ATLASea

- **Un atlas des génomes marins**

Challenges liés à l'acquisition et la gestion des données au sein du PEPR ATLASea

Annie LEBRETON

Atelier GFM

16/09/2025

Photos crédits: Arwan Amice – LEMAR – CNRS Photothèque

ATLASea – un atlas des génomes marins

~ 370 000 espèces marines eucaryotes sont actuellement décrites *(marinespecies.org)*
dont ~50 000 en ZEE française
mais peu de génomes de bonne qualité sont disponibles

PEPR ATLASea
2023 - 2030

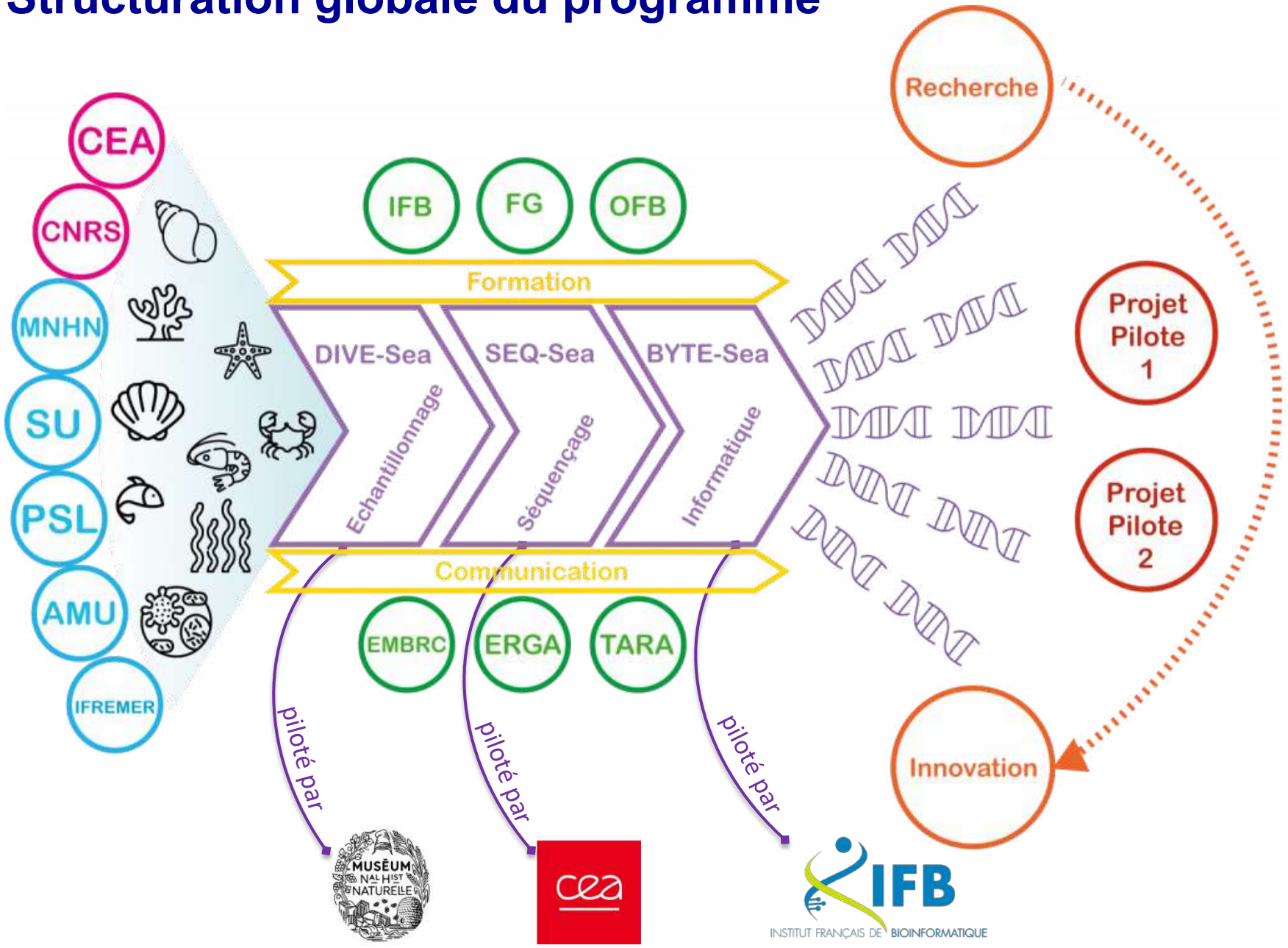
Objectif: Séquencer 4 500 espèces eucaryotes marines des eaux françaises

Animaux, Plantes, Chromistes, Protozoaires, Champignons

ZEE Française

Phylopics images from phylopic.org
by T. Michael Keesey., Nix Illustration. Harold N Eyster. Maxime Dahirel. Matt Crook.

Structuration globale du programme



De nombreux partenaires

4 projets ciblés :

- *DIVE-Sea*
- *SEQ-Sea*
- *BYTE-Sea*
- *WEEL-Sea (gouvernance)*

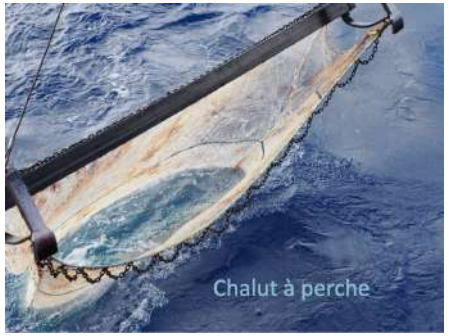
Collecte des espèces

Origine des échantillons

- Missions
- Stations
- Collections
- Ambassadeurs

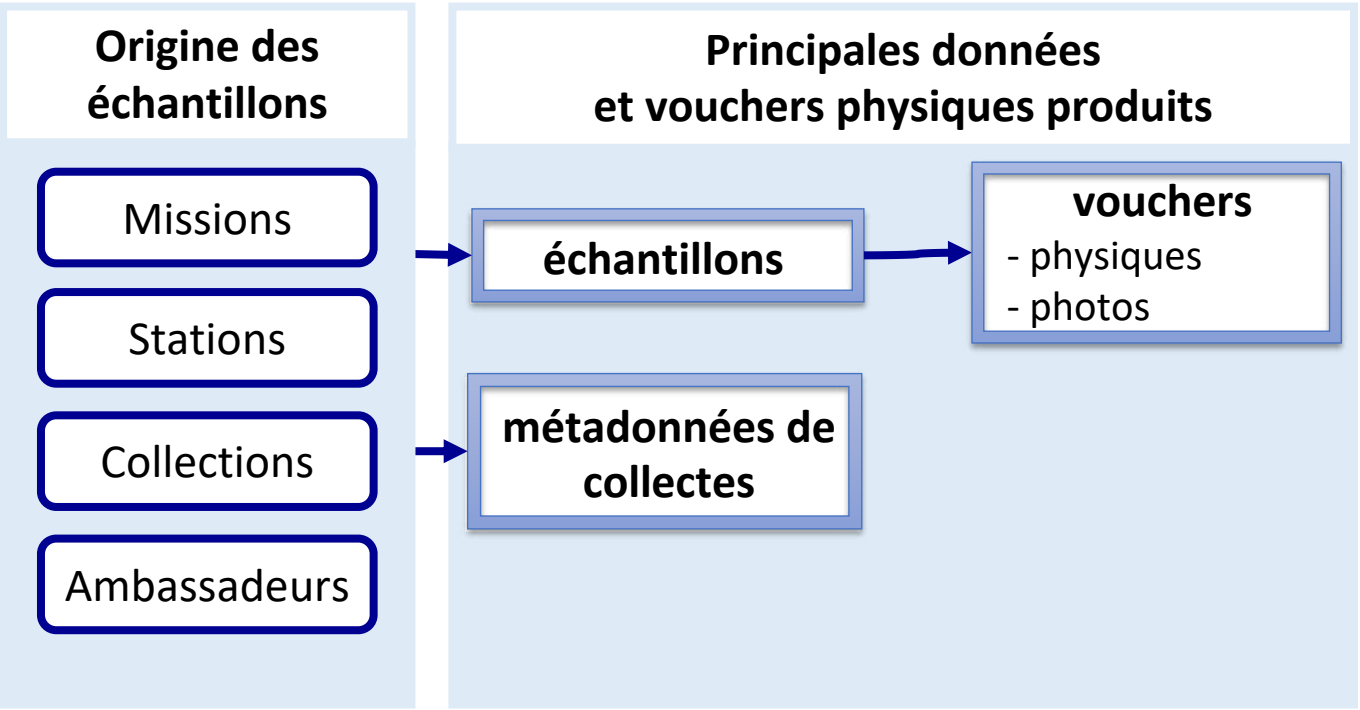
Grande diversité d'espèces
(taxonomie, taille, mode de vie...)

Grande diversité de méthodes d'échantillonnage
(chalut, plongée, pêche à pied...)

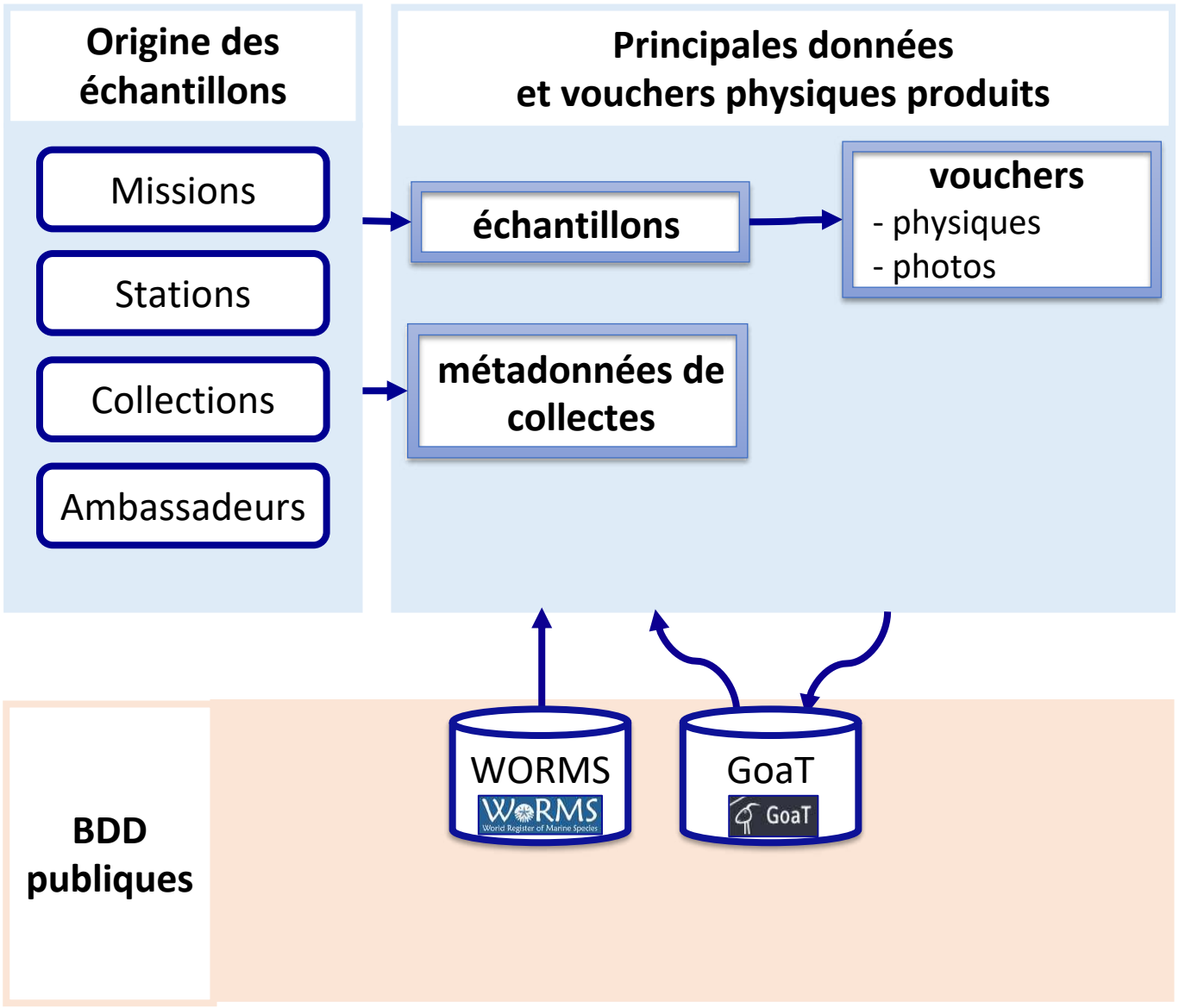


credits photos: team ATLASea

Principaux délivrables liés à l'échantillonnage



Principaux délivrables liés à l'échantillonnage

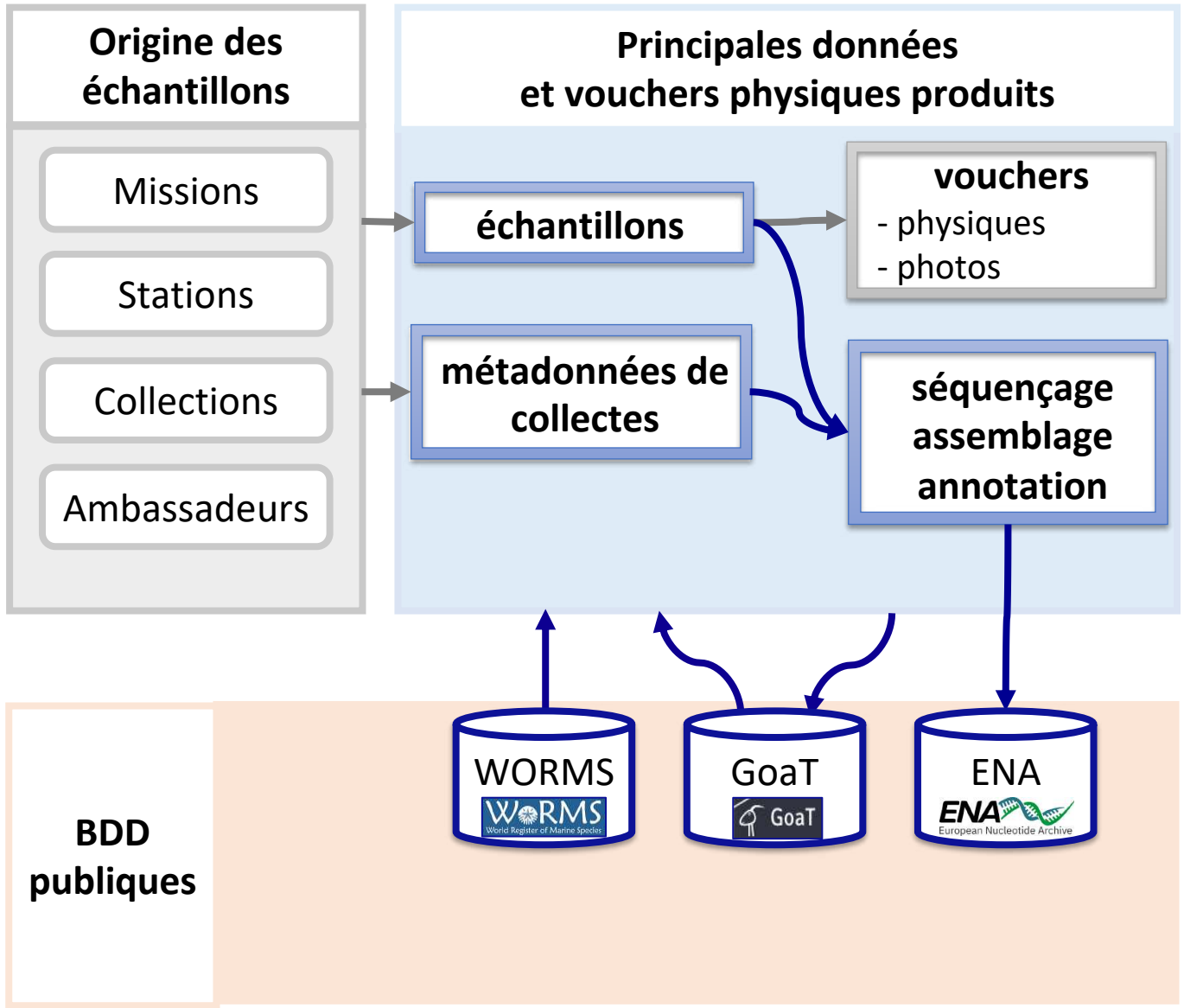


Seules les espèces qui ne sont pas entrées dans un processus de séquençage dans un autre projet (ex. dTOL) sont utilisées pour ATLASea

Genomes on a Tree (GoaT)
=> **suivi des espèces qui sont en cours de séquençage dans le monde**
Outil conçu dans le cadre d'Earth Biogenome Project (EBP)

Wold Register of Marine Species
Classification et catalogue des espèces marines

Principaux délivrables liés à la production de génomes



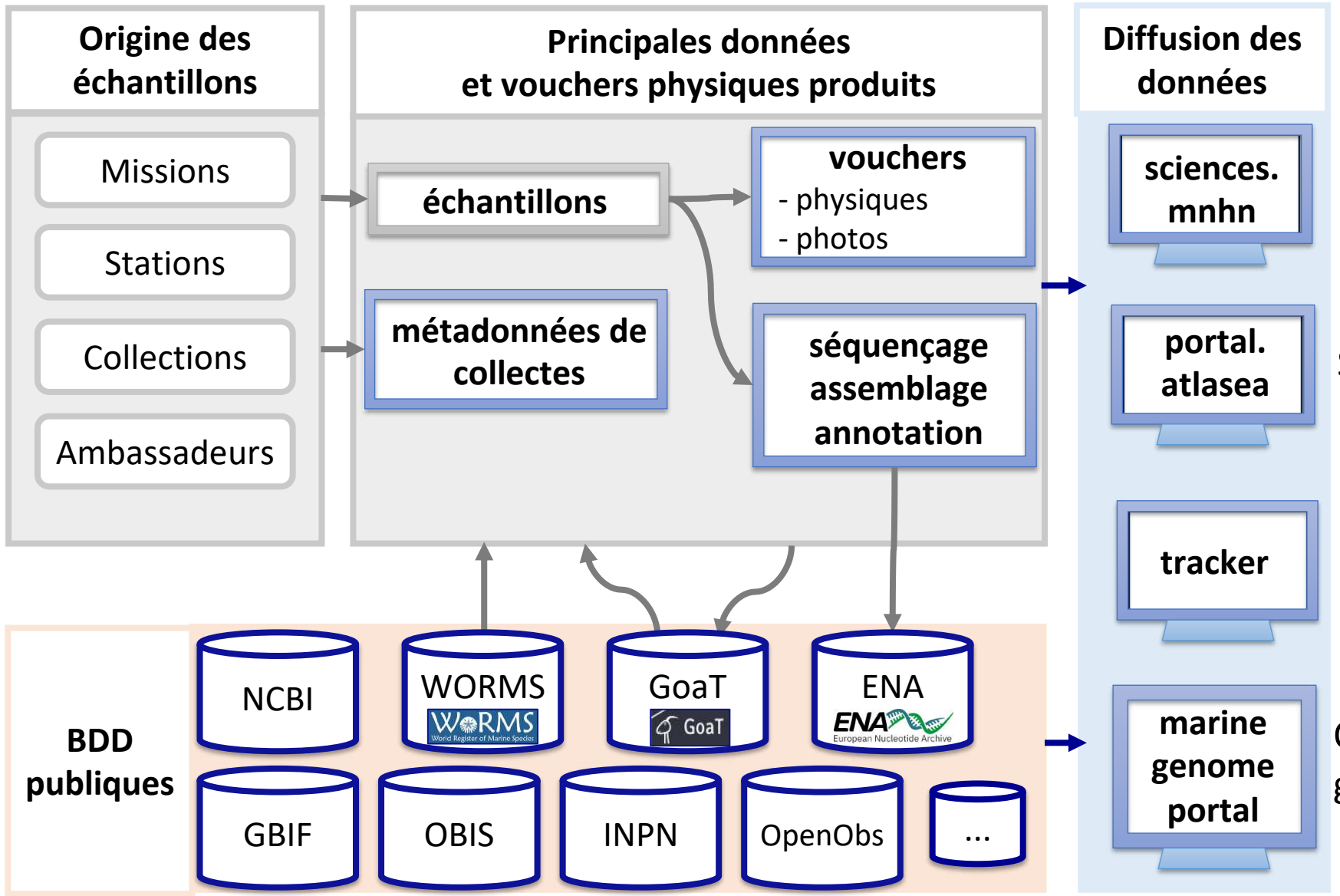
- Alimentation de GoaT fréquente
  (enjeux vis à vis des autres programmes de séquençage)

- Soumission à l'ENA dès que possible

ENA European Nucleotide Archive

Base de données de référence mondiale archivant les produits de séquençage et informations dérivées
(e.g. genomes, annotation genomiques)

Diffusion et exploitation des données



Informations sur les expéditions, les specimens en collection etc.
pérenne

Suivi public

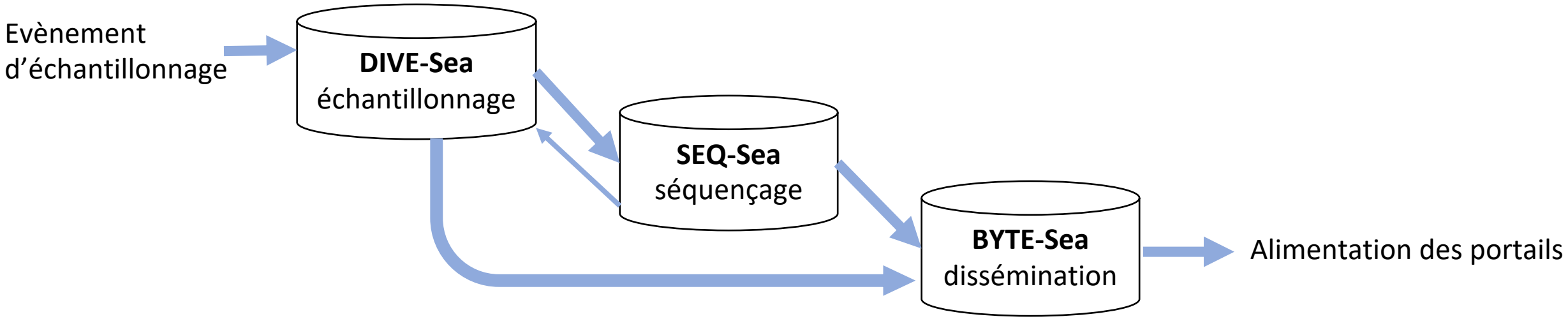
Suivi privé

durée de vie du programme (8 ans)

Outil permettant d'analyser les génomes d'organismes marins
pérenne

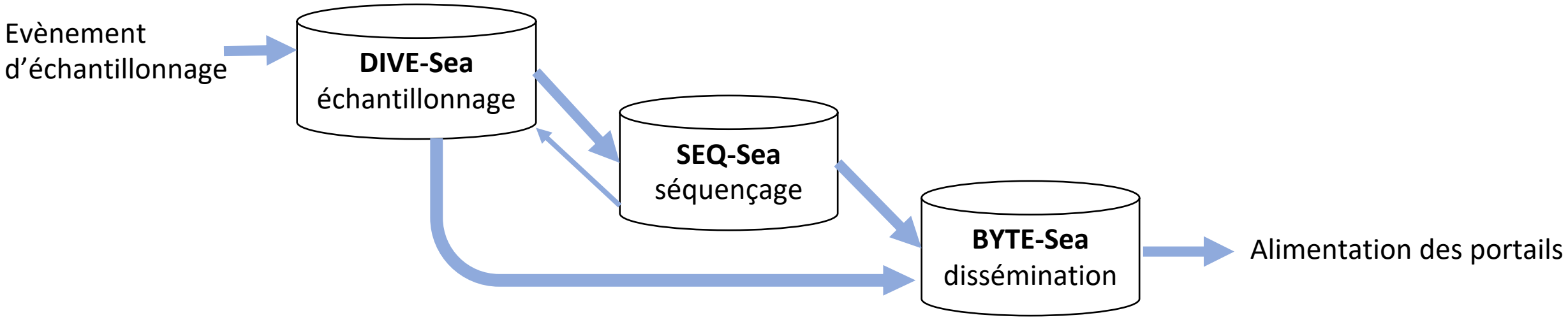
Flux de données principaux entre les systèmes d'informations du programme

Chaque projet ciblé a une base de données principale et est maître d'un ensemble données



Flux de données principaux entre les systèmes d'informations du programme

Chaque projet ciblé a une base de données principale et est maître d'un ensemble données



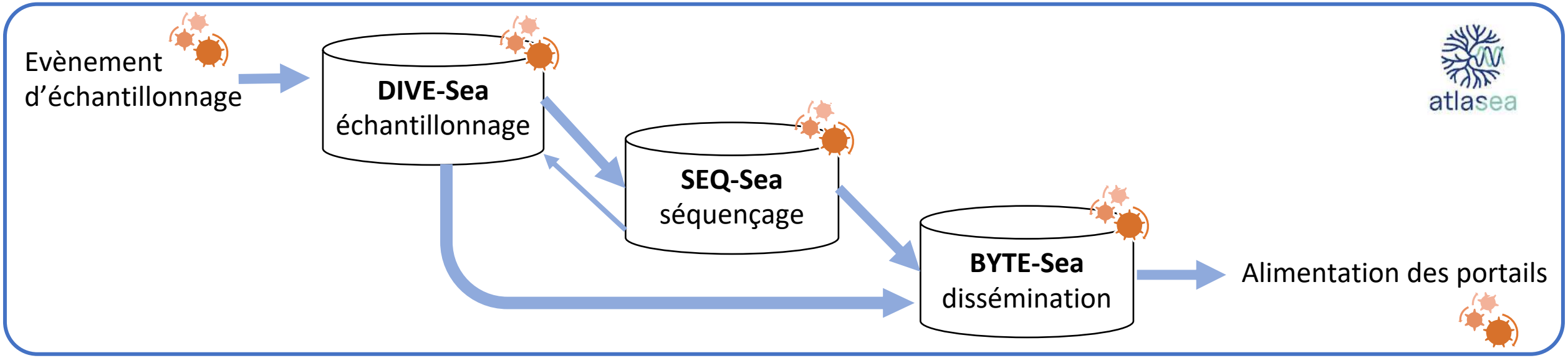
Actuellement en cours d'année 3 (sur 8) : les systèmes d'informations (SI) sont en cours de développement et automatisation...

...mais la production de données a commencé depuis le début du programme

Quelques challenges associés a la gestion des flux de données

Réception de données à flux tendu

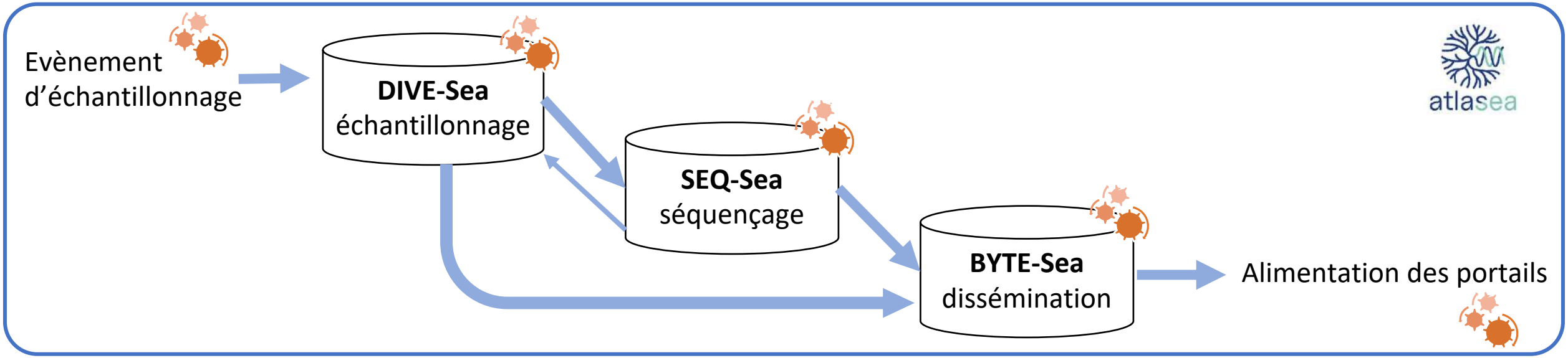
SI en cours de maturation, dont certains ne relèvent pas de notre périmètre d'action (partenaires ou concurents !)



Quelques challenges associés a la gestion des flux de données

Réception de données à flux tendu

SI en cours de maturation, dont certains ne relèvent pas de notre périmètre d'action (partenaires ou concurents !)



NCBI

- Déclaration des projets de séquençage
- MàJ de nom d'espèces

ENA

Procédure d'intégration des données soumises

- temps requis par l'ENA soumission-accès sur la BDD
- implique des changements de formats/identifiants

ENSEMBL

Annotations par ensembl rapid release ?
Sur quel calendrier ?
(ATLASea génère ses propres annotations en parallèle)

Quelques challenges associés à la collecte des métadonnées en mission

Objectif

Tracer les informations sur
- le lieu de collecte
- la collecte
- l'échantillon
- comment l'échantillon a été traité
- qui a fait quoi

Quelques challenges associés à la collecte des métadonnées en mission

Objectif

Tracer les informations sur
- le lieu de collecte
- la collecte
- l'échantillon
- comment l'échantillon a été traité
- qui a fait quoi

Contraintes terrain

- Temps limité, échantillonnage à flux tendu
- Contrainte de temps pour la préservation des tissus (chaine de froid)
- Beaucoup de cas particuliers

Contraintes outil

compréhensible et adaptable à la volée

Quelques challenges associés à la collecte des métadonnées en mission

Objectif	Contraintes terrain	Contraintes outil
Tracer les informations sur - le lieu de collecte - la collecte - l'échantillon - comment l'échantillon a été traité - qui a fait quoi	- Temps limité, échantillonnage à flux tendu - Contrainte de temps pour la préservation des tissus (chaine de froid) - Beaucoup de cas particuliers	compréhensible et adaptable à la volée
	- Plusieurs équipes en parallèle - Pas d'accès internet (ou très contraint)	challenge associé à la synchronisation/fusion des données
	- les équipes changent d'une mission à l'autre, certains participants ne sont impliqués que dans une seule mission	formation à faire au début de chaque mission

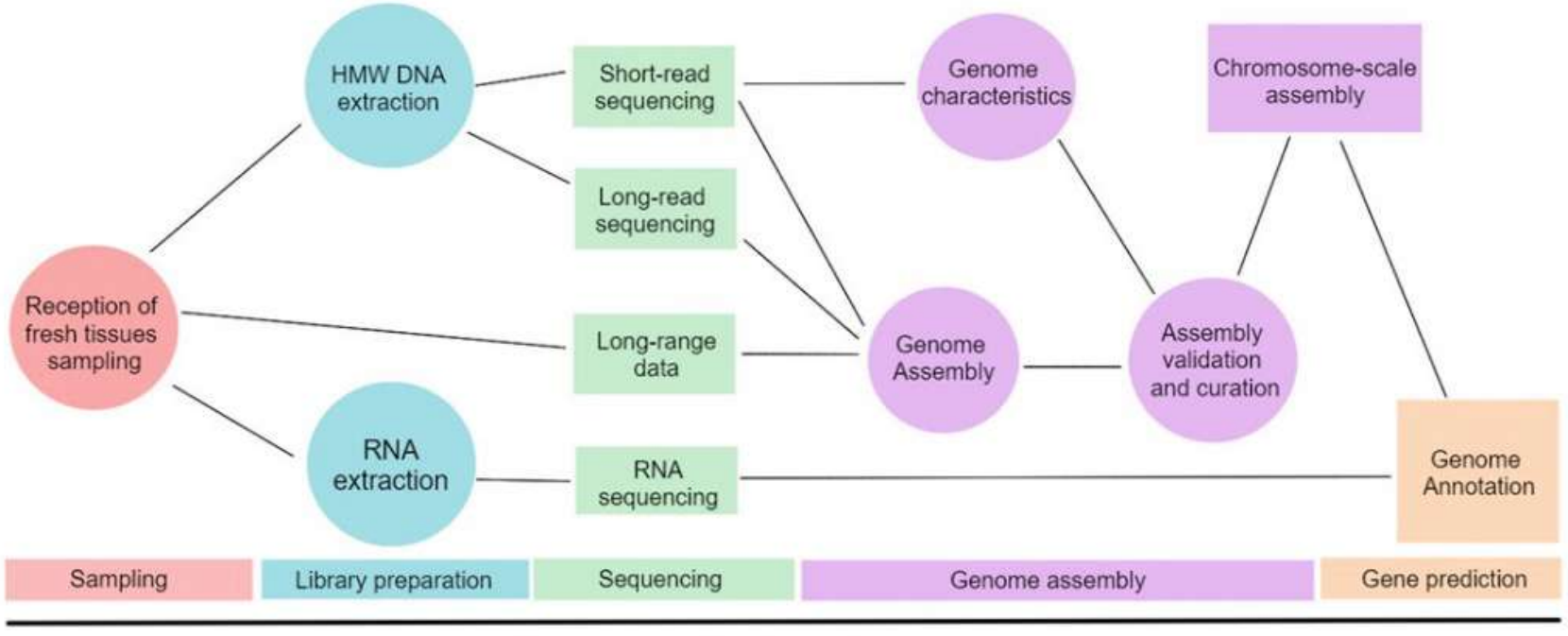
Quelques challenges associés à la collecte des métadonnées en mission

Objectif	Contraintes terrain	Contraintes outil
Tracer les informations sur - le lieu de collecte - la collecte - l'échantillon - comment l'échantillon a été traité - qui a fait quoi	- Temps limité, échantillonnage à flux tendu - Contrainte de temps pour la préservation des tissus (chaine de froid) - Beaucoup de cas particuliers	compréhensible et adaptable à la volée
	- Plusieurs équipes en parallèle - Pas d'accès internet (ou très contraint)	challenge associé à la synchronisation/fusion des données
	- les équipes changent d'une mission à l'autre, certains participants ne sont impliqués que dans une seule mission	formation à faire au début de chaque mission

en cours de collecte	CardObs - descriptions des sites de collectes - collecte des métadonnées
fin de journée	- script pour détecter les plus gros problèmes de remplissage - fusion des documents des différents groupes - compilation des données pour suivi mission

Developpement d'un outil *en cours* :
- compiler les données en cours de mission
- vérifier les données rentrées
  (set de règles de validations)
- requête des données en cours de mission

Quelques challenges associés au séquençage



Full traceability of the different stages

Nécessite des protocoles et méthodologies adaptés pour les différents types d'organismes

=> Echantillons tests

Tracer tout ce qui a été réalisé et automatiser les processus :

rythme attendu 10 génomes / semaine fin 2025

Quelques challenges associés au suivi des métadonnées des échantillons

Une histoire de nom...

Echantillonnage

2022

crypticus sp.1

credits photo: team ATLASea

(noms d'espèces non contractuels)

Reception des échantillons

...

crypticus sp. 4

crypticus sp.5

barcode + vérification taxonomie

crypticus identificatus

lien avec les données publiques

REF tax.

WORMS

NCBI

nom NCBI (obsolete)

crypta idata

* les informations sont également susceptibles de changer si une précision/correction est dédectée

- La taxonomie NCBI est la référence pour le dépôt de données génomiques
- Clef qui lie à d'autres sites e.g. GBIF

Quelques challenges associés au suivi des métadonnées des échantillons

Une histoire de nom...

Echantillonnage

2022

crypticus sp.1

credits photo: team ATLASea

(noms d'espèces non contractuels)

Reception des échantillons

...

crypticus sp. 4

crypticus sp.5

barcode + vérification taxonomie

crypticus identificatus

lien avec les données publiques

REF tax.

WORMS

NCBI

nom NCBI (obsolete)

crypta idata

* les informations sont également susceptibles de changer si une précision/correction est dédectée

Sur les portails:

- **Affiche le nom WORMS (à jour)** ***crypticus identificatus***
- **utilise le nom NCBI (obsolète)** ***crypta idata***

pour faire le lien avec d'autres BDD externes

- La taxonomie NCBI est la référence pour le dépôt de données génomiques
- Clef qui lie à d'autres sites e.g. GBIF

On doit être en capacité de suivre les changements de noms depuis l'évènement de collecte si besoin (cas d'interrogation sur les métadonnées de collecte)

Quelques challenges associés au volume et diversité des données

Suivi et automatisation des étapes à tous niveaux : établir des procédures (SOP), documenter les retours d'experience

Quelques challenges associés au volume et diversité des données

Suivi et automatisation des étapes à tous niveaux : établir des procédures (SOP), documenter les retours d'experience

Infrastructure permettant l'organisation des données, leur stockage, les calculs et le déploiement d'outils (browsers, sites webs)

stockage e.g. les génomes (compressés): 9T en fin de projet
sites webs : e.g. questionnement autour de la sécurité des sites

Quelques challenges associés au volume et diversité des données

Suivi et automatisation des étapes à tous niveaux : établir des procédures (SOP), documenter les retours d'experience

Infrastructure permettant l'organisation des données, leur stockage, les calculs et le déploiement d'outils (browsers, sites webs)

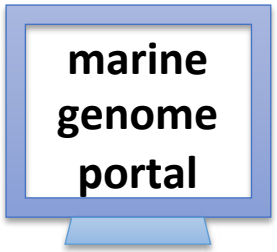
stockage e.g. les génomes (compressés): 9T en fin de projet
sites webs : e.g. questionnement autour de la sécurité des sites

Identifier des outils adaptés

Un grand nombre d'outils bioinformatiques ne permettent pas d'exploiter un si grand nombre de données

(cas du marine genome portal)

=> tests d'outils et méthodologies qui supporteraient la mise à l'échelle.



Outil permettant d'analyser les génomes d'organismes marins

pérenne

Quelques challenges associés à l'implication d'un grand nombre de personnes

Beaucoup de réunions !

Communication au sein du groupe
=> Rocket.Chat
=> Réunions mensuelles

Communication entre les groupes
=> Hackathons
=> AG

Répartition des tâches
=> Création de groupes de travail avec des réunion récurentes

Etablir la localisation et l'organisation des données et documents de suivi
=> MàJ avec gestion des historiques
=> Etablir des procédures pour avertir qu'il y a une MàJ

Conserver le lien avec l'évolution des référentiels métier
=> Groupes de travail sur l'établissement de normes (e.g. ERGA)
=> points de contact pour les SI avec lesquels on interragit souvent (e.g. GoaT/ENA/NCBI/WORMS...)

Conclusion

Programme ATLASea (2023-2030)

Collecte => séquençage => diffusion de 4500+ génomes marins

2023 -> sept. 2025 :

5 campagnes d'échantillonnage

+ 1400 échantillons dans les pipelines atlasea

~100 assemblages de génome publiés

Quelques retours suite à l'expérience d'ALTASea

- Tout tracer et documenter : la création des documents et la reprise d'historique est souvent lourde mais une information manquante peut rendre l'échantillon inexploitable

- Essayer de planifier les données et formats d'entrée et de sortie des différentes sous parties du projet pour faciliter les développements parallèles

- Mettre l'accent sur la communication entre les personnes impliquées

Le programme exploratoire de recherche ***« ATLASea : Atlas des génomes marins »*** *et ses projets ciblés :*

- ***WHEEL-Sea (gouvernance)***
- ***DIVE-Sea***
- ***SEQ-Sea***
- ***BYTE-Sea***

*bénéficient d'une aide de l'État gérée par l'****Agence Nationale de la Recherche (ANR)*** *dans le cadre du plan d'investissement d'avenir* ***« France 2030 »****.*