Plan de Gestion des Données : PEPR ATLASea : Atlas des génomes marins

Erwan Corre, Jean-Marc Aury, Line Le Gall, Patrick Wincker, Hugues Roest Crollius

Plan de Gestion des Données

PEPR ATLASea : Atlas des génomes marins

Version 1.2

Juin 2024

Table des matières

I. Introduction

A. Contexte du programme

Le programme ATLASea vise à explorer la biodiversité des écosystèmes marins naturels du littoral français (Zone Économique Exclusive - ZEE) en échantillonnant et en séquençant le génome de plusieurs milliers d'espèces. Cette étude aura des implications significatives pour la compréhension de la diversité génétique et la conservation des espèces.

B. Objectifs de la gestion des données

Le présent document établit un Plan de Gestion des Données (PGD) pour décrire la manière dont les données collectées au cours du programme seront traitées, stockées et partagées pendant le programme et après son achèvement, en assurant la qualité, l'intégrité et la ré-utilisabilité des données générées. La collecte de données comprend à la fois les données générées par les activités du programme, mais aussi les données réutilisées à partir de recherches antérieures, de référentiels ouverts et de sources similaires. Conformément au modèle de convention de subvention d'Horizon Europe, le PGD d'ATLASea est considéré comme un document évolutif. Cela signifie que le PGD sera régulièrement mis à jour et revu au fur et à mesure de l'avancement du programme afin de refléter les nouvelles informations disponibles ou tout ajustement ou ajout nécessaire. Les informations relatives aux versions seront incluses dans le processus de révision et indiquées clairement sur la page de titre.

C. Portée du Plan de Gestion des Données

Ce plan couvre toutes les étapes de gestion des données, de la collecte des échantillons sur le terrain ou au sein des stations marines jusqu'à l'archivage à long terme des données, en incluant la gestion de la qualité, la sécurité des données, le partage des données et la conformité éthique.

Plus spécifiquement le Plan de Gestion de Données couvre :

- la description générale des sources, des types et des formats de données collectées (réutilisées et générées) au cours du programme,
- l'utilité et finalité des données collectées,
- les besoins de stockage prévus, ainsi que les options de stockage des données à court et à long terme,
- l'alignement à l'échelle du programme sur les principes FAIR et les pratiques scientifiques ouvertes, telles que l'utilisation d'identifiants permanents, de métadonnées riches, de référentiels de domaine ouvert et de normes approuvées par la communauté,
- la gestion d'autres types de résultats de la recherche (par exemple, logiciels, flux de travail, échantillons physiques...),
- les informations concernant l'allocation des ressources, la sécurité des données, les politiques en matière d'éthique et de propriété intellectuelle, et d'autres questions pertinentes.

Outre les données scientifiques, ce Plan de Gestion des Données couvre la gestion des données personnelles et administratives en relation avec :

- le recrutement du personnel nécessaire à la réalisation des objectifs scientifiques (y compris des doctorants, étudiants Master et stagiaires),
- la coordination des travaux scientifiques du PEPR ATLASea (données personnelles des participants du programme ATLASea),
- la communication générale du programme (organisation d'événements telles que les conférences, journées scientifiques, écoles thématiques d'été ou webinaires, la mise en place de la newsletter avec inscription obligatoire ou la mise en place des réseaux sociaux, supports de communication, vidéos de promotion, etc.),
- le reporting scientifique vis-à-vis de l'ANR (reporting scientifique, administratif et financier, indicateurs annuels),

- les appels à projets de recherche en partenariat public privé pour l'innovation,
- les appels à projets pilotes.

D. Contributeurs

Nom	**Projet Ciblé**	**Affiliation**	**Rôles**
Hugues Roest Crollius https://orcid.org/0000-0002-8209-173X	WHEEL-Sea	CNRS	Co-directeur du programme ATLASea
Patrick Wincker https://orcid.org/0000-0001-7562-3454	WHEEL-Sea	CEA	Co-directeur du programme ATLASea
Line Le Gall https://orcid.org/0000-0001-7807-4569	DIVE-Sea	MNHN	Coordinatrice du PC DIVE-Sea. Relectrice PGD
Jean-Marc Aury https://orcid.org/0000-0003-1718-3010	SEQ-Sea	CEA	Coordinateur du PC SEQ-Sea. Relecteur PGD
Erwan Corre https://orcid.org/0000-0001-6354-2278	BYTE-Sea	CNRS	Coordinateur du PC BYTE-Sea. Rédacteur du PGD
Kamil Szafranski	WHEEL-Sea	CNRS	Chef de programme ATLASea
Caroline Belser https://orcid.org/0000-0002-8108-9910	SEQ-Sea	CEA	Relectrice PGD
Patrick Durand https://orcid.org/0000-0002-5597-4457	BYTE-Sea	IFREMER	Relecteur PGD
Franck Bellugeon https://orcid.org/0009-0002-6201-5112	DIVE-Sea	MNHN	Relecteur PGD
Bertrand Bed'Hom https://orcid.org/0000-0002-0825-0886	DIVE-Sea	MNHN	Relecteur PGD
Mélanie Van Weddingen https://orcid.org/0009-0009-2960-2578	DIVE-Sea	MNHN	Relectrice PGD
Benjamin Girard https://orcid.org/0009-0002-0428-1041	DIVE-Sea	MNHN	Relecteur PGD
Loraine Gueguen https://orcid.org/0000-0002-8640-4190	BYTE-Sea	CNRS	Relectrice PGD
E'Krame Jacoby https://orcid.org/0000-0002-3185-1364	SEQ-Sea	CEA	Relectrice PGD
Alexandra Louis https://orcid.org/0000-0001-7032-5650	BYTE-Sea	CNRS	Relectrice PGD

Claude Scarpelli https://orcid.org/0000-0002-2458-9775	SEQ-Sea	CEA	Relecteur PGD
Anthony Bretaudeau https://orcid.org/0000-0003-0914-2470	BYTE-Sea	INRAE	Relecteur PGD

II. Description du programme

A. Objectifs scientifiques

Le programme ATLASea propose de déchiffrer et d'exploiter les informations issues d'espèces marines du littoral Français. Quatre mille cinq cents espèces marines, avec une forte couverture des espèces de métropole mais en incluant aussi des espèces des territoires ultramarins, seront collectées par le Projet Ciblé (PC1) DIVE-Sea. Les tissus nécessaires pour réaliser les extractions d'ADN et ARN seront confiés aux équipes du Projet Ciblé (PC2) SEQ-Sea pour produire des assemblages de génome associés à des annotations structurales de haute qualité. Le Projet Ciblé (PC3) BYTE-Sea s'appliquera à centraliser toutes les données de génomiques produites par le programme ATLASea dans un seul portail et à intégrer les génomes d'organismes connexes séquencés par d'autres consortiums.

De plus, à travers deux Projets Pilotes, l'exploitation des données servira des objectifs ambitieux dans deux domaines :

- La caractérisation de voies métaboliques amenant à des molécules connues, de nouvelles molécules, métabolites, matériaux et voies de synthèses, à partir de données génomiques,
- le suivi de la dynamique des écosystèmes marins, en particulier ceux susceptibles d'être perturbés par l'invasion d'espèces exogènes.

B. Description de l'échantillonnage, du séquençage, de l'assemblage et annotation et de la mise à disposition des génomes

Premier maillon du PEPR ATLASea, l'objectif du Projet Ciblé (PC1) « DIVE-Sea Ressources Biologiques et bancarisation » est de fournir une banque de spécimens pour réaliser le séquençage de génomes de référence pour 4,500 espèces marines françaises avec une forte couverture des espèces de métropole mais en incluant aussi des espèces des territoires ultramarins. L'échantillonnage sera effectué à partir d'une variété d'habitats naturels, en collectant des échantillons biologiques divers tels que des individus, des tissus ou des cellules.

Le Projet Ciblé (PC2) SEQ-Sea prévoit de séquencer le génome des organismes collectés, en s'appuyant sur l'expertise du Genoscope dans l'extraction d'ADN de haute qualité, couplée à une veille technologique permanente. L'ADN et l'ARN de chacun des génomes seront séquencés en s'appuyant sur des technologies de pointe dans l'infrastructure France Génomique. Un environnement informatique assurera la génération des assemblages et de leur annotation, le stockage et la diffusion de ces génomes. Ceux-ci seront produits à un niveau de qualité dit «de référence » afin de générer une ressource de haute confiance pour la biologie et les biotechnologies du futur. L'assemblage et l'annotation des génomes sera réalisée en utilisant des pipelines ouverts et documentés, utilisant des suites logicielles libres.

Le Projet Ciblé (PC3) BYTE-Sea s'appliquera à centraliser toutes les données de génomique produites par le programme ATLASea dans un seul portail et à intégrer les génomes d'organismes connexes séquencés par d'autres consortiums. Il permettra de garantir l'interopérabilité et la sécurité des données, et de faciliter leur diffusion et leur utilisation conformément aux principes FAIR et Open Science.

C. Liste des espèces cibles

Une liste préliminaire d'espèces cibles a été établie en fonction de leur importance écologique, de leur rareté ou de leur statut de conservation à partir de l'inventaire du Patrimoine Naturel (https://inpn.mnhn.fr/accueil/index) qui recense environ 13000 espèces présentes sur le littoral français. Elle sera déposée avec un statut « long-list » du programme ATLASea sur le site Genome On A Tree (https://goat.genomehubs.org/). Cette liste sera actualisée en cours de programme.

III. Collecte et organisation des données

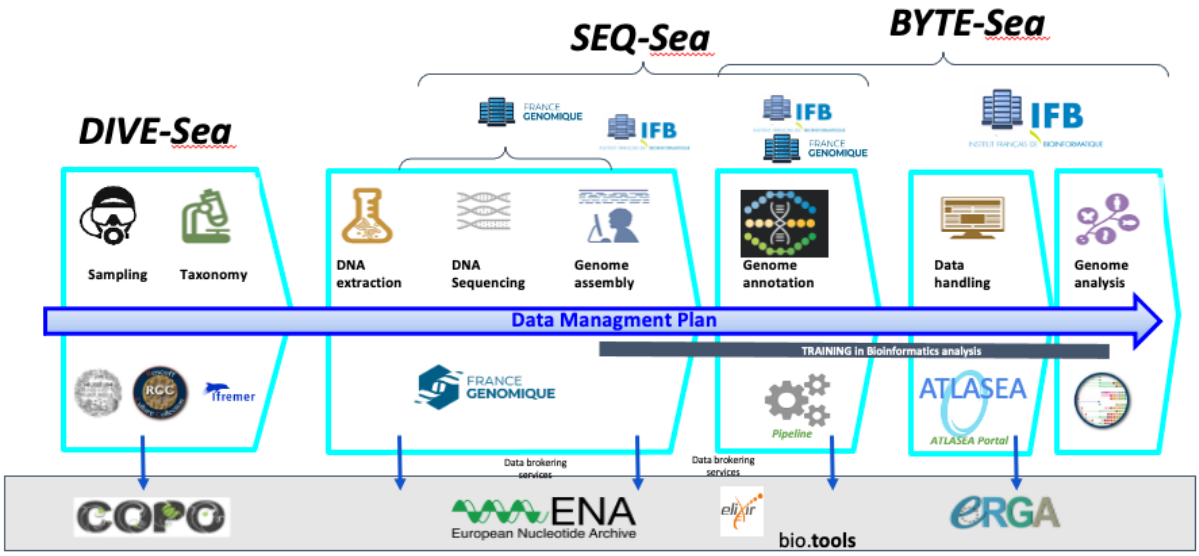
A. Réutilisation de données existantes

L'inventaire complet des jeux de données réutilisés sera construit tout au long du programme pour chaque produit de recherche identifié, et consigné dans les versions successives du présent PGD.

En l'état, il est par exemple établi que :

- Le Projet Ciblé WHEEL-Sea réutilisera des données du référentiel taxonomique TaxRef (https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref?lg=fr) dans le cadre du Comité de Sélection des Espèces, des vidéos et des photos à des fins de communication.
- Le Projet Ciblé DIVE-Sea réutilisera des données de l'inventaire national du Patrimoine naturelle (https://inpn.mnhn.fr/accueil/index), du référentiel taxonomique TaxRef (https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref?lg=fr), du World Register of Marine Species (https://www.marinespecies.org/index.php), des bases de données du MNHN (http://collections.mnhn.fr/wiki/Wiki.jsp?page=Bases_Collections).
- Le Projet Ciblé SEQ-Sea réutilisera des données issues d'une part des protocoles de préparation des échantillons et des SOP de programmes existants visant à séquencer la biodiversité (comme le Darwin Tree of Life), notamment ceux disponibles dans protocols.io. SEQ-Sea utilisera également des données publiques de séquençage RNA-Seq (ENA, NCBI) pour l'annotation structurale mais aussi des gènes marqueurs connus (GENBANK, BOLD) pour valider le nom d'espèce attribué à chacun des échantillons collectés. D'autre part, des logiciels de la communauté et des chaînes de traitement mise au point au Genoscope seront réutilisés dans le cadre du Projet Ciblé SEQ-Sea.
- Le Projet Ciblé BYTE-Sea réutilisera des données déjà publiées par la communauté scientifique dans des entrepôts publics (ENA, ENSEMBL, etc…) : génomes d'organismes marins avec une haute qualité d'assemblage.

B. Production des données



La production des données du programme ATLASea suit une chaîne de traitement à travers laquelle une partie des données produites par un projet ciblé correspond à la donnée collectée par un des autres projets ciblés en suivant le logigramme schématique présenté ci-dessus.

C. Types de données

Les données scientifiques

Les types de données scientifiques utilisées dans le cadre du programme ATLASea varient d'un projet ciblé à l'autre. On retrouve malgré tout deux types de données distinctes :

- Les données biotiques au sein du programme ATLASea concernent principalement les données de génome, les gènes marqueurs, mais aussi les métadonnées générées suite au traitement de ces données.
- Les données abiotiques comprennent les données environnementales qui seront collectées en même temps que l'échantillonnage ainsi que certains documents administratifs concernant les données.

Quelques exemples de types de données attendus et de leurs formats correspondants figurent dans le tableau 1 et seront mis à jour dans les versions ultérieures.

Data types	PC	Data formats	Publication
Voucher	1	Images, physical sample, DNA/RNA, tissue	science.mnhn.fr
Environmental data (collected during community sampling)	1	Text data (txt, csv)	
Personal data	1	Text data (txt, csv, PDF, html)	
Campaign data	1	Text data (txt, csv)	BASEXP
Collection data	1	Text data (txt, csv)	JACIM
Sample image	1	Tiff, Jpeg, Png	science.mnhn.fr
Collecting permits	1	pdf	BASEXP
Contact details, prior informed consent	1	Text data (txt, csv, PDF, html)	BASEXP
Bioproject data	2	Text data (txt, xml)	ENA - EBI
Biosample data	2	Text data (txt, xml)	ENA - EBI
Genome size, ploidy and heterozygosity rate estimates	2	Text data (txt, csv)	GoAT
Sequencing experiments (DNA, RNA)	2	FASTQ and BAM	ENA - EBI
Barcode sequence	2	FASTA, SCF files	BOLD
Genome sequence	2	FASTA	ENA - EBI
Genome structural annotation	2	GFF3, FASTA	ENA - EBI
Genome functional annotation	2	TXT, EMBL	ENA - EBI
Expert structural and functional annotation	3	TXT, EMBL	ENA - EBI
Gene trees	3	TXT, TSV, NHX	ATLASea-PORTAL
Species trees	3	TXT, NWK	ATLASea-PORTAL
Comparative genomic datasets (Orthologs gene lists, multiple alignments)	3	TXT, MULTI-FASTA	ATLASea-PORTAL
Specimen metadata (assembly metrics, number of chromosomes, gene statistics)	3	Text data (txt, csv)	GoAT
Sequencing methods	2	Text data (txt, pdf)	protocols.io, PEPR-ATLASea Git dedicated protocols' journals
Data quality control procedures	2	Text data (txt, pdf)	PEPR-ATLASea Git
Statistics of generated data	2,3	Text data (txt, csv)	NA

Assembly workflows	2	Text data (txt, pdf)	PEPR-ATLASea Git
Structural Annotation workflows	2	Text data (txt, pdf)	PEPR-ATLASea Git
Functional Annotation workflows	3	Text data (txt, pdf)	PEPR-ATLASea Git
Comparative Genomics workflows	3	Text data (txt, pdf)	PEPR-ATLASea Git
Ancestral reconstruction workflows	3	Text data (txt, pdf)	PEPR-ATLASea Git
Additional software used within the project will be open sourced as much as possible	1,2,3	Text data	PEPR-ATLASea Git
Test data sets from software developments	2,3	FASTA, fastq, tab	PEPR-ATLASea Git
Pedagogical documentation	1,2,3	Text data (txt, pdf)	PEPR-ATLASea Portal- Moodle ATLASea

Concernant les projets financés dans le cadre du PEPR ATLASea (ex.: projets pilotes financés suite à un appel ouvert), les types d'objets numériques, leur format et leur volume sont propres aux activités de chacun des projets et seront décrits dans les PGD relatifs aux dits projets.

Les données personnelles

Aucune donnée sensible (qui révèle la prétendue origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique) ne sera collectée dans le cadre du PEPR ATLASea.

Des données personnelles non sensibles sont collectées, mais en suivant impérativement le Règlement Général sur la Protection des Données (RGPD : https://www.cnil.fr/fr/reglement-europeen-protection-donnees), avec notamment la recherche de consentement pour la réutilisation (ex. adresses e-mail pour la newsletter ; nom, prénom, affiliation et adresse e-mail pour le formulaire adressé aux ambassadeurs des espèces à séquencer ; CVs et lettres de motivation soumis par les candidats aux postes proposés dans le cadre des recrutements annoncés publiquement par les équipes partenaires du PEPR ATLASea).

D. Protocoles d'échantillonnage sur le terrain (DIVE-Sea)

Des protocoles standardisés seront utilisés pour la collecte des échantillons, incluant les informations de station (localisation géographique), l'habitat, la date, le collecteur, la technique de collecte, la méthode d'identification et l'identificateur.

E. Métadonnées associées aux spécimens (DIVE-Sea)

Chaque spécimen sera accompagné de métadonnées détaillées pour faciliter l'interprétation et la réutilisation des données. Les métadonnées comprendront des informations sur l'espèce, la date, le lieu de prélèvement, le type d'écosystème, les conditions de conservation, etc. Ces données seront stockées dans les bases développées par les partenaires MNHN, IFREMER et EMBRC et accessibles depuis la base de données ATLASea.

F. Protocoles d'extraction d'ADN et de séquençage du génome (SEQ-Sea)

Les protocoles d'extraction d'ADN et d'ARN et de séquençage seront documentés avec précision, en spécifiant les techniques utilisées, les paramètres de séquençage, les kits réactifs, etc. Les protocoles utilisés se rapprocheront des protocoles déjà publiés dans protocole.io

G. Protocoles d'assemblage et annotation des génomes et génomique comparative (SEQ-Sea ; BYTE-Sea)

Les protocoles d'assemblage et d'annotation structurale et fonctionnelle des génomes ainsi que les protocoles de génomique comparative seront également documentés, en spécifiant les techniques, les suites logicielles utilisées et les paramètres d'optimisation.

Les différents protocoles seront mis à disposition de la communauté au sein d'entrepôts git (propres au programme : https://github.com/PEPR-ATLASea; https://gitlab.com/pepr-atlasea).

H. Structure de stockage des données brutes et des données traitées (DIVE-Sea ; SEQ-Sea ; BYTE-Sea)

L'organisation des données repose sur des espaces de stockage caractérisés par des typologies de données à accueillir et des règles de préservation spécifiques incluant un ou plusieurs dispositifs (matériel tolérant aux pannes, sauvegarde court-terme sur disque, sauvegarde long-terme sur bande, sauvegardes sur bande dupliquées et externalisées). Cette architecture de stockage sera mise en place pour conserver les données brutes issues du séquençage, de l'assemblage et de l'annotation structurale par le partenaire CEA. Une architecture de stockage sécurisée (sauvegardée et redondée) sera mise en place pour les données obtenues après les analyses bioinformatiques (annotation fonctionnelle, catalogue de gènes orthologues, analyses phylogénétiques) par le partenaire IFB. Une documentation claire sur l'organisation des données sera communiquée aux partenaires du programme ATLASea en charge de l'exploitation des données produites.

IV. Traitement et analyse des données

A. Traitement des données d'échantillonnage (DIVE-Sea ; SEQ-Sea)

Collecte

Chaque espèce collectée sera identifiée par un parrain taxonomiste.

Étiquetage

Pour chaque contenant (tubes, sachets, etc) une étiquette assurera la traçabilité.

Un numéro de lot unique sous forme de code barre sera attribué à chaque lot d'individu d'une même espèce collecté à une localisation et date identique.

Un numéro unique sous forme de code barre sera attribué à chaque individu du lot prélevé.

Chaque prélèvement de l'individu (entier ou partie disséquée de l'individu) sera conditionné dans un contenant (tube) avec un code unique sous forme de code barre collé sur le contenant.

Les métadonnées de collecte, de l'espèce et du conditionnement seront associées à chaque contenant par son code barre.

La traçabilité des métadonnées et les liens entre ces codes sera assurée lors de la collecte sur le terrain dans un fichier csv pour être ensuite intégrée dans les systèmes d'information du programme ATLASea.

Stockage transitoire et bancarisation

Les échantillons seront soumis à différents protocoles d'extraction, pour permettre l'utilisation de technologie de séquençage longues-lectures ainsi que de données dites "long-range" (Hi-C) permettant de reconstruire les chromosomes. D'autre part, pour chacun des organismes, une extraction et un séquençage des ARNs messagers sera réalisée. Ces données seront utilisées pour la phase d'annotation structurale du génome. Lorsque cela sera possible (matériel suffisamment abondant) un voucher éthanol sera conservé au MNHN et du matériel sera cryopréservé dans une banque de tissus localisée au MNHN.

B. Prétraitement des données de séquençage (SEQ-Sea)

Les données de séquençage brutes seront soumises à un processus de prétraitement pour confirmer la taxonomie des spécimens, éliminer les adaptateurs, filtrer les séquences de faible qualité et éliminer les artefacts de séquençage. Les protocoles suivis sont décrits dans les 2 publications suivantes :

- Alberti, A., Poulain, J., Engelen, S. et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Sci Data 4, 170093 (2017). https://doi.org/10.1038/sdata.2017.93
- Belser, C., Poulain, J., Labadie, K. et al. Integrative omics framework for characterization of coral reef ecosystems from the Tara Pacific expedition. Sci Data 10, 326 (2023). https://doi.org/10.1038/s41597-023-02204-0

C. Assemblage du génome et annotation (SEQ-Sea ; BYTE-Sea)

Les séquences d'ADN obtenues avec une technologie longue-lecture seront assemblées pour reconstruire le génome de chaque espèce. Dans un second temps, des données comme le Hi-C seront utilisées pour faciliter l'obtention de la structure des chromosomes de chaque génome. Une curation et une validation manuelle sera ensuite appliquée à chacun des assemblages avant leur mise à disposition. Pour finir, à partir d'un assemblage, des données de RNA-Seq correspondant et de données externes, une annotation structurale et fonctionnelle sera réalisée pour caractériser et identifier les gènes, les éléments régulateurs et assigner des fonctions.

D. Analyses bioinformatiques spécifiques (diversité génétique, phylogénie, etc.) (BYTE-Sea)

Différentes analyses bioinformatiques seront effectuées, telles que la détermination de la diversité génétique intra- et inter-espèces, la recherche de gènes spécifiques, l'identification de voies

métaboliques, etc. Ces analyses seront réalisées dans le cadre du Projet Ciblé BYTE-Sea ou dans le cadre des projets pilotes lancés en année 3 du programme.

E. Outils et logiciels utilisés

Une liste des outils et logiciels bioinformatiques utilisés pour le traitement et l'analyse des données (voir les sections IV.B, IV.C, IV.D) sera documentée au sein de dépôts git propres au programme ATLASea (https://github.com/PEPR-ATLASea et https://gitlab.com/pepr-atlasea) pour faciliter la reproductibilité et la diffusion des résultats.

F. Données personnelles

Les Projets Ciblés du programme ATLASea collecteront des données personnelles dans la mise en œuvre des diverses activités prévues. La gouvernance d'ATLASea collectera principalement des données personnelles au travers de ces actions de coordination et de communication générale du programme.

La collecte et l'utilisation de données personnelles se fera dans le respect du Règlement Général sur la Protection des Données (RGPD).

Les lettres d'engagement (modèle standard validé par l'ANR repris pour tous les projets d'ATLASea) signées par les partenaires du projet de pilotage et gouvernance et des projets financés font office d'accord de consortium et prévoient les dispositions de protection des données personnelles.

Le paragraphe 9 de l'Annexe II de la Lettre d'engagement stipule que chacune des Parties s'oblige à se conformer à toutes dispositions en vigueur relatives au traitement de données personnelles prévues par les textes législatifs et réglementaires applicables en France, par le droit de l'Union européenne, y compris le règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (« RGPD »), la loi n°78-17 modifiée du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (loi Informatique et libertés) et de toute réglementation nationale prise en application, concernant les traitements de données à caractère personnel mis en œuvre dans le cadre du Projet (ci-après la « Réglementation »). Chaque Partie est responsable des traitements, au sens de la Réglementation, qu'elle met en œuvre seule.

V. Gestion de la qualité des données

A. Contrôle qualité lors de la collecte des échantillons, du séquençage des génomes et des traitements des données bioinformatiques

Une attention particulière sera portée sur le choix et la qualité des standards et métadonnées qui sont essentiels pour assurer la compréhension, la fiabilité et la réexploitation des données du programme.

Standard données d'échantillonnage : Darwin-core (https://www.gbif.org/darwin-core) et (https://www.gbif.org/es/data-quality-requirements-sampling-events)

Standard données de séquences de genomes : Standards promus par le GDC (Genomic Data standards du type https://docs.gdc.cancer.gov/Data/Data_Model/GDC_Data_Model/)

Standard données de séquences de métabarcode : (https://mbmg.pensoft.net/article/58056/)

Le contrôle de la qualité des données sera enrichi dans une version ultérieure du Plan de Gestion de Données.

B. Mesures d'assurance qualité mises en place

Des procédures d'assurance qualité seront appliquées tout au long du programme pour détecter et corriger les erreurs potentielles et garantir la fiabilité des données par exemple :

Échantillonnage : La traçabilité des informations de collecte est assurée tout au long de la chaîne de tri du spécimen jusqu'à sa mise en tube en conformité avec les procédures internes.

Conservation des échantillons : Traçabilité par code barre, et informatisation des données des échantillons.

Assignation taxonomique : Une première assignation sera réalisée par des experts taxonomistes présents lors de l'échantillonnage. Une analyse par barcoding moléculaire sera réalisée par la suite sur les échantillons pour confirmer le positionnement phylogénétique.

Extraction d'ADN : voir les méthodes décrites dans Alberti, et al. (2017). https://doi.org/10.1038/sdata.2017.93 et Belser, C. et al. (2023). https://doi.org/10.1038/s41597-023-02204-0

Séquençage : Les données issues des expériences de séquençage suivront une procédure qualité utilisée dans des projets et décrite dans deux publications (Alberti, et al. (2017) et Belser, C. et al. (2023)). Les outils utilisés pour ce suivi qualité sont disponibles librement et listés dans les deux publications.

Ces données de séquençage permettront aussi de valider certaines caractéristiques des génomes, comme leur taille ou encore leur taux d'hétérozygotie (GenomeScope, LocoGSE). Une fois les assemblages réalisés, plusieurs critères seront évalués pour permettre l'attribution d'un niveau de qualité de chacun des assemblages. On peut citer par exemple, la contiguité des assemblages (NGX, LGX, auN), la qualité du consensus obtenu (Merqury), la complétion du génome (Merqury) ou encore la complétion en gènes (BUSCO). D'autre part, lorsque cela sera possible, ils seront comparés à des génomes d'espèces proches. De la même façon, la qualité des catalogues de gènes obtenus sera évaluée, notamment en termes de complétion (BUSCO) et de conservation des protéines orthologues (Orthofinder).

VI. Stockage et sauvegarde des données

A. Infrastructure de stockage des données

Le stockage et la sauvegarde des données tout au long du programme seront conformes aux politiques définies par les plateformes qui hébergent les données sur la durée du projet. Un soin particulier sera apporté à sensibiliser les partenaires à la nécessité de disposer de moyens de préservation pérennes avec des moyens dédiés au stockage et à la sauvegarde des données. Une infrastructure de stockage spécifique et adaptée aux données scientifiques sera mise en place par les partenaires des Projets Ciblés SEQ-Sea et BYTE-Sea pour accueillir les données générées, en tenant compte de leur volume potentiellement important. Des démarches seront entreprises pour proposer des solutions d'hébergement des données s'appuyant sur l'infrastructure dédiée au programme dans les cas où celles-ci n'existent pas pour des jeux de données qu'il n'est pas possible de collecter et ou acquérir à nouveau en cas de perte.

Des données opérationnelles (y compris des données personnelles et administratives) seront stockées et partagées via Osmose - une plateforme collaborative des communautés professionnelles de l'État. Plusieurs plateformes d'échange sécurisé des fichiers ont été testées pour assurer l'accès aux données protégées aux partenaires du programme et gérer leurs droits d'accès. Osmose semble être la mieux adaptée aux besoins du programme.

B. Politique de sauvegarde des données

Une politique de sauvegarde régulière des données scientifiques sera établie pour minimiser le risque de perte de données en cas d'incident technique ou de panne.

Les données du Projet Ciblé WHEEL-Sea seront sauvegardées dans la plateforme collaborative Osmose.

Les données du Projet Ciblé DIVE-Sea seront sauvegardées dans le système d'information du MNHN.

Les données brutes du Projet Ciblé SEQ-Sea, une fois générées, seront mises à disposition sur disque et une double copie sera réalisée sur bande, permettant la restauration des données (notamment en cas de sinistre). Après validation de l'assemblage et réalisation de l'annotation, les données issues du séquençage, l'assemblage et le catalogue de gènes seront déposés d'une part dans les banques publiques (EBI) et également mises à disposition à travers le portail ATLASea. Les données sur l'espèce collectée (biosample) seront alors également soumises à l'EBI selon une trame ERC (trame de métadonnées obligatoires et optionnelles de l'EBI).

Les données du Projet Ciblé BYTE-Sea seront sauvegardées dans 3 centres distincts (Paris, Rennes et Plouzané) pour assurer leur disponibilité.

C. Sécurité et confidentialité des données

Des mesures de sécurité seront mises en place pour protéger les données contre l'accès non autorisé et garantir la confidentialité des informations sensibles.

VII. Partage et accès aux données

A. Politique de partage des données

Une politique de partage des données sera définie pour encourager le partage des données avec la communauté scientifique tout en respectant les droits des auteurs et les contraintes éthiques.

Les ensembles de données seront déposés dans un certain nombre de référentiels de confiance, en fonction de la nature et de l'objectif spécifique des données.

Ces référentiels de confiance sont mis en œuvre selon des protocoles de communication Internet standards et intègrent un grand nombre des caractéristiques souhaitées pour ce type de référentiels, telles que la gestion de l'identité et l'authentification, une documentation étendue et la prise en charge des normes approuvées par la communauté. Par exemple ELIXIR life science login (ELIXIR LS Login) (https://elixir-europe.org/AAI-migration).

Aucun transfert de données à caractère personnel vers un pays tiers ou à une organisation internationale n'est pas prévu dans le cadre du programme ATLASea.

B. Licences et restrictions d'accès

Les conditions de partage, y compris les licences et les restrictions d'accès, seront clairement définies au cours du programme pour les données et développement logiciels qui pourraient nécessiter des restrictions particulières. Les développements logiciels s'appuieront sur une licence d'utilisation adaptée au droit français et en accord avec la politique des instituts porteurs CEA et MNHN et internationalement acceptée comme CC-BY ou CC-BY-SA.

C. Plateformes de partage de données utilisées

Les plateformes ou bases de données spécifiques sur lesquelles les données seront déposées et rendues accessibles seront identifiées.

Voici une liste non exhaustive des principaux référentiels spécifiques à un domaine qui sont utilisés dans le cadre du programme ATLASea :

- ATLASea portal (https://portal.atlasea.fr)
- European Nucleotide archives (ENA)
- Global genome Biodiversity network (GGBN)
- Global Biodiversity Information Facility (GBIF)
- Collaborative Open Plant Omics (COPO)
- Les collections du Muséum Nationale d'Histoire Naturelle (MNHN)
- Le Système d'Information de l'Inventaire du Patrimoine Naturel (SINP)
- Portails de biobanques (RCC, …).
- Genome on a Tree (GoAT)
- World Register of Marine Species (WORMS)
- Software Heritage
- WorkflowHub

Si certains types de données n'ont pas leur place dans les référentiels susmentionnés d'autres référentiels généraux seront utilisés, tels que ceux énumérés ci-dessous :

- Inventaire des Données de la Recherche en environnement et Sociétés InDoRES
- HAL (https://hal.science/ATLASEA)
- Zenodo

D. Contraintes Juridiques

Dans l'état actuel de l'inventaire des données manipulées dans le cadre du programme ATLASea, et en dehors des données à caractère personnel, traitées dans la section III.C, aucune donnée à caractère sensible n'a été identifiée.

Les modalités d'accès aux produits de recherche (données, logiciels) doivent être définies au niveau du consortium et consignées dans l'accord de consortium ou les lettres d'engagement en faisant office. Ces modalités devront tenir compte des préconisations des financeurs en la matière, en favorisant tant que faire se peut l'accès ouvert aux produits de recherche, notamment au travers du choix des licences qui leur seront applicables.

VIII. Archivage à long terme

A. Stratégie d'archivage des données

Une stratégie d'archivage à long terme des données sera élaborée pour garantir leur disponibilité et leur pérennité à long terme. Toutefois en dehors des entrepôts de référence, l'infrastructure qui va être mise en place dans le Projet Ciblé BYTE-Sea n'a pas vocation à archiver les données du programme ATLASea. L'infrastructure nationale IFB contribuera à indexer ces données pour favoriser leur moissonnage dans le catalogue datagouv.

Ce point sera précisé lors d'une prochaine itération du Plan de Gestion de Données.

B. Choix des archives et des entrepôts de données

Les archives et les entrepôts de données appropriés seront choisis pour assurer une conservation adéquate des données. Sous condition de conformité avec les termes de l'accord de consortium, l'usage d'entrepôts sera préconisé. Pour certaines disciplines scientifiques, de tels entrepôts existent déjà (par exemple, la plateforme ENA - European Nucleotide Archive proposée par l'EBI pour les données de type séquence, la plateforme (Eur)OBIS - Ocean Biodiversity Information System pour les données de type biodiversité), et seront donc retenus. Les campagnes d'échantillonnage seront identifiées dans les référentiels des partenaires MNHN (BASEXP) et IFREMER (SISMER) (ex. : SeaNOE (https://www.seanoe.org/) opéré par le pôle Océan - ODATIS de l'IR Data Terra dédié aux données marines ou Sextan (https://sextant.ifremer.fr/).

Le paysage des entrepôts pour les données de type imagerie reste encore à éclaircir. Ce travail sera mené tout au long du programme ATLASea.

Pour des données orphelines, il pourra être fait appel à des entrepôts plus généralistes (ex. : SeaNOE ou Sextan), allant jusqu'au dataverse Data Gouv.

Pour les logiciels, et toujours sous condition des termes de l'accord de consortium, une plateforme assurant un suivi des versions sera préconisée (GitLab), avec un référencement dans la plateforme du Software Heritage Initiative.

C. Métadonnées d'archivage

Les métadonnées nécessaires pour faciliter la découverte et la réutilisation des données seront fournies lors de l'archivage en utilisant les standards préconisés par la communauté qui seront précisés dans une prochaine itération du PGD.

IX. Gestion des versions et des mises à jour

A. Gestion des versions des données

Un système de gestion des versions notamment des protocoles et des versions des assemblages sera mis en place pour suivre l'évolution des données tout au long du programme, en permettant de conserver un historique des modifications apportées.

B. Gestion des mises à jour des données

Une procédure pour la gestion des mises à jour des données sera établie, en précisant les responsabilités et les délais pour mettre à jour les enregistrements de données lorsque cela est nécessaire.

C. Documentation des changements

Tous les changements apportés aux données seront dûment documentés, y compris les raisons des modifications et les auteurs responsables.

X. Responsabilités et rôles

A. Identification des responsables de la gestion des données

Les membres de l'équipe impliqués dans la gestion des données au sein des différents projets ciblés seront identifiés, et leurs rôles et responsabilités seront clairement définis.

Projets Ciblés	**Noms**
PC 0 : WHEEL-Sea	Hugues Roest Crollius et Kamil Szafranski
PC 1 : DIVE-Sea	Line Le Gall et Franck Bellugeon
PC 2 : SEQ-Sea	Claude Scarpelli et Jean-Marc Aury
PC 3 : BYTE-Sea	Erwan Corre et Patrick Durand

B. Rôles et tâches des membres de l'équipe

Les rôles spécifiques des membres de l'équipe, tels que les responsables de l'échantillonnage, du séquençage, de l'analyse bioinformatique et de la gestion des données, seront énoncés.

XI. Formation et sensibilisation

A. Formation des membres de l'équipe aux bonnes pratiques de gestion des données

Le PEPR ATLASea dispose d'une forte composante en personnel spécialisé en informatique et bioinformatique tout au long du processus de collecte, production et partage des données, déjà fortement sensibilisés aux questions de traitement, partage, sécurisation des données. Nous nous assurerons au fil du projet qu le personnel puisse completer et renforcer ces compétences en s'appuyant notamment sur les formations dispensées par l'infrastructure nationale IFB. Si de nouvelles réglementation se mettent en place, une communication adaptée sera assurée afin de diffuser l'information aux personnes responsables.

B. Sensibilisation à l'importance de la gestion des données

Une sensibilisation sera réalisée auprès de tous les membres de l'équipe pour souligner l'importance de la gestion rigoureuse des données pour le succès du programme.

XII. Budget et ressources

A. Estimation des coûts liés à la gestion des données et aux ressources matérielles et logicielles nécessaires

Une estimation des coûts liés à la gestion des données, y compris l'infrastructure, les logiciels, la formation, etc., a été évaluée au sein du programme et a été financée sur la durée de ce programme. Au-delà de la durée du programme il sera nécessaire d'évaluer le coût de maintien en conditions opérationnelles et de jouvence de l'infrastructure et solliciter des financements récurrents.

XIII. Éthique et consentement

A. Conformité aux règles éthiques et réglementations en matière de recherche

Le projet sera mené conformément aux règles éthiques et aux réglementations applicables en matière de recherche, en obtenant les autorisations nécessaires. La collecte et l'utilisation de données personnelles se fera dans le respect du Règlement Général sur la Protection des Données (RGPD).

A date, seuls certains jeux de données ou produits de recherche ont été identifiés comme étant susceptibles d'être concernés par la protection des données à caractère personnel.

La collecte et l'utilisation de données personnelles se fera dans le respect du Règlement Général sur la Protection des Données (RGPD). Il s'agit principalement des jeux de données relevant du périmètre des Projets Ciblés DIVE-Sea et SEQ-Sea qui incluent des traces numériques pouvant contenir des données à caractère personnel (ex. : les noms des personnes ayant participé à la collecte, l'acquisition et/ou l'analyse des données, les parrains taxonomistes).

Les responsables de ces bases de données seront sensibilisés aux bonnes pratiques de la protection des données à caractère personnel, afin d'assurer la conformité de ces bases avec les réglementations en vigueur.

Les données personnelles incluses dans les métadonnées sont le nom de la personne de contact, l'adresse électronique de la personne de contact, l'échantillon prélevé par la personne (nom), la personne responsable du stockage (nom). Seuls le nom et l'adresse électronique de la personne de contact seront rendus publics. Avant la publication des données personnelles avec les métadonnées, les personnes concernées seront invitées à remplir un formulaire GDPR et à indiquer si elles autorisent que leurs données personnelles apparaissent dans des bases de données publiques avec les métadonnées. Ce programme n'impliquant pas d'échantillons de tissus humains, les données personnelles ne seront en aucun cas traitées. Les données personnelles ne sont présentées dans les métadonnées qu'à des fins de traçabilité, de valorisation du travail ou de communication à la personne responsable et d'assurance que toutes les procédures ont été respectées.

Toutes les données incluses dans l'ensemble de données génomiques sont la propriété intellectuelle de ATLASea (des partenaires de ATLASea). Les métadonnées des collectes sont une propriété conjointe de ATLASea et des stations partenaires. Les données et métadonnées générées dans le cadre de ATLASea seront librement accessibles.

L'APA est un cadre pour l'accès et l'utilisation des ressources génétiques. Le Projet Ciblé DIVE-Sea contactera les référents pour chaque territoire et déclarera ce que ATLASea à l'intention de collecter, stocker et distribuer des échantillons et des informations de séquences numériques. Les règles APA seront suivies par chacun des partenaires collecteurs et la documentation connexe sera stockée dans le système de fichier ATLASea maintenu à l'ENS et dans la base de données ATLASea.

Ce point sera mis à jour lors d'une prochaine itération du Plan de Gestion de Données.

B. Obtention du consentement éclairé pour la collecte et l'utilisation des échantillons

Toutes les procédures de collecte d'échantillons seront effectuées en respectant le consentement éclairé des parties concernées.

Ce point sera mis à jour lors d'une prochaine itération du Plan de Gestion de Données.

XIV. Calendrier

A. Échéancier des différentes étapes du Plan de Gestion des Données

Un calendrier détaillé basé sur les tableaux de Gantt construit lors de la rédaction initiale des projets ciblés sera établi, en indiquant les dates clés pour chaque étape de la gestion des données dans une prochaine itération du PGD.

B. Révisions régulières du plan en fonction de l'avancement du programme

Le Plan de Gestion des Données sera révisé régulièrement pour tenir compte de l'évolution du programme et des besoins émergents.