprendre et gérer l'évolution du climat et de la biodiversité. Cela dit, la pertinence d'une observation « essentielle » est contrebalancée par une contrainte de faisabilité de la mesure, c'est-à-dire réalisable à l'aide de méthodes éprouvées *in situ*, économiquement viable et durables dans le temps. Ces notions soulignent le rôle transformateur de facteurs technologiques - et de leur appropriation - dans l'évolution de la recherche et des objectifs de suivi et d'évolution des socio-écosystèmes.

Verrous

Le principal jalon des données acquises localement par de l'instrumentation *in situ* concerne la représentativité spatiale de la mesure (micro ou méso échelle). Celui-ci a été levé à travers la mise en place de réseaux de mesures. Les enjeux d'observation à distance par des réseaux de capteur *in situ* couvrent l'ensemble des étapes de la conception du capteur, de la définition d'une stratégie d'observation, du déploiement *in situ*, jusqu'à la FAIRisation des données générées. Les verrous peuvent ainsi être classés selon des critères technologiques, techniques, de stratégie d'observation et de définition du contenu informatif.

Les systèmes de mesure à distance permettent de limiter la présence humaine sur les sites (notamment dans les zones à risques) et d'accompagner les tâches répétitives (telles que le téléchargement des données ou le remplacement de batterie). Ces contraintes mettent en avant un double verrou d'autonomie, c'est-à-dire de fonctionnement sur des sources d'énergie intermittentes et de télétransmission des données. Ce verrou répond également aux enjeux de réduction de l'empreinte carbone. La démultiplication des types de capteurs et la densification des systèmes de mesure *in situ* pose la question de la maintenabilité des outils et de leur suivi. Si la transmission des données favorise la maintenance préventive, la présence sur site reste nécessaire pour assurer le bon fonctionnement, remplacer les capteurs et les étalonner le cas échéant.

La crédibilité de la donnée acquise impose une qualification des capteurs et des protocoles de mesure. Les nouvelles technologies et le développement du numérique ont transformé la collecte des données, avec une diversification importante des types de capteurs (notamment bas coût) et d'acquisition des données (sciences participatives). Ces nouveaux outils et nouvelles pratiques ont fait émerger de nouvelles questions relatives à la confiance dans les données, mais également de nouvelles questions légales, éthiques et d'acceptabilité sociale dans l'usage des capteurs et des données.

Les données doivent être accompagnées de métadonnées qui décrivent les conditions, matériels et protocoles d'acquisitions. En termes métrologiques, toute donnée doit être accompagnée d'une estimation des incertitudes (qualité d'un résultat de mesure par rapport à la réalité), *a minima* d'une erreur (représentation de la différence entre une valeur mesurée d'une grandeur et une valeur de référence). Cette étape est cruciale puisque les incertitudes donnent une valeur à la donnée, son apport informatif, qui est à comparer directement à notre capacité d'expliquer les données (par exemple par des modèles).

À titre d'exemple, le développement récent des méthodes d'analyse du bruit sismique a permis de faire un bon fondamental en sciences de la Terre (Lecocq *et al.*, 2017), et cela a été possible grâce à une bancarisation systématique de toutes les données enregistrées par les sismographes, y compris en dehors des périodes d'activité sismiques. Ainsi, des données perçues aujourd'hui comme bruit, c'est-à-dire non expliquées, pourraient demain devenir une information à la lumière de nouveaux développements.

Les capteurs nous offrent d'étendre nos sens et d'être opérationnels sur site en continu. Il se pose ainsi la question de l'immersion dans le système observé. Le contrôle à distance implique que le système de mesure autorise le déclenchement d'actions plus élaborées. Il s'agit de générer des alertes appelant à une opération humaine, ou bien de faire évoluer le système d'observation lui-même lorsque les conditions l'imposent (conditions extrêmes ou singulières) : augmenter les cadences de prise de données, activer un outil spécifique (ex. échantillonner de l'eau) qui permettra d'enrichir les observations, sur le volet biotique notamment.
La gestion des flux de données générées reste également un verrou important qui pose la question de la pertinence des données générées et de leur contenu informatif. Ainsi, l'exemple de capteurs passifs du type pièges photographiques est démonstratif. Le capteur génère des flux de données importants, alors que l'information pertinente reste l'occurrence d'un individu d'une certaine espèce. Les verrous sont ainsi doubles : à l'échelle du capteur, il semble nécessaire d'aller vers des capteurs qui embarquent une certaine intelligence synthétisant les données sous la forme d'une information directement utilisable. À l'échelle du réseau de capteurs, il s'agit de développer des indicateurs pour des usages spécifiques qui intègrent les données générées par un ensemble de capteurs. Par exemple, un système d'alerte d'un évènement de salinisation des écosystèmes côtiers lors d'une surcote, nécessite d'intégrer à la fois une mesure de l'altitude relative entre le niveau marin et le domaine continental, mais également l'évolution de la conductivité de l'eau dans les eaux de surface et les eaux profondes. Encore une fois, un traitement des données au plus proche des capteurs permet de donner du sens à celles-ci. Ces systèmes d'intelligence déportée sont également cohérents avec les enjeux de sobriété énergétique et numérique.

Perspectives et recommandations

Les verrous identifiés mettent au centre des préoccupations une question de cohérence entre l'existence de produits/solutions techniques adaptables, cohérents avec les contraintes *in situ*, mais qui permettent également d'enrichir les approches scientifiques par de nouvelles données, plus précises, plus complètes, pour répondre aux besoins du développement des approches systémiques. Il s'agit ainsi de fédérer des communautés d'une manière très concrète et favoriser les interactions entre technologues, personnes de terrains, et chercheurs thématiques. Cette structuration transverse peut être opérationnalisée de deux manières :

- d'une part autour de réseaux métiers, tels que le Réseau Technologique sur les Capteurs en Environnement (RTCE*, https://www.reseau-capteurs.cnrs.fr/) qui rassemble déjà une large communauté autour de la métrologie *in natura* ;
- d'autre part des objets à observer (observatoires ou sites expérimentaux) au sein des infrastructures de recherche.

Le marché des capteurs « environnementaux » est en plein essor, et s'articule autour de nouvelles opportunités technologiques, notamment autour de capteurs bas coût. Au-delà de l'effet d'aubaine, l'enjeu est de nous donner la capacité de définir la portée de ces innovations et de juger de leur pertinence et de leur acceptabilité. Il est également important de réfléchir à des outils permettant :

- d'aller vers une homogénéisation du parc instrumental national et de favoriser le partage des expertises ;
- de renforcer l'interopérabilité des systèmes existants, notamment dans un cadre de transmission automatique des données. L'instrumentation frugale consiste à associer des composants élémentaires, génériques ou spécifiques, autour d'un cahier des charges précis sur un cas d'usage, tenant compte en amont des contraintes de terrains pour maximiser leur pertinence et durée de vie. L'intérêt pour la communauté doit également se concrétiser

* *Glossaire*

par de nouveaux outils de cartographie et de partage de ressources documentaires, solutions techniques et savoir-faire. Ces développements, associant pleinement technologues et chercheurs au sein des réseaux métiers, des projets et des observatoires, ont le double intérêt (1) de favoriser l'agilité dans le développement et la reproduction des systèmes opérationnels et (2) de construire et fédérer des communautés d'utilisateurs pour que les outils répondent toujours mieux aux besoins exprimés.

L'observation à distance intègre des questions fondamentales sur les données produites par les capteurs et leur utilisation. Si les enjeux de FAIR-isation sont pleinement intégrés et bien décrits dans le volet « données », l'enjeu de la gestion des flux de données en continu générées par les systèmes *in situ* reste une préoccupation importante pour faire émerger le contenu informatif des données produites pour générer des alertes et appelant à une réponse instrumentale ou humaine. La question soulevée s'articule autour des relations données – modèles, l'information la plus pertinente étant non prévisible, ou défiant l'attendu. L'intelligence artificielle est un des outils pouvant répondre à certains de ces enjeux, il s'agira à nouveau de favoriser les interactions avec des communautés numériques pour alimenter les systèmes de traitement embarqués.

Discussion et perspectives

Synthèse

Cette synthèse présente, pour chacune des modalités de captation de données à distance sur lesquelles les chercheurs ont une influence directe (exception faite des données collectées au travers des sciences participatives), l'état des lieux, les principaux verrous et enjeux à court et moyen termes.

Que les verrous soient éthiques, réglementaires, technologiques, structurels et organisationnels, ou encore financiers, des solutions existent et les développements tendanciels laissent présager une levée plus ou moins rapide de ceux-ci. Pour cela, l'innovation technologique est un prérequis. Néanmoins, l'innovation sociale l'est tout autant et mérite une vigilance particulière : que ce soit en interne à la communauté scientifique (mise à disposition des informations, éviter les guerres de clochés, pratique de non-partage immédiat par les chercheurs encore terriblement présente…) ou vis-à-vis de la société civile (acceptabilité et perception des modalités d'acquisition), un accompagnement est indispensable pour favoriser et faciliter les changements de paradigmes évoqués.

L'observation à distance pour relever le défi du changement d'échelle ?

Ces nouvelles données et leur densité spatiale ou/et temporelle permettent d'envisager le passage d'une caractérisation d'une propriété biophysique, géochimique ou sociale à la caractérisation de processus socio-écologiques, physiques… Par ailleurs, la complétude de ces différentes modalités d'observation à distance permet d'envisager de mettre en relation des données ponctuelles pouvant présenter des résolutions spatiales ou temporelles plus fines avec des données 2D/3D, couvrant des étendues plus vastes, mais également de mieux appréhender les *continuums* (terre-mer par exemple) jusque-là forts cloisonnés. Ainsi, il devient possible d'expliciter finement des données acquises par télédétection satellitaire ou drone (état, dynamique) à l'aide de données ponctuelles (*bio-logging*, *in situ*) et inversement, de pouvoir extrapoler des informations ponctuelles sur la base de leur similitude/base d'apprentissage colocalisée avec des données *raster*. Les mêmes principes sont envisageables entre des données de télédétection présentant des résolutions spatiales et/ou temporelles différentes (drones/satellites) (Alvarez-Vanhard *et al.*, 2021). *In fine*, nous pourrons espérer relever le vieux défi du changement d'échelle dans le suivi, l'analyse et la compréhension des processus biotiques/abiotiques/anthropiques et de leurs interactions.

De la nécessaire implication de la société

Dans cet atelier, les données dites sociales ont été peu abordées. Deux types de données peuvent se distinguer : celles relatives à la participation de la société à la captation de données pour densifier la collecte de données et celles relevant de la préférence, de la perception, des pratiques (déplacements, traces numériques) et des choix des individus et de la société. La non prise en compte de ce type d'observation à distance ici s'explique probablement parce que le monitoring et la captation de ce type de données relève, selon certains et peut-être de façon stéréotypée, plus de CNRS Sciences humaines & sociales que de CNRS Écologie & Environnement d'une part ; et plus certainement par les compétences requises (non représentées par les porteurs de cet atelier). La réglementation contraignante (RGPD) rend ces données plus rares malgré la démocratisation récente des smartphones et des sciences participatives. Néanmoins, il paraît de plus en plus évident que ces données vont devenir essentielles pour mesurer la propension de la société à accepter des mesures politiques et réglementaires visant par exemple l'atténuation ou l'adaptation au changement global, une transition sociale, énergétique et environnementale. Ceci renforce clairement le besoin de développer des recherches socio-environnementales, où la donnée sociale et humaine et les enjeux de captations inhérents soient priorisés.

Observation à distance sur le long terme : science ouverte, résolutions et sobriété

Volontairement, l'accès à la donnée acquise à distance n'a pas été abordé dans cet atelier puisqu'un atelier traitait spécifiquement de cette question. Néanmoins, compte tenu des enjeux de mise en relation des données issues des diverses modalités d'acquisition, il est fondamental de rendre ces données accessibles ouvertement, c'est-à-dire de veiller à leur FAIRisation. Pour ce faire, il est nécessaire de promouvoir la mise en relation des bases de données FAIR (GBIF, Data Terra, et autres Infrastructures de Données Spatiales) et de mettre en place une politique de la données cohérentes à l'échelle des divers instituts du CNRS. La mise à disposition est un préalable pour le suivi à long terme en écologie et environnement ainsi qu'à la valorisation des données. S'il y a un enjeu structurel fort (en ressources humaines, logicielles et matérielles), il est également fondamental d'agir sur les pratiques des scientifiques, encore archaïques parfois (« *ce sont mes données* », « *je ne le mettrai à disposition qu'une fois l'article publié* »…), au travers de méthodes incitatives voire pourquoi pas coercitives.

De plus, la question de la volumétrie des données qui sera générée par une démultiplication des capteurs et modalités de captation est clairement soulevée ici, et renvoie notamment à l'atelier « Données, après l'acquisition ». Si la pertinence de la donnée (et leurs résolutions spatiale et temporelle) dépend avant tout de la question scientifique traitée, il semble que les résolutions les plus fines (dans la mesure du possible) doivent être envisagées. Elles peuvent être sélectionnées ou dégradées si nécessaire pour favoriser une certaine sobriété « scientifique ». Cela permettrait qu'elles puissent être revisitées ultérieurement par de nouvelles méthodes afin de détecter des processus/signaux faibles que les méthodes actuelles ne permettent pas.

Enfin, le capteur reste un moyen d'acquérir et d'articuler des connaissances. Les données acquises peuvent être considérées comme stratégiques pour documenter le fonctionnement et la trajectoire des socio-écosystèmes, notamment dans les observatoires, au sein des territoires. D'autre part, les besoins en matériaux et énergie pour construire, faire fonctionner et traiter les données ne sont pas neutres. La question de la place de l'instrumentation dans la recherche actuelle et dans les transitions socio-environnementales reste ouverte, mais nous oriente vers le nécessaire développement d'une instrumentation frugale, c'est-à-dire de systèmes d'observation adaptés aux conditions de terrain, ayant une longue durée de vie, capables de « grappiller » de l'énergie dans le milieu naturel, et générant des flux de données gérables.

Positionnement et enjeu à l'échelle européenne : vers une profonde innovation ?

Dans le contexte actuel de structuration européenne des infrastructures de recherche (IR) au travers du projet eLTER (H2020 PPP* et PLUS), la place de l'instrumentation d'acquisition de données à long terme est centrale et constitue même un critère d'éligibilité des sites scientifiques de suivi à long terme des systèmes socio-écologiques de la zone critique. Dès lors, bien que cela se place dans une perspective de compétitivité interne à la recherche européenne, l'observation à distance et les données qu'elle produit devient un enjeu de plus en plus important, bouscule les pratiques, questionne l'organisation même du CNRS mais finalement devrait favoriser une innovation sociale et politique de la recherche française à l'aulne des enjeux actuels et de plus en plus prégnants des changements globaux.

RÉFÉRENCES

- Alvarez-Vanhard, E.G., Houet, T., Mony, C., Lecoq, L., Corpetti, T. (2020) Can UAVs fill the gap between in situ surveys and satellites for habitat mapping?, Remote Sensing of Environment. 243, 12p, https://doi.org/10.1016/j.rse.2020.111780
- Alvarez-Vanhard, E.G., Houet, T., Corpetti, T. (2021) UAV & Satellite synergies for optical remote sensing applications: a review. Science of Remote Sensing, vol. 3, 14 p. https://doi.org/10.1016/j.srs.2021.100019
- Bretagnolle, V., Berthet, E., Gross, N., Gauffre, B., Plumejeaud, C., Houte, S., et al. (2018). Towards sustainable and multifunctional agriculture in farmland landscapes: lessons from the integrative approach of a French LTSER platform. Science of the Total Environment, 627, 822-834.
- Gaillardet, J., Braud, I., Hankard, F., Anquetin, S., Bour, O., Dorfliger, N., et al. (2018). OZCAR: The French network of critical zone observatories. Vadose Zone Journal, 17(1), 1-24.
- Lecocq, T., Longuevergne, L., Pedersen, H. A., Brenguier, F., Stammler, K. (2017). Monitoring ground water storage at mesoscale using seismic noise: 30 years of continuous observation and thermo-elastic and hydrological modeling. Scientific reports, 7(1), 1-16.
- Lindeman, RL. (1942). The trophic-dynamic aspect of ecology. Ecology, 23: 399–417.
- Litleskare, S.E., MacIntyre, T., Calogiuri, G. (2020). Enable, reconnect and augment: a new ERA of virtual nature research and application. International Journal of Environmental Research and Public Health, 17(5), 1738.
- Melero, Y., Evans, L.C., Kuussaari, M., Schmucki, R., Stefanescu, C., Roy, D. B., et al. H. (2022). Local adaptation to climate anomalies relates to species phylogeny. Communications Biology, 5(1), 143.
- Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, et al. (2013). Essential biodiversity variables. Science 339, 277–278. https://doi.org/10.1126/science.1229931
- Prange, S., Gehrt, S. D., Hauver. S. (2011). Frequency and duration of contacts between free-ranging raccoons: uncovering a hidden social system. Journal of Mammalogy, 92, 1331–1342.
- Ragueneau, O., Raimonet, M., Maze, C., Coston-Guarini, J., Chauvaud, L., Danto, A. et al. (2018). The impossible sustainability of the Bay of Brest? Fifty years of ecosystem changes, interdisciplinary knowledge construction and key questions at the science-policy-community interface. Frontiers in Marine Science, 5, 124.
- Tansley, AG., (1935). The use and abuse of vegetational concepts and terms. Ecology, 16, 284–307.

* *Glossaire*

Données, après l'acquisition

Auteurs : Arnaud Elger (LEFE), Émilie Lerigoleur (GEODE), Bruno Mansoux (BBEES), Alain Queffelec (PACEA)

3 PRIORITÉS SCIENTIFIQUES À ABORDER D'ICI 2030

- **Co-construire avec les divers acteurs une stratégie de gestion des données aux différentes étapes de leur cycle de vie, dans une perspective d'efficacité et de sobriété numérique**
- **Mettre en place une politique de formation des personnels à la démarche de planification, de gestion et de diffusion des données FAIR - prioritairement ciblée vers les correspondants OpenDoRES des unités**
- **Développer et permettre l'appropriation par la communauté de nouveaux outils de traitement collaboratif des données (ex. lacs de données, environnements virtuels de recherche)**

État des lieux

Contexte et problématique

Selon l'OCDE* (2007), les données de la recherche sont définies comme des « enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche ».

Les scientifiques produisent des données dès qu'ils réalisent une observation ou prélèvent un échantillon sur le terrain ou en conditions contrôlées. Les données sont également de plus en plus souvent issues de systèmes automatisés (ex. capteurs connectés), générées par des simulations numériques ou produites par les citoyens au travers des « sciences participatives » ou des « observations opportunistes ». L'ensemble de ces données, ainsi que diverses informations renseignant le contexte dans lequel elles ont été acquises (métadonnées), sont stockées dans des bases de données et/ou disponibles sous forme de fichiers.

Dans le cadre du Plan national pour la science ouverte (PNSO 2018, 2021) et des diverses feuilles de route institutionnelles, les données (au moins celles vouées à la diffusion) doivent être facilement trouvables, accessibles, interopérables et réutilisables selon les principes FAIR* (Wilkinson *et al.*, 2016). Elles doivent pouvoir être identifiées de façon uniques et pérenne en leur associant des PID*, afin de permettre leur suivi dans le cadre d'une utilisation ultérieure. Il faut également qu'elles soient décrites selon des formats standardisés (ex. INSPIRE, *Dublin core*, *Darwin core*…) et des thésaurus*/ontologies* partagés, et stockées dans des formats de fichier ouverts (ex. CSV, TXT, XML, JSON) garantissant leur réutilisation hors du contexte dans lequel elles ont été acquises, et ce pour une durée pouvant théoriquement aller jusqu'à plusieurs décennies.

Les données sont au cœur d'un cycle qui, de leur acquisition à leur réutilisation, passent par des étapes clés comme la documentation, la conservation et l'exposition. Ce cycle de vie implique par ailleurs diverses compétences techniques liées notamment à la curation des données, encore insuffisamment (re)connues au sein des communautés scientifiques : sécurisation, formatage, accessibilité, pérennité et écoresponsabilité.

De nombreux outils et services sont mis à disposition de la communauté scientifique pour faciliter le flux d'informations entre chaque étape de ce cycle de vie. Les lacs de données et les environnements virtuels de recherche (EVR*) font partie des solutions d'avenir pour une gestion intégrée et interopérable des données qu'il reste à explorer.

Comment la communauté scientifique et technique de CNRS Écologie & Environnement s'approprie-t-elle les nouvelles pratiques de gestion FAIR (autant que possible) des données dans le contexte de la science ouverte ? Comment les scientifiques pratiquent-ils cette gestion au travers des plans de gestion de données* ? Quels sont les entrepôts de données* ou les infrastructures de recherche* (IR) où ils exposent et diffusent leurs données ? En quoi les pratiques des autres Instituts du CNRS peuvent nous conduire à faire évoluer les nôtres ?

Un paysage d'infrastructures et d'entrepôts en pleine évolution

Il existe une multitude d'infrastructures et d'entrepôts internationaux et nationaux dont une partie est identifiée dans la feuille de route nationale des IR du Ministère de l'Enseignement supérieur et de la Recherche (Figure 1).

Certaines offrent le stockage et la sauvegarde à moyen terme des jeux de données grâce au service d'entrepôt, d'autres permettent le catalogage de données, soit stockées physiquement dans l'entrepôt adjacent soit dans des entrepôts distants. Un catalogue permet en effet d'exposer les jeux de données à travers leurs métadonnées et, par système de moissonnage, les catalogues sont interconnectés pour donner une plus grande visibilité aux données. Certaines IR permettent également le traitement de données par l'intermédiaire d'un EVR par exemple.

* *Glossaire*

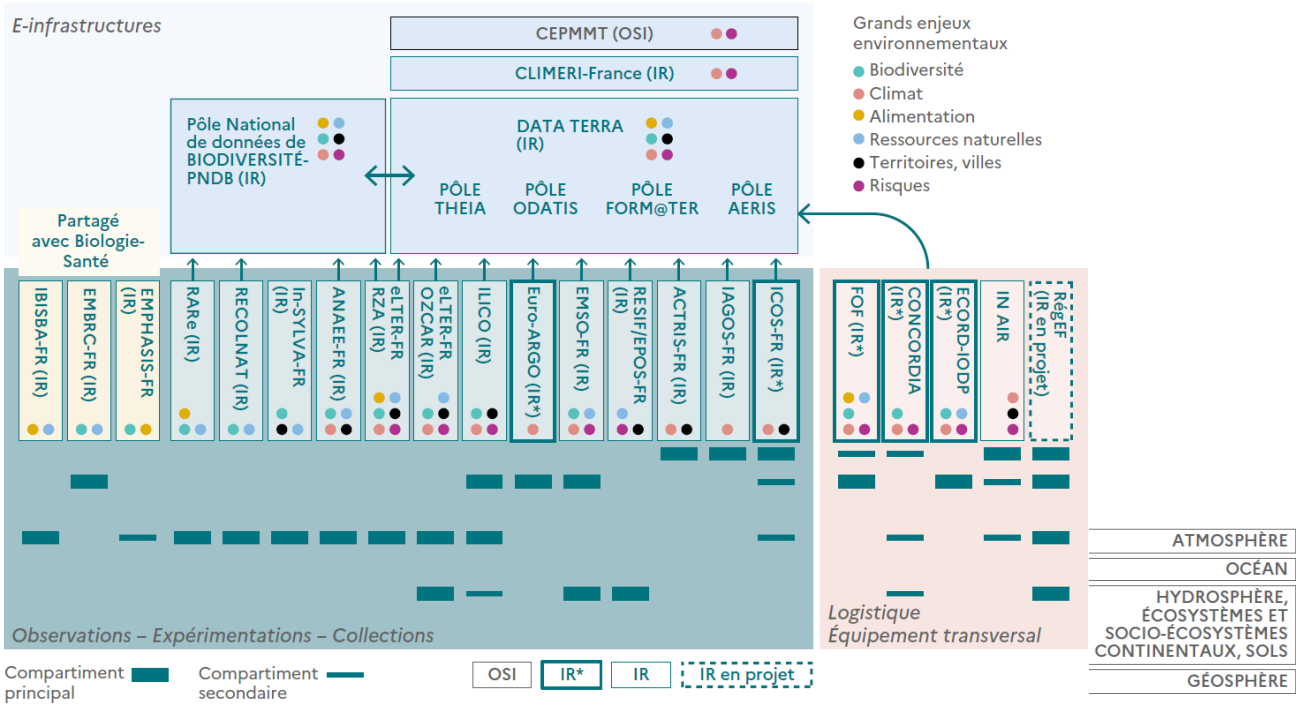


Figure 1.
Feuille de route nationale des Infrastructures de Recherche.
Source : Stratégie nationale des infrastructures de Recherche, 2021.

CNRS Écologie & Environnement développe des dispositifs spécifiques pour lesquels il a été proposé un cadre définissant les principes généraux d'utilisation, de stockage, de diffusion et de réutilisation des données (Politique des données des dispositifs et infrastructures de CNRS Écologie & environnement, juillet 2022 Callou *et al.*, 2022). Ces dispositifs permettent d'observer, d'expérimenter, de modéliser le passé et le présent afin de comprendre le fonctionnement des socio-écosystèmes et d'en prédire l'évolution ; ils sont organisés en réseaux et leurs moyens sont mutualisés (voir annexes). On peut citer : le Réseau des Zones Atelier (RZA), les Observatoires Hommes-Milieux (OHM) du LabEx DRIIHM, le réseau des stations d'Ecologie expérimentale (ReNSEE), les sites d'étude en écologie globale (SEEG) et AnaEE France.

La communauté CNRS Écologie & Environnement interagit également avec d'autres infrastructures de recherche portées ou coportées par d'autres instituts du CNRS ou par des organismes étrangers (Figure 2). Le bilan ci-dessous reflète la complexité du paysage auquel sont confrontés les producteurs de données de CNRS Écologie & Environnement représentés par les participants à l'atelier.

Le projet structurant GAIA Data (EquipEx+ PIA3 2022-2030) est porté par trois ingénieurs de recherche (IR) inscrits sur la feuille de route nationale des IR CLIMERI-France, PNDB*, Data Terra, et 21 autres partenaires. Il a pour ambition de développer et mettre en œuvre une plateforme intégrée et distribuée de services /données pour l'observation, la modélisation et la compréhension du système Terre, de la biodiversité et de l'environnement (Les projets du programme d'investissement d'avenir, Data Terra**).

Comment, du côté des infrastructures/dispositifs de recherche (co)portés par CNRS Écologie & Environnement, se met en place la formation aux bonnes pratiques de planification/stockage/diffusion des données et l'interopérabilité des solutions technologiques ? Quelles nouvelles compétences et quels nouveaux métiers émergent autour de la donnée ?

*** Les Les projets du programme d'investissement d'avenir. https://www.data-terra.org/activites/projets-techniques-scientifiques/projets-nationaux/les-projets-du-programme-dinvestissements-davenir/*

Figure 2.
Liste des infrastructures de recherche et des entrepôts utilisés par les participants à l'atelier.

Infrastructures de Recherche thématiques nationales
InDoRES (Inventaire des Données de Recherche en Environnement) qui comporte deux outils : un catalogue de métadonnées (cat.InDoRES) et un entrepôt de données disciplinaire (data.InDoRES)
DATA TERRA : avec son catalogue et son futur entrepôt de données orphelines en sciences de la terre et de l'environnement (cf. projet EquipEx+ GAIA Data 2022-2030)
PNDB : Pôle National de Données de Biodiversité
OneWater Data : dédiée aux eaux continentales et en projet dans le cadre du PEPR OneWater (2022 - 2032)
Nakala (Huma Num CNRS-INSHS) : pour les données en sciences humaines et sociales

Entrepôts de données généralistes
Dryad
Figshare
OpenAIRE
Open Science Framework OSF
Zenodo

Plateformes ou infrastructures de Recherche thématiques internationales
DEIMS-SDR : Dynamic Ecological Information Management System - Site and Dataset Registry
EMSO : European Multidisciplinary Subsea Observatory, réseau européen d'observatoires sous-marins pour l'environnement
GBIF : pour les données mondiales de biodiversité
GenBank (NCBI) : pour les séquences génétiques et génomiques
Pangaea : pour les données du système Terre et environnement

Infrastructures de Recherche thématiques spécifiques
OZCAR : Données de la zone critique
ICOS : Mesure des flux et des concentrations de gaz à effet de serre
ACTRIS : Observation et exploration des aérosols, des nuages et des gaz réactifs et de leurs interactions
ILICO : Recherche littorale et côtière
PRIDE : Données de protéomique issues de spectrométrie de masse

État des connaissances et des pratiques de la communauté CNRS Écologie & Environnement

Cet atelier a permis d'évaluer le niveau de connaissance et les pratiques des participants via un sondage interactif comportant 25 questions et trois sujets principaux :

- le PDG*,
- les principes FAIR,
- les nouvelles solutions collaboratives pour la gestion et le traitement des données, telles que les lacs de données ou les EVR*.

L'atelier a réuni une cinquantaine de personnes (dont 21 DR/prof., 15 CR/MCF, 7 IR, 6 IE, 1 AI) dont les champs disciplinaires relèvent majoritairement de l'écologie, la génomique, l'environnement, la géographie, la microbiologie, l'écotoxicologie, la géomatique/télédétection, la bioinformatique, l'archéologie et l'anthropologie.
Le panel de participants manipule des données tabulaires (90 %), textuelles (58 %, incluant les séquences génomiques), iconographiques (40 %), géographiques (35 %), vidéos (21 %) et sonores (10 %). La nature des données manipulées est donc extrêmement variée (Figure 3).

Concernant la planification de la gestion des données, plus de la moitié des personnes interrogées a contribué à au moins un DMP dans le cadre d'un projet de recherche (23 pers. - 68 %), pour une IR (7 pers. - 20 %) ou pour une unité (4 pers. - 12 %). Le DMP permet une meilleure anticipation de la gestion des données en optimisant leur structuration et leur pérennisation. Toutefois, le DMP est jugé comme trop chronophage pour la quasi-totalité des personnes sondées, les rubriques à remplir ne sont pas toujours claires et sont parfois inadaptées. Les leviers potentiels pour limiter ces inconvénients sont présentés en deuxième partie de la synthèse.

* *Glossaire*

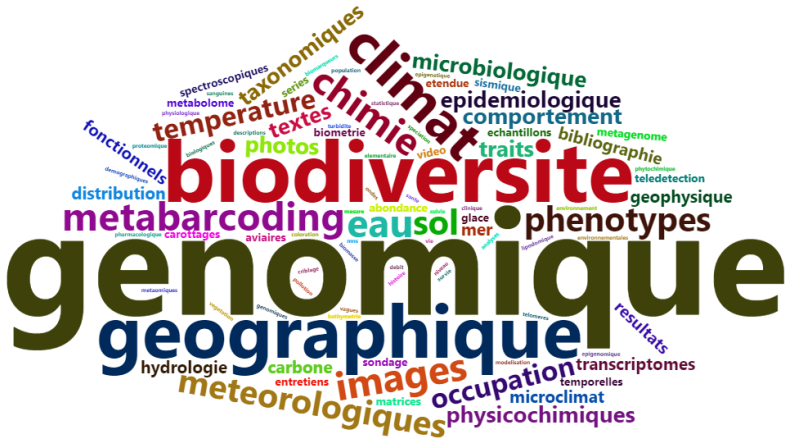


Figure 3.
Des données de nature très variée (Sondage réalisé durant l'atelier, 168 réponses de 50 participants).

Alors que le terme FAIR est connu de tous, les enjeux et le détail de ces principes sont différemment perçus. Il n'est pas aisé de qualifier ses données comme étant FAIR ou pas FAIR. Un consensus émerge dans les discussions : il serait préférable d'exprimer un degré de « FAIR-isation » ou de « FAIRitude » des données. Ce paramètre peut être différemment mesuré selon l'outil-métier (FAIR-Checker, F-UJI…).

L'exploitation des métadonnées décrivant les sites LTER de DEIMS-SDR révèle un degré de FAIRisation plutôt faible avec le « F » et le « A » opérationnels mais les dimensions « I » et « R » quasi inexistantes. Souvent, les métadonnées sont mal dimensionnées (parfois inutilement trop précises, par exemple une emprise géographique trop finement découpée), ou manquantes (pas de dictionnaire de données*, pas d'explication sur la provenance), voire erronées. La question de l'évaluation du degré et de la qualité de la « FAIR-isation » des données prend donc tout son sens.

Il existe aussi un amalgame entre données FAIR et données ouvertes. L'atelier a été l'occasion de rappeler que des données FAIR ne sont pas automatiquement ouvertes, l'accessibilité directe aux données n'étant pas obligatoire (accès via demande au producteur des données).

Un deuxième amalgame a émergé des discussions : le fait de déposer ses données dans un entrepôt de données ou dans une IR ne signifie pas que les données seront automatiquement FAIR. En effet, un entrepôt comme Zenodo ne procède à aucune modération ou curation des données en amont du dépôt. Il permet certes de fournir un DOI et un accès à la donnée, mais sans garantie de qualité, donc un « FA » *a minima* avec des métadonnées relativement pauvres. À l'inverse, certaines IR imposent un format pivot standardisé en entrée, des métadonnées riches et une vérification systématique des données, ce qui va augmenter considérablement le degré et la qualité de la « FAIRisation » (ex. IR OZCAR, entrepôt Pangaea).

Le sondage révèle que l'entrepôt généraliste Zenodo est majoritairement utilisé (figure 4), tout comme GenBank pour les données génomiques. Le nuage de mots résultant mélange des noms d'entrepôts/IR, des noms d'applications (ex. GeoNetwork, Dataverse) et des forges telles que GitHub utilisées essentiellement pour les codes sources, illustrant la confusion dans les outils de stockage, description et exploitation des données et dans le vocabulaire associé. Il est à noter que l'entrepôt Recherche Data Gouv est absent des réponses, cela étant sans doute lié à son existence récente (juillet 2022) ou à la méconnaissance de son existence.

Un dernier amalgame est la confusion entre données FAIR et données reproductibles. Certaines données d'observation comme des séries temporelles peuvent être FAIR tout en étant non reproductibles par nature.

Figure 4.
Identification des entrepôts utilisés par les participants (Sondage réalisé durant l'atelier comportant 168 réponses de 38 participants).

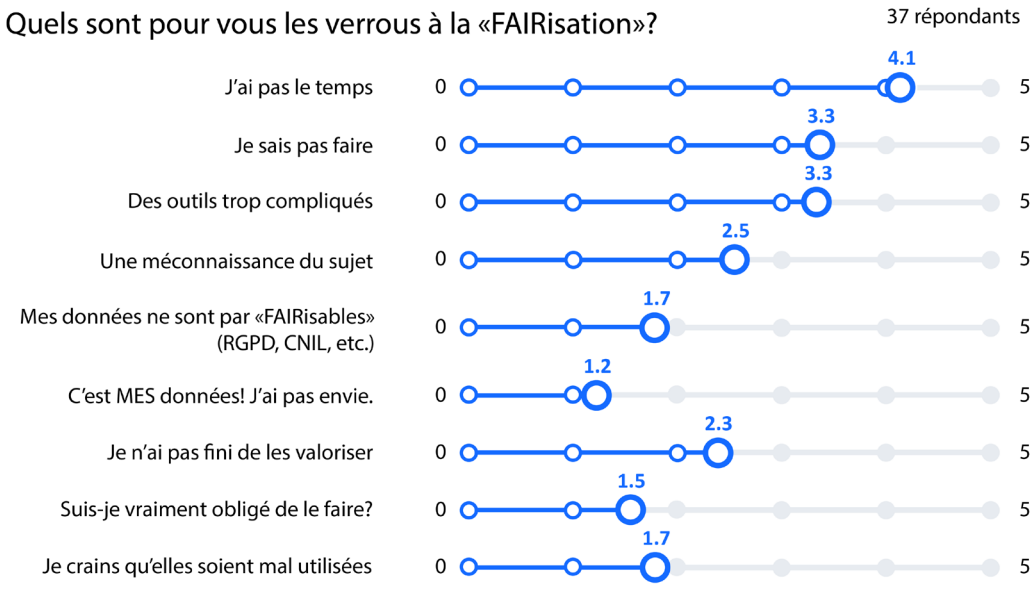


Figure 5.
Les verrous à la « FAIRisation » des données (résultat du sondage, 37 répondants).

Parmi les verrous identifiés dans la « FAIRisation » des données (Figure 5), beaucoup de personnes sondées expriment ne pas avoir le temps. D'autres reconnaissent des lacunes techniques ou ont besoin d'un délai pour le faire afin d'avoir le temps de les valoriser en amont. Les leviers potentiels sont abordés en deuxième partie.

Les lacs de données - encore très peu connus de la communauté de CNRS Écologie & Environnement (la moitié des 40 répondants indiquent ne pas connaître cette notion) - se distinguent des entrepôts de données par le caractère dynamique du stockage des données. Ils prennent en charge les données sous leur forme brute, alors que les entrepôts nécessitent au préalable leur formatage strict. Les lacs de données sont donc capables de collecter et stocker facilement des données de tout type, ce qui les rend très utiles pour le traitement des flux de données générés et transmis de manière automatisée par une large gamme d'instruments de mesure. Les lacs de données sont généralement structurés en plusieurs zones, au sein desquelles les données peuvent subir différents traitements automatisables, permettant dans un second temps leur mise en forme pour des usages variés ainsi qu'une analyse en temps quasi réel (notamment par des algorithmes d'apprentissage automatique), par exemple pour générer des alertes lors de la détection d'anomalies.

Concernant les Environnements virtuels de recherche (EVR), parmi les 33 réponses obtenues, un peu plus de la moitié des participants n'en connaît pas l'existence ou n'en utilise pas. Si l'on prend l'exemple de Jupyter (EVR utilisé par certains participants), qui se définit comme un environnement pour le calcul et la production collaborative, le partage et la publication de documents interactifs, on peut estimer que ces environnements adressent un possible nouveau défi, qui est la production directe et quasi automatique de publications dès la fin de l'analyse des données.

L'hétérogénéité des réponses et les discussions durant l'atelier montrent que les verrous à lever en matière de gestion des données de recherche concernent différentes échelles et réseaux d'acteurs, et que plusieurs actions sont nécessaires pour avancer.

Verrous à lever

Politique des données à l'échelle de CNRS Écologie & Environnement et de ses laboratoires

À l'issue de cet atelier, nous souhaitons tout d'abord faire remonter à la direction de CNRS Écologie & Environnement le besoin de produire un second document sur la politique des données de l'Institut reprenant les grandes orientations stratégiques en lien avec les enjeux de science ouverte et d'écoresponsabilité des données, étendant ainsi le document existant, limité actuellement à ses dispositifs (Politique des données des dispositifs et infrastructures de CNRS Écologie & Environnement, juillet 2022).
La première recommandation au personnel de CNRS Écologie & Environnement producteur et utilisateur de données/codes est d'ouvrir *a minima* les métadonnées pour l'ensemble des jeux de données et codes diffusables. CNRS Écologie & Environnement pourrait également inciter les personnels à publier autant que possible les jeux de données et codes (sous réserve des conditions d'accès à évaluer au cas par cas) via les dispositifs, IR* ou entrepôts de données recommandés et clairement listés dans ce document.
La définition des termes employés est cruciale pour éviter les confusions (données, jeux de données, principes FAIR, stockage, sauvegarde, archivage, entrepôt de données, catalogue de données, PGD, EVR, IR…). Ces définitions seraient à annexer dans la politique de données.

Lever les amalgames autour des principes FAIR et encourager les personnels à rendre leurs données FAIR, autant que possible

Toutes les données et codes produits par le personnel de CNRS Écologie & Environnement ne sont pas à « FAIRiser ». Un tri et une priorisation des données sont nécessaires. Par exemple, les données et codes en lien avec les publications sont prioritaires, secondairement les données qui remontent dans les dispositifs et IR de CNRS Écologie & Environnement, puis les données d'observation uniques (non reproductibles). On peut aussi partager les requêtes effectuées sur une base de données en ligne ou locale plutôt qu'une base de données dans sa totalité.
Ce choix des données à « FAIRiser » va également dépendre des droits associés : droits de propriété, droits de diffusion… Le point de vue du juriste est ici fondamental (A. Robin, 2022).
CNRS Écologie & Environnement peut recommander dans sa politique des données, pour l'ensemble du spectre disciplinaire, l'utilisation de tel ou tel entrepôt de données pour le stockage physique à moyen terme et la diffusion des données, afin d'assurer un niveau minimal de « FAIRisation ». *Data.InDoRES* peut être l'entrepôt par défaut pour les données de recherche en écologie et environnement, si la communauté disciplinaire concernée n'en dispose pas déjà. Pour les codes, CNRS Écologie & Environnement pourrait encourager leur dépôt dans *Software Heritage* ; une étude de faisabilité pourrait être engagée à ce sujet.

Améliorer la stratégie d'exposition des données et codes pour CNRS Écologie & Environnement

Pour rendre les données en écologie et environnement les plus visibles et les plus facilement trouvables, CNRS Écologie & Environnement peut réfléchir à une stratégie de catalogage. C'est par une exposition large des métadonnées dans des catalogues consultés fréquemment par les communautés scientifiques et citoyennes que les données sont visibles. Il pourrait promouvoir l'usage du catalogue InDoRES (couplé à l'entrepôt Data.InDoRES, mais aussi ouvert au catalogage d'autres ressources) qui est lui-même « moissonné » par le géocatalogue national, et prochainement Recherche Data Gouv et Data Gouv, sans oublier le catalogue européen d'EOSC*.
Les IR (co)portées par CNRS Écologie & environnement pourraient poursuivre, ou engager selon les cas, ces dynamiques de moissonnage pour maximiser le rayonnement et l'exposition des données cataloguées. Pour les laboratoires utilisant des entrepôts disciplinaires non portés par le CNRS, une réflexion serait à engager pour vérifier la stratégie de catalogage et estimer le niveau de « trouvabilité » des données.

La communication à développer sur la Politique des données et l'existence du réseau des correspondants Données OpenDoRES des laboratoires de CNRS Écologie & Environnement

Le réseau OpenDoRES vise à regrouper un ou plusieurs correspondants Données de chaque laboratoire ayant comme tutelles le CNRS et le MNHN. Créé en septembre 2022, il est coordonné par l'unité BBEES avec pour objectif de sensibiliser et former la communauté aux questions d'ouverture et de gestion des données de recherche (cycle de vie de la donnée, questions juridiques, dépôt dans un entrepôt, licence…). Les missions de ce réseau sont encore peu connues de la communauté de CNRS Écologie & environnement. Les contours des rôles des correspondants ainsi que l'articulation opérationnelle entre chercheurs/correspondants OpenDoRES, dispositifs/infrastructures, IR nationales et internationales thématiques restent à clarifier et à définir pour chaque laboratoire. Le document sur la Politique des données de CNRS Écologie & environnement pourrait permettre ces clarifications.
La mise en œuvre de la politique des données au sein des laboratoires et l'accompagnement dans la « FAIRisation » des données/codes requièrent d'importants moyens humains et financiers, et une synergie avec les autres instituts du CNRS et les autres organismes de recherche dans le cas de laboratoires multitutelles. De nouveaux métiers émergent autour des données, de nouvelles compétences sont à acquérir dans la communauté en écologie et environnement.

Propositions d'actions en lien avec la politique des données et la gestion RH de CNRS Écologie & Environnement

- Missionner un groupe de travail pour compléter la Politique des données des dispositifs et infrastructures et ainsi couvrir tous les types de données produits par la communauté de CNRS Écologie & environnement. L'objectif serait de proposer une trame de politique de données à l'échelle de chaque unité ;
- mettre en place un comité de suivi des opérations pour veiller à la bonne articulation de l'organisation mise en œuvre entre les différents acteurs : chercheurs, correspondants OpenDoRES, dispositifs/infrastructures, IR nationales et internationales thématiques ;
- suivre régulièrement les dispositifs et infrastructures de CNRS Écologie & Environnement dans leur effort d'accompagnement/formation des producteurs de données, le contrôle qualité des métadonnées et la démarche éventuelle de certification ;
- conseiller les unités quant au choix des IR les plus adaptées pour leurs données ;
- soutenir des actions œuvrant dans la création de vocabulaires standardisés pour les données en écologie et en environnement, avec si possible la dimension web sémantique, comme par exemple les travaux en lien avec le GDR SémanDiv (Sémantique de la Biodiversité) pour l'« *Ecological Trait-data standard* » ou AnaEE pour les écosystèmes ;
- soutenir des actions œuvrant dans l'évaluation de la qualité des données diffusées et la complétude des métadonnées ;
- la dimension « e » (économique, écologique et énergivore) est aujourd'hui absente du FAIR alors même que toutes les étapes du cycle de la donnée ont un coût énergétique. Une étude visant à dresser un rapport d'expertise sur cette question pourrait être lancée ;
- missionner un groupe de travail sur la multidis-ciplinarité/l'interdisciplinarité pour identifier et recommander l'usage d'outils (lacs de données, EVR, autres) facilitant les croisements et les analyses de données pluridisciplinaires ;
- prévoir des profils de postes ouverts au recrutement (ou accompagnement de personnels en réorientation professionnelle) de *data stewards* et autres métiers liés. La formation des correspondants Données est également à soutenir ;
- poursuivre la formation des correspondants OpenDoRES et soutenir des actions de formation de type ANF* et écoles thématiques communes aux personnels chercheurs et ingénieurs et techniciens producteurs et utilisateurs de données (action inter-instituts/organismes) :
  - planification de la gestion des données (et par conséquent, savoir remplir un PGD) ;
  - documentation des données à l'aide de métadonnées ;
  - apprentissage et intégration de vocabulaires métiers (*thésaurus*, ontologies) ;
  - utilisation de lacs de données ou de EVR.

* *Glossaire*

Recommandations auprès des unités de CNRS Écologie & Environnement pour encourager la planification de la gestion des données et la « FAIRisation » des données et codes, avec la mise en œuvre autant que possible de mesures écoresponsables

Une recommandation générale est d'encourager chaque unité à identifier un correspondant Données pour participer au réseau OpenDoRES.

Concernant les personnels producteurs ou gestionnaires de données

Les recommandations ci-dessous sont issues des discussions lors de l'atelier et post-atelier. Elles sont structurées selon les étapes du cycle de vie des données du guide de bonnes pratiques sur la gestion des données de la recherche réalisé par un groupe de travail de la MITI* (Hadrossek *et al.*, 2023) :

Concevoir/Planifier :

- identifier et utiliser les référentiels communs pour les jeux de données à collecter, acquérir ou générer ;
- identifier en amont où stocker/sauvegarder les données d'intérêt et la stratégie de « FAIRisation » et diffusion des données, en lien avec les recommandations de l'Institut et les directives.

Réaliser/Collecter :

- pour la collecte des données de terrain le cas échéant, définir une stratégie instrumentale et de stockage/pérennisation cohérente. Se rapprocher du réseau EcoInfo pour étudier leurs recommandations écoresponsables ;
- concernant la gestion raisonnée des échantillons, des données et des codes : trier, ne pas tout stocker et sauvegarder, apprendre à jeter ;
- utiliser des cahiers de laboratoire électroniques par exemple pour saisir des métadonnées dès les phases de collecte/création des données.

Réaliser/Traiter et Réaliser/Analyser :

- explorer et si besoin se former sur les nouveaux outils tels que les lacs de données et les EVR.

Préserver/Archiver :

- identifier parmi les données triées lesquelles rendre FAIR autant que possible, selon des niveaux de priorité recommandés par CNRS Écologie & Environnement ;
- sauvegarder les données sélectionnées après tri selon la règle du 3-2-1 (à modérer avec les mesures d'écoresponsabilité) qui signifie de disposer de trois copies des données : stocker ces copies sur deux supports différents et conserver une copie de la sauvegarde hors site.

Publier/Diffuser :

- documenter aussi finement que possible les données, au moins celles vouées à être partagées/diffusées, dans un cadre scientifique et en tenant compte des destinataires de la donnée (précision des mesures, volume…) ;
- choisir un entrepôt de « confiance » où publier les données :
  - pour les données reproductibles : partager voire ouvrir les codes associés aux données pour permettre aux utilisateurs de reproduire les données ;
  - pour tout type de données : annexer aux données tout protocole qui explique leur provenance.
- valoriser les données, notamment celles rendues le plus FAIR possible par exemple sous la forme de technical papers pour les codes sources, et de *data papers* pour les jeux de données.

Concernant les dispositifs et infrastructures de recherche/données de CNRS Écologie & Environnement

- veiller à fournir les outils et méthodes permettant de rendre le plus FAIR possible les données afin de pouvoir le réaliser de façon itérative ;
- encourager les IR à se rapprocher des exigences des entrepôts certifiés CoreTrustSeal et à étudier la faisabilité de candidater à cette certification ;
- les dispositifs et IR pourraient fournir si possible de manière automatisée ou semi-automatisée des éléments alimentant le DMP et la saisie des métadonnées ;

* *Glossaire*

- les dispositifs et IR devraient poursuivre et renforcer la collaboration avec les scientifiques de leur communauté pour maximiser la qualité des données et métadonnées ;
- un système de repérage des doublons pourrait être imaginé ;
- veiller, du point de vue du catalogage, à leur interopérabilité avec les infrastructures thématiques partenaires qu'il reste à clairement identifier ainsi que les futurs catalogues de Recherche Data Gouv et d'EOSC. Doit-on interroger l'ensemble des dépôts via un seul portail et est-ce utile ? Quelles sont leurs interactions avec la future plateforme intégrée issue du projet GAIA Data ? Chaque IR doit s'interroger sur l'interopérabilité mise en œuvre ou à mettre en œuvre. Une cartographie des interactions entre les différentes IR et dispositifs (co)portés par CNRS Écologie & Environnement serait à faire connaître auprès des directeurs d'unités et leurs correspondants Données via le réseau OpenDoRES.

RÉFÉRENCES

- Callou, C., Charpentier, I., Clavreul, A., Hénon, A., Joly, D., et al. (2022). Politique des données des dispositifs et infrastructures de l'INEE. HAL, hal-04000652.
- Deuxième Plan national pour la Science Ouverte (2021). Guide de bonnes pratiques sur la gestion des données de la Recherche. v2 2023. Ministère de l'enseignement supérieur, de la recherche et de l'innovation, p. 32. Disponible sur : https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte/
- Hadrossek, C., Janik, J., Libes, M., Louvet, V., Quidoz, M., et al. (2023). Guide de bonnes pratiques sur la gestion des données de la Recherche. v2 2023. HAL, hal-03152732v2.
- Plan national pour la Science Ouverte (2018). Ministère de l'enseignement supérieur, de la recherche et de l'innovation, p. 12. Disponible sur : https://www.ouvrirlascience.fr/plan-national-pour-la-science-ouverte
- Robin, A. (2022). Droit des données de la recherche - Science ouverte, innovation, données publiques. Larcier. HAL, hal-03630680.
- Stratégie nationale des infrastructures de Recherche (2021). Ministère de l'enseignement supérieur, de la recherche et de l'innovation, p. 270. Disponible sur : https://www.enseignementsup-recherche.gouv.fr/fr/la-feuille-de-route-nationale-des-infrastructures-de-recherche-2021-84056
- Wilkinson, M.D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3:160018. doi: 10.1038/sdata.2016.18.

Les défis méthodologiques du **phénotypage haut débit**

Auteurs : Violaine Llaurens (ISYEB), Benjamin Marie (MCAM) et Christophe Salon (Agroécologie INRAE)

3 PRIORITÉS SCIENTIFIQUES À ABORDER D'ICI 2030

- **Développer les plateformes d'acquisition de données à haut débit utilisables pour une large diversité d'espèces vivantes**
- **Former les acteurs de la recherche aux traitements des données brutes, notamment par les méthodes d'apprentissage automatiques**
- **Favoriser les échanges entre disciplines pour développer le phénotypage intégratif**

Introduction

Les méthodes de séquençage d'ADN ont connu des avancées technologiques majeures durant les dernières décennies, ouvrant l'accès aux génomes complets de multiples organismes, y compris les organismes non modèles. De nombreux développements ont en parallèle été réalisés pour l'analyse de ces grands jeux de données. Cependant, si l'avènement de la génomique a donné lieu à des avancées majeures dans les domaines de la biologie évolutive, la caractérisation à large-échelle des phénotypes des organismes séquencés soulève maintenant des défis méthodologiques et techniques importants. En effet, l'étude des variations des traits reste centrale pour résoudre une grande partie des questionnements actuels sur l'écologie et l'évolution des espèces, appelant à développer des outils permettant le phénotypage à haut-débit des organismes.
Le but de cet atelier était d'établir l'état des lieux actuel des méthodes, outils, initiatives passées ou en cours concernant le phénotypage à haut-débit ainsi que les perspectives attendues. Au sein de CNRS Écologie & Environnement, les chercheurs et chercheuses étudient un grand nombre d'organismes auxquels sont associés des traits variés, cela implique de nombreux défis méthodologiques pour l'acquisition des données et leurs analyses, mais aussi leur stockage et leur partage.

État des lieux

Le phénotypage d'organismes non-modèles

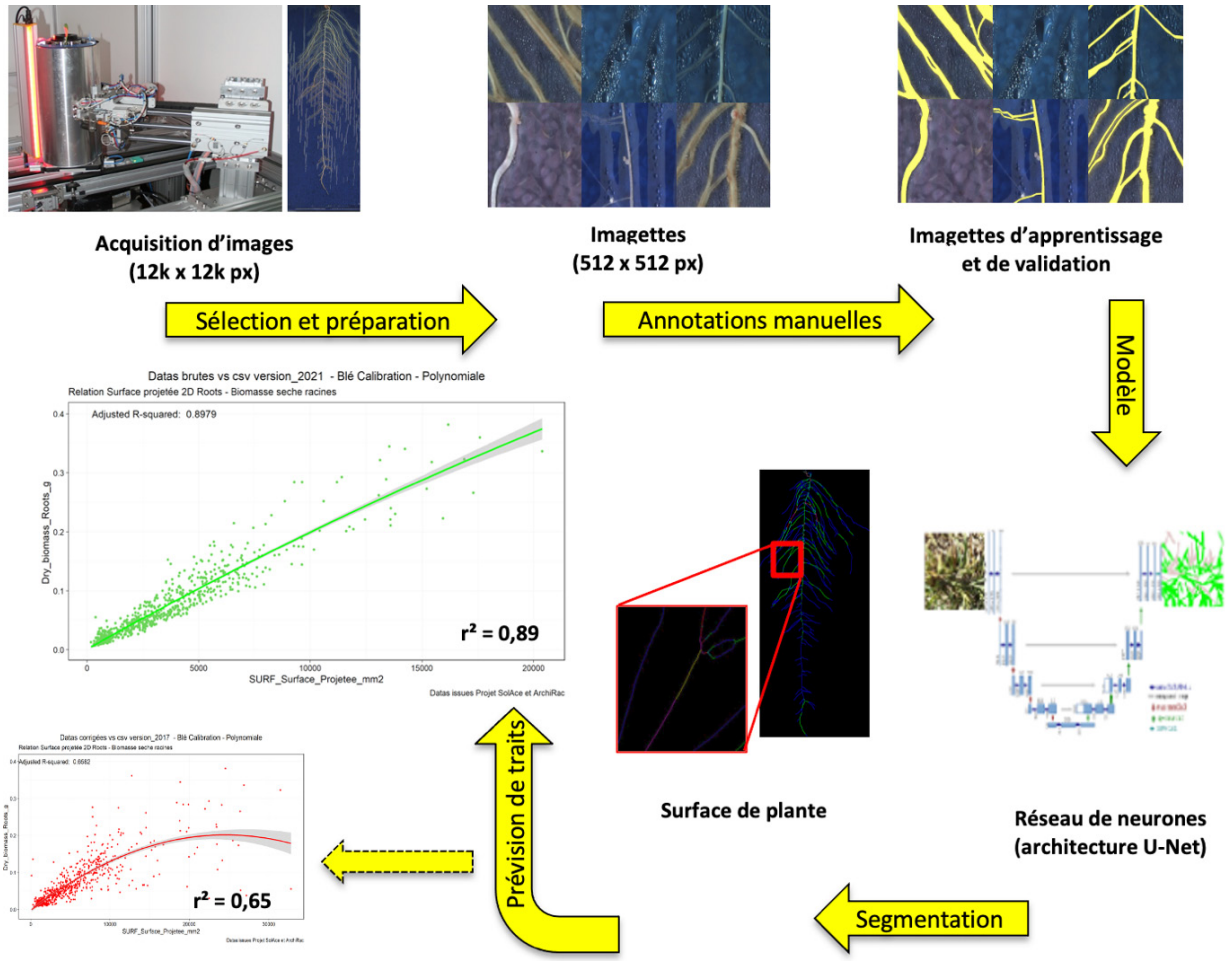
Les questions méthodologiques liées au phénotypage haut-débit intéressent les chercheur. es travaillant sur des espèces actuelles comme fossiles, à différentes échelles (des molécules aux communautés) et sur des espèces couvrant largement l'arbre du vivant. Ces questionnements requièrent souvent l'étude de phénotypes très divers, et les méthodes développées sur des organismes modèles ne sont pas forcément adaptées à la grande diversité des organismes étudiés. L'éventail des traits phénotypiques d'intérêt (morpho-anatomie, expression génique, composition chimique, son, comportement, traits de vie et performances…) reste ainsi assez différent selon les espèces étudiées pour leur écologie ou leur évolution.
Dans le domaine du végétal, les approches et les outils de phénotypage à haut-débit sont bien développés pour une diversité d'espèces, concernant les dispositifs de cultures (Jeudy *et al.*, 2016), l'acquisition de données (ex. par imagerie dans différentes longueurs d'onde) ou bien les variables environnementales et leur traitement. Le phénotypage à haut-débit des parties aériennes ou souterraines des plantes est réalisé au sein de plateformes robotisées, en conditions contrôlées ou en conditions naturelles (Atkinson *et al.*, 2019). Ces outils et infrastructures permettent de quantifier des traits phénotypiques morphométriques ou fonctionnels d'espèces dites "modèles" telles qu'*Arabidopsis thaliana* ou *Medicago Truncatula*, ainsi que de plantes des agroécosystèmes (Figure 1). De telles approches sont aussi développées pour des champignons ou animaux à intérêt agronomique, ainsi que pour les modèles classiques utilisés dans la recherche fondamentale en biologie, tels que la souris, la drosophile ou la bactérie.

Il semble cependant peu réaliste de développer des plateformes spécifiques pour chacun des organismes à étudier car l'éventail des traits phénotypiques reste globalement spécifique à ces différents modèles. De plus la compatibilité des modalités de maintien des organismes au laboratoire en vue de l'automatisation de l'acquisition et du traitement de données de phénotypage a aussi été identifiée comme une nécessité qui reste donc conditionnée par les capacités d'acclimatations aux conditions de laboratoire. Ainsi les structures de phénotypage à haut-débit

d'organismes relativement aisés à élever, tels que les bactéries ou les vers nématodes, connaît un certain essor ; cependant ces développements restent limités à une certaine gamme de traits pouvant être aisément abordés chez ces organismes. Par exemple des méthodes de comptage cellulaire ou de détection automatique permettent l'accès à des traits tels que la croissance des populations, d'activité locomotrice ou de comportement.

Dans le but de comprendre les mécanismes impliqués dans l'évolution des traits ou de leur implication dans les interactions écologiques avec d'autres espèces, le phénotypage des traits s'avère ainsi constituer un défi majeur de nos communautés pour les années à venir.

Figure. 1. Canal d'acquisition et d'analyse de données phénotypiques à haut-débit, pour l'étude des phénotypes racinaires de plantes (4PMI INRAE Dijon). Les images acquises par le robot dans la cabine d'imagerie sont coupées en imagettes plus petites qui servent ensuite à l'apprentissage de vérité terrain.
Par *deep learning*, la segmentation des images est ainsi fortement améliorée et permet d'obtenir une meilleure précision entre les traits, ici la surface projetée de racines, estimés par le modèle d'après les images (axe X) et ceux mesurés de manière destructive (axe Y).



Des molécules aux traits d'histoire de vie : des méthodes d'acquisition très diverses avec leurs défis spécifiques

Notre atelier a offert un aperçu de la large gamme de traits étudiés par les différents chercheurs et chercheuses de CNRS Écologie & Environnement : ainsi la notion de phénotypage à « haut-débit » ne concerne pas uniquement le nombre d'individus étudiés et leur variabilité génétique, mais aussi le nombre et la diversité des traits mesurés et la fréquence de mesure de ces traits, ainsi que le traitement des données ainsi générées. Nous avons ici identifié plusieurs catégories de traits et les principaux défis méthodologiques associés, cette liste restant non-exhaustive.

Traits moléculaires

De nombreux traits concernent l'échelle moléculaire : ainsi les patrons d'expression des gènes, mais également les variations des abondances des très diverses familles de composés chimiques qui constituent les organismes (ou mêmes les microbiomes/communautés dans le cas des écosystèmes microbiens) sont fréquemment étudiés. Les analyses globales, ciblées ou non, telles que le permettent les approches de transcriptomique ou de métabolomique, sont à présent maîtrisées par des plateformes spécialisées, permettant d'obtenir des données moléculaires sur des grandes séries d'échantillons (ex. métabolomique - Figure 2). Ces méthodes donnent accès à des grands jeux de données d'abondances relatives d'une multitude de différents traits chimiques (ex. métabolites primaires ou accessoires, ARN messagers), jusqu'à plusieurs dizaines, voire centaines de milliers de traits par analyses.

Des approches globales d'analyses multivariées permettent en premier lieu d'aborder dans leur ensemble la multitude des variations de ces différents traits afin de définir grâce à des approches statistiques dédiées lesquels présentent un intérêt biologique spécifique. Cependant, la caractérisation de la pertinence fonctionnelle de l'ensemble de ces différents traits reste souvent difficile à aborder. La caractérisation des métabolites, par exemple, se heurte à des problématiques spécifiques d'annotation qui reste encore limitée aux molécules les mieux connues. Elle requiert ainsi en parallèle le développement de corpus de données, qui restent trop souvent le fruit d'initiatives spécifiques. Ainsi, ces corpus peuvent être plus ou moins génériques (métabolisme primaire, expression de gènes de ménages) ou spécifiques aux différents organismes et à leurs singularités respectives.

À l'échelle des cellules, leur marquage chimique, ou celui de leur expression génique permet de caractériser plusieurs centaines de génotypes différents, tandis que la plupart du temps, seul un nombre limité de traits peut être mesuré en

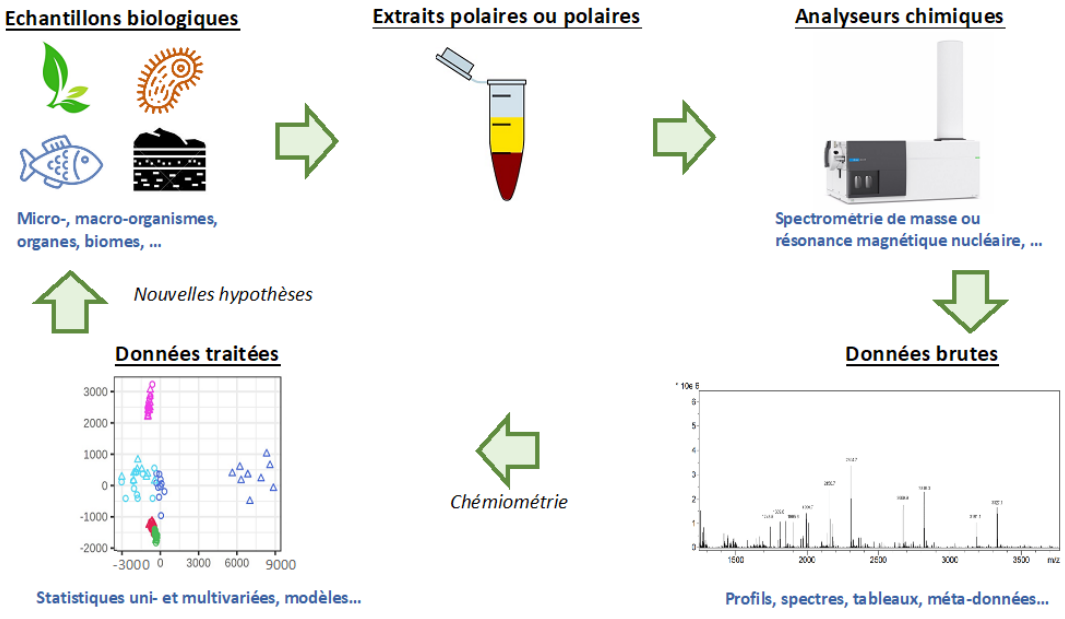


Figure 2 : Schéma conceptuel du pipeline analytique de métabolomique décliné aux sciences de l'écologie et de l'environnement.

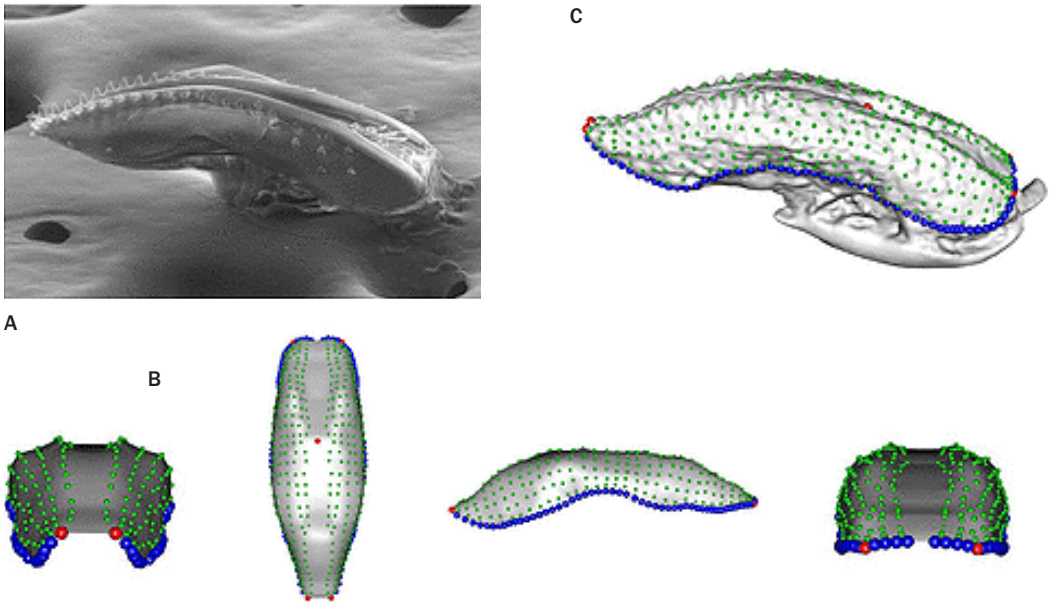
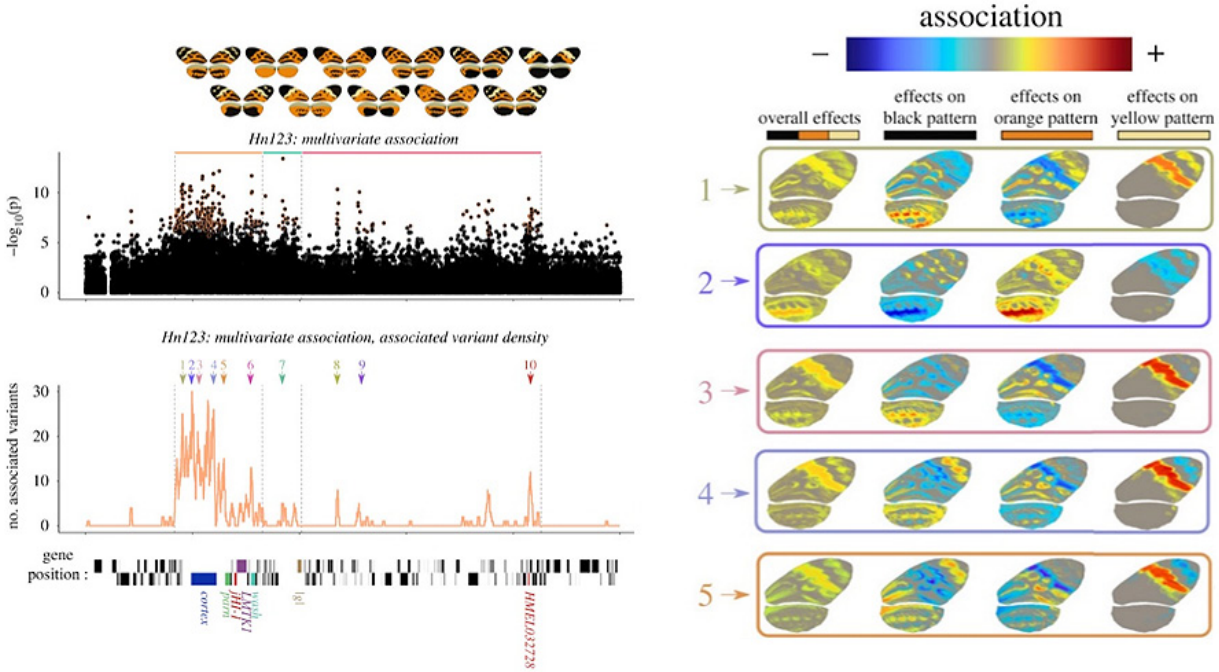


Figure 3. Exemple de caractérisation morphologique en 3 dimensions de l'ovipositeur de l'espèce invasive Drosphila suzukii (d'après Gonzalez-Varon *et al.* 2020). Dans cette espèce, on observe des ovipositeurs très résistants permettant la ponte sur des fruits sur pied, tandis que les espèces sœurs ont des ovipositeurs ne permettant pas de percer les fruits sur pieds et la ponte a ainsi principalement lieu sur les fruits pourris tombés au sol. L'étude de l'évolution de ce trait a requis sa caractérisation par microscopie électronique (A) Image d'un ovipositeur obtenue par microscopie électronique (B) Reconstruction en trois dimensions de l'ovipositeur et (C) construction d'un modèle de référence de la forme de l'ovipositeur. Placement des points-repères placés : *landmarks* en rouge, *semilandmarks* en bleu et en vert les *semilandmarks* de surface.

Figure 4 : Exemple d'étude d'association entre variations génétiques et phénotypiques permise par le développement de méthode d'analyse d'image chez le papillon *Heliconius numata* (d'après Jay *et al.*, 2022). Dans cette espèce, une région unique – appelée supergène P- contrôle les variations de motifs, et présentent des inversions de l'ordre de ses gènes. Identification des associations entre la variation génétique au sein de l'inversion Hn123 et les variations de motifs de coloration des ailes. Effets phénotypiques des principaux variants génétiques détectés.



même temps (une fluorescence révélant l'induction d'un gène, ou une valeur sélective) ce qui rend ces approches plus propices à l'analyse de phénotypes spécifiques.

Traits morphologiques

Le phénotypage à haut débit de traits morphologiques, déjà largement développé pour les plantes au sein de différentes platefores (voir supra) est aussi en plein essor pour d'autres modèles biologiques, en particulier grâce à la possibilité de caractériser les variations morphologiques en trois dimensions (Figure 3).
Les nouvelles méthodes d'analyse d'images, dans le domaine du visible, comme dans l'hyperspectral, permettent également la caractérisation de phénotypes de coloration. Par exemple, l'analyse d'images de motifs de coloration des ailes variables au sein d'une espèces permet de réaliser des associations statistiques entre variations génétiques et phénotypiques (Figure 4).

Traits sonores

Le développement des enregistrements sonores dans différents environnements, ainsi que de leur traitement automatisé, connaît également un essor important (Figure 5). Ces techniques d'analyse, basées notamment sur le traitement du signal et l'apprentissage automatisé permettent le suivi de communautés naturelles et de mesurer très concrètement, par exemple, l'impact anthropique sur les interactions écologiques entre espèces (Folliot *et al.,* 2022).

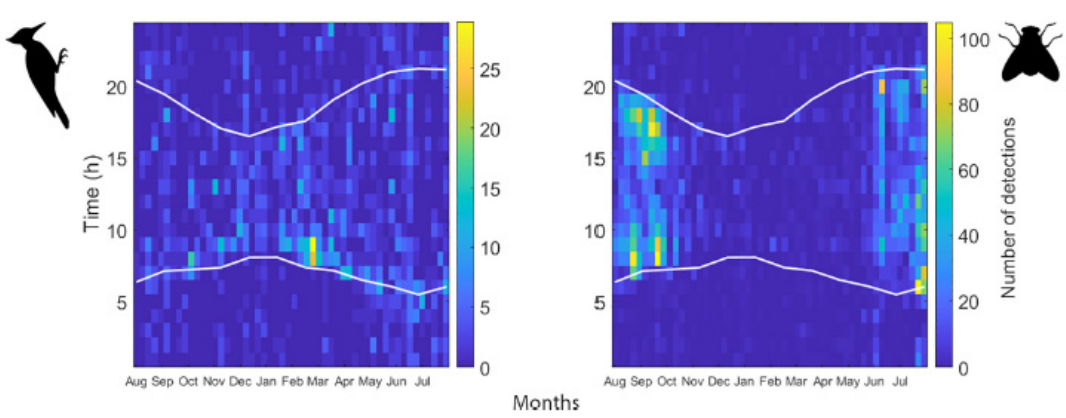
Traits comportementaux

Chez les animaux, la dimension comportementale implique le développement de techniques de phénotypage spécifique. Ainsi le développement de méthodes de suivi automatique basé sur de l'apprentissage automatisé et supervisé (Figure 6) permet d'analyser un grand nombre de séquences vidéo (Mathis *et al.* 2018) et de caractériser le comportement d'un grand nombre d'individus dans des conditions pré-établies. Cependant, la calibration de l'espace dans lequel les animaux sont filmés, ainsi que la nécessité d'acquérir un jeu de données suffisant lors de la phase d'apprentissage, limitent en général le nombre de conditions pouvant être déjà investiguées : on testera ainsi dans un premier lieu le comportement d'individus dans un contexte bien contrôlé (ex. cage expérimentale) avant de pouvoir étendre l'utilisation de ces outils à des conditions plus complexes et fluctuantes.
Des développements spécifiques d'identification individuel permettent également d'étudier le comportement de certains animaux *in naturae*. Ainsi, la méthode développée par Ferreira *et al*. (2020) permet, quant à elle, de quantifier chez les oiseaux le nombre de visites de chaque individu d'une population vers son nid ainsi que vers différentes sources de nourritures.

Traits d'histoire de vie

Une dernière gamme de traits, les traits d'histoire de vie, apparaît très pertinente dans le cadre des recherches menées au sein de CNRS Écologie &

Figure 5. Variations annuelles et journalières dans l'activité des pics-vert et des insectes pollinisateurs au sein d'une forêt alpine, mesurées par la détection des sons de type « martèlement » (panneau de gauche) et « bourdonnement » (panneau de droite) par apprentissage automatisé. La gamme de couleurs indique l'intensité des interactions estimée par le modèle (Folliot *et al.* 2022).



Environnement. En effet, les traits d'histoire de vie (tels que le temps de développement, la longévité, l'investissement reproducteur) ont généralement des effets directs sur la valeur sélective des individus suscitant ainsi l'intérêt des chercheurs abordant les questions évolutives et les interactions avec les conditions environnementales, qu'elles soient biotiques (compétition, prédation, mutualisme) ou abiotiques (effet du changement climatique par exemple).

Contrôle du stade développemental

Pour permettre des comparaisons pertinentes entre individus ou entre espèces, le phénotypage haut-débit implique de pouvoir caractériser ou contrôler de manière fine le stade développemental auquel le trait est mesuré. Ceci est d'autant plus difficile lorsque que la fenêtre temporelle au cours de laquelle le trait comportemental peut être mesuré est courte. Pour des questions liées à la biologie du développement notamment, le phénotypage haut-débit peut également nécessiter des mesures de traits répétées au cours du développement des différents individus étudiés.

Contrôle des conditions environnementales

Enfin, les phénotypes mesurés dépendent fortement des conditions environnementales dans lesquelles les individus se développent et sont observés (plasticité phénotypique). Ainsi, une approche haut-débit peut également impliquer la mesure d'un même trait dans plusieurs expériences avec des conditions environnementales similaires pour tous les essais ou, bien au contraire, en faisant varier les conditions environnementales pour obtenir des courbes de réponses des traits phénotypiques. La précision de mesure et de contrôle des paramètres environnementaux est ainsi un élément clef afin de parvenir à une comparaison non biaisée des variations phénotypiques entre espèces. Certaines plateformes sur lesquelles CNRS Écologie & Environnement joue un rôle majeur, telles que la station expérimentale de Moulis ou l'écotron de Foljuif, permettent des élevages en conditions contrôlées d'organismes terrestres ou aquatiques et ouvrent de nombreuses possibilités sur le phénotypage d'un grand nombre d'organismes, placés dans des conditions environnementales variées.

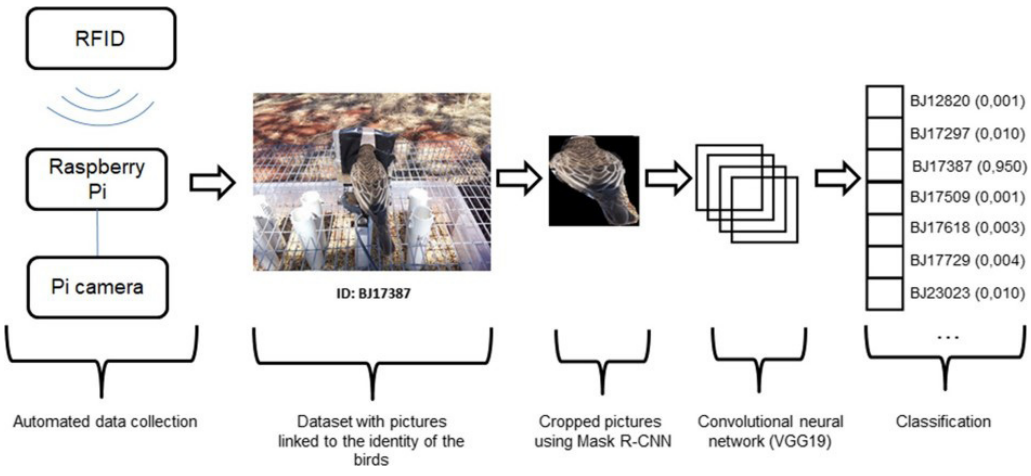


Figure 6. Aperçu de la façon dont des oiseaux visitant une mangeoire peuvent être identifiés individuellement grâce à un algorithme d'analyse d'image, basé sur de l'apprentissage automatisé (Ferreira *et al.*, 2020).

Questions futures et verrous à lever

Défis techniques liés aux grands jeux de données phénotypiques

Le phénotypage à haut-débit engendre de très grands jeux de données, posant tout d'abord le problème de leur stockage et de leur transfert. Ce problème est d'autant plus important que si des bases de données internationales de référence existent et sont bien établies, pour l'ensemble de la communauté, en ce qui concerne les données de génomique ou de transcriptomique (*National Center for Biotechnology Information, European Nucleotide Archive*), la nature

beaucoup plus hétérogène des données de phénotypage et la diversité des communautés qui les génèrent et les utilisent, limitent pour le moment la mise en place de grandes bases de données de phénotypage à large échelle et nous questionnent quant aux capacités de stockage et de bancarisation de ces données au sein des unités de recherche de CNRS Écologie & Environnement (voir aussi Atelier « Données, après l'acquisition »). Il existe en revanche des bases de données phénotypiques centrées sur des organismes modèles, telles que *Flybase*, qui permettent de rassembler des données de phénotypes divers, obtenues sur les drosophiles. D'autres bases de données génétiques, comme *barcode of life* (BOLD), donnent accès non seulement à des séquences génétiques permettant l'identification des espèces mais aussi à des photos pouvant permettre des études sur la morphologie.

Ontologie des phénotypes

La comparaison entre espèces, mêmes proches, ainsi que l'intégration de traits mesurés à différentes échelles sur les mêmes individus requièrent de définir précisément les traits mesurés, ainsi que les techniques d'acquisition. Ainsi, même pour des traits simples, tels que la taille corporelle chez les animaux ou la hauteur pour une plante, la définition d'un cadre commun s'avère nécessaire pour répondre à certaines questions d'écologie et évolution. On pourra également utiliser des critères développementaux ou fonctionnels selon les questions abordées.
Certains traits plus complexes, tels que la composition génétique des communautés de microorganismes qui constituent un microbiote, sont de plus en plus perçus comme de véritables phénotypes spécifiques à l'hôte ou à l'écosystème qui les abritent ; mais ces utilisations font encore débat au sein même des différentes communautés scientifiques qui les étudient.
L'atelier a ainsi mis en évidence la diversité du vocabulaire employé dans les différentes communautés scientifiques et l'intérêt des échanges entre disciplines pour la caractérisation de phénotypes intéressant différentes équipes travaillant sur des questions distinctes. Ainsi, la notion de « biomarqueur » pourra correspondre à un métabolite bien défini jouant un rôle biologique potentiel pour un spécialiste de l'écologie chimique, alors qu'un écologiste pourra aborder une signature chimique même imparfaitement caractérisée comme bio-indicatrice d'un phénomène particulier, dès lors qu'il puisse rationnellement montrer que ce signal soit biologiquement discriminant.

Défis techniques rencontrés lors du traitement des données brutes

L'atelier a permis de mettre en évidence que le traitement des données phénotypiques requiert fréquemment des développements de méthodes bioinformatiques spécialisées. À titre d'exemple, des progrès des bases de données de références d'écologie chimique nous invite à ré-interroger les données brutes précédemment acquises bien que ces dernières soient très volumineuses et de nature hétérogène (car dépendant du fabricant de l'analyseur utilisé) et donc peu compatibles avec les plateformes déjà existantes (et utilisant des formats compressés), nécessitant ainsi le développement de solutions adaptées et dimensionnées en fonction de ces seules contraintes. De la même manière, l'application de méthodes d'apprentissage profond pour l'acquisition de données comportementales nécessite des compétences en codage et en manipulation de grands jeux de données, soulignant le besoin de formation dans ces domaines. D'autre part, les chercheurs et chercheuses participant à l'atelier ont souligné l'intérêt des échanges sur ces défis entre équipes ou entre laboratoires s'intéressant soit à des traits similaires, mais sur des organismes différents (acquisition de données sur le vol chez les insectes ou les oiseaux), ou à des traits de nature très différente sur des organismes similaires, où le partage d'expérience sur les conditions d'élevage et de réalisation des expérimentations au laboratoire peut avoir un fort impact sur l'acquisition et le traitement des données.

Défis analytiques et statistiques multivariées

La majorité des échanges de l'atelier a notamment porté sur le développement des méthodes d'analyses des données multivariées. Pour certains traits, comme ceux mesurés en morphologie, des méthodes spécifiques, comme la morphométrie géométrique, ont été largement développées et il existe une communauté de chercheurs et de chercheuses actifs et actives, échangeant réguliè-

rement sur les aspects techniques comme scientifiques, notamment à travers les Symposium de Morphométrie et d'Évolution des Formes. Notons par ailleurs que le développement de la chimiométrie en France a pu grandement bénéficier de l'essor de la société française de chimiométrie qui a déjà ainsi alimenté les travaux menés dans de nombreux domaines (médical, industrie…) auxquels ceux des sciences de l'écologie et de l'évolution se sont jointes plus récemment. En revanche, pour d'autres traits, dont l'acquisition a été facilitée par des avancées technologiques plus récentes, les méthodes d'analyses restent encore assez limitées. L'application des méthodes de statistiques multivariées sur des jeux de données où les variables ne sont pas indépendantes et ont des échelles drastiquement différentes introduit des biais, au même titre que les biais liés aux échantillons choisis, au nombre de réplicats ainsi qu'aux conditions d'acquisition. Certaines solutions d'analyses multiblocs dernièrement développées tendent à permettre l'essor de ces approches complexes et une communauté dynamique de statisticiens, développeurs et utilisateurs avertis s'est constituée pour assurer des formations pratiques sous forme d'ateliers d'accompagnement du traitement des jeux de données des utilisateurs (ex. *package mixomics* sous R).
D'autre part, l'étude de certains phénotypes demandant des compétences très spécialisées, par exemple en chimie pour les données métabolomiques, l'intégration de phénotypes variés dans des approches intégratives, soulignent le besoin de collaboration entre équipes, depuis l'élaboration initiale du projet jusqu'à son aboutissement et sa valorisation.

Les chercheurs et chercheuses de l'atelier ont ainsi montré un grand intérêt pour la mise en commun des méthodologies et résultats obtenus dans le cadre d'études de phénotypage à haut-débit, à travers par exemple la mise en place de séminaires réguliers qui pourraient entre autres être soutenus par un RT spécifique.

Défis scientifiques : intégration phénotypique

Tout comme pour la génomique, l'ouverture de nouvelles possibilités de phénotypage, permise par les innovations technologiques récentes, permet d'envisager des études exploratoires à la recherche, par exemple, de nouvelles molécules. Cela étant, les études impliquant le phénotypage à haut-débit permettront également de répondre à des questions fondamentales telles que la coévolution entre morphologie et comportement ou entre espèces mutualistes. De telles études nécessiteront une synergie entre différentes équipes de recherches travaillant sur des phénotypes très différents, dont l'acquisition repose sur des techniques présentant des difficultés et contraintes spécifiques, ainsi que sur des organismes très divers.

RÉFÉRENCES

- Ferreira, A. C., Silva, L. R., Renna, F., Brandl, H. B., Renoult, J. P., Farine, D. R., ... & Doutrelant, C. (2020). Deep learning-based methods for individual recognition in small birds. Methods in Ecology and Evolution, 11(9), 1072-1085.
- Jay, P., Leroy, M., Le Poul, Y., Whibley, A., Arias, M., Chouteau, M., & Joron, M. (2022). Association mapping of colour variation in a butterfly provides evidence that a supergene locks together a cluster of adaptive loci. Philosophical Transactions of the Royal Society B, 377(1856), 20210193.
- Folliot, A., Haupert, S., Ducrettet, M., Sèbe, F., & Sueur, J. (2022). Using acoustics and artificial intelligence to monitor pollination by insects and tree use by woodpeckers. Science of the Total Environment, 838, 155883.
- Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature neuroscience, 21(9), 1281-1289.
- Varón-González, C., Fraimout, A., Delapré, A., Debat, V., & Cornette, R. (2020). Limited thermal plasticity and geographical divergence in the ovipositor of Drosophila suzukii. Royal Society open science, 7(1), 191577.