MODELCET ATL

Conception d'un modèle prédictif de distribution et d'abondance des cétacés en Manche et en Atlantique

Description des méthodes de collecte, de compilation, et d'harmonisation des données

Auteurs : Yves Le Bras, Cindy Gidoin, Marion Pillet, Floriane Plard, Matthieu Authier

UO 1.1.1 – Version du 30/09/2025



Image générée par une Intelligence Artificielle

La Rochelle Université

cnrs

PELAGIS UAR 3462 - LRUniv - CNRS

biotope

Floriane PLARD

OFB OFFICE FRANÇAIS DE LA BIODIVERSITÉ

Financeurs



Office Français de la Biodiversité
Direction générale
12, cours Louis Lumière
94300 VINCENNES
Contact : guillaume.paquignon@ofb.gouv.fr

Partenaires du projet



Observatoire Pelagis
UAR 3462 CNRS/LRUniv
5 allée de l'Océan
17000 La Rochelle
www.observatoire-pelagis.cnrs.fr
Contact : matthieu.authier@univ-lr.fr



Biotope - Département Diversification et Innovation
22 Boulevard Maréchal Foch
BP58 - 34140 MÈZE - FRANCE
www.biotope.fr
Contact : ylebras@biotope.fr

Floriane PLARD
Barraque de la Pinatelle
15500 Molompize
www.florianeplard.com
Contact : floriane.c.plard@gmail.com

Résumé

Le projet MODELCET ATL a pour but de créer un modèle prédictif de distribution et d'abondance des cétacés en Manche et en Atlantique. La première phase consiste à collecter, compiler et harmoniser les données existantes concernant les cétacés. Ce document présente les cinq étapes de la méthodologie employée à cette fin : l'inventaire des sources de données, leur hiérarchisation, leur collecte, leur harmonisation et l'assemblage final. L'inventaire a permis de recenser 108 sources de données protocolées (suivis nautiques, aériens visuels et "digitaux", données acoustiques exclues) et 13 sources de données de présence seulement. Pour obtenir une base de données conséquente pour la modélisation en respectant le temps imparti, les sources ont été hiérarchisées afin de sélectionner les plus pertinentes. La hiérarchisation se base sur des critères tels que la couverture de l'échantillonnage spatio-temporel ainsi que la qualité et la disponibilité des données. Les étapes de collecte, d'harmonisation et d'assemblage ont été appliquées sur un total de 68 jeux de données sélectionnés. Le document décrit les méthodes employées et les difficultés rencontrées. Le travail effectué sert de base pour le plan de gestion des données (UO 1.1.3), la base de données (UO 1.1.4) et la réalisation d'un modèle de distribution des cétacés (UO 1.2.3).

Mots clés

Cétacés, Données, Atlantique, Manche

Mise en garde

Ce rapport est confidentiel jusqu'à sa version finale

Citation

Veuillez citer ce travail comme suit :

> Yves Le Bras, Cindy Gidoin, Marion Pillet, Floriane Plard, Matthieu Authier. 30/09/2025. Description des méthodes de collecte, de compilation, et d'harmonisation des données. (UO 1.1.1). *MODELCET ATL : Conception d'un modèle prédictif de distribution et d'abondance des cétacés en Manche et en Atlantique*. Biotope.

Table des matières

1 Introduction **5**
- 1.1 Le projet MODELCET ATL . . . 5
- 1.2 Typologie des données . . . 5
- 1.3 Objectifs du document . . . 6

2 Inventaire des données **6**
- 2.1 Démarche générale . . . 6
- 2.2 Inventaire des données protocolées . . . 7
- 2.3 Inventaire des données de présence seulement . . . 14

3 Hiérarchisation et sélection des jeux de données **16**
- 3.1 Démarche générale . . . 16
- 3.2 Données protocolées . . . 17
  - 3.2.1 Données de suivis des parcs éoliens offshore . . . 18
  - 3.2.2 Données fournies par Pelagis . . . 19
  - 3.2.3 Données OFB . . . 20
- 3.3 Données de présence seulement . . . 20
  - 3.3.1 Global Biodiversity Information Facility . . . 21
  - 3.3.2 iNaturalist . . . 21
  - 3.3.3 OBSenMER . . . 21
  - 3.3.4 Sources non sélectionnées . . . 22

4 Collecte des données **22**
- 4.1 Procédure de demande . . . 22
- 4.2 Conditions d'utilisation . . . 24

5 Harmonisation des données **24**
- 5.1 Flux de travail pour l'ajout d'un nouveau jeu de données . . . 25
- 5.2 Extraction des données d'effort et des observations . . . 26
  - 5.2.1 Données d'effort . . . 27
  - 5.2.2 Observations biologiques . . . 28
  - 5.2.3 Jointure entre les observations et l'effort d'échantillonnage . . . 28
- 5.3 Données de configuration . . . 29
  - 5.3.1 Métadonnées . . . 29
  - 5.3.2 Spécification du format des données . . . 29
  - 5.3.3 Synonymes taxonomiques . . . 30

6 Assemblage des données **30**
- 6.1 Concaténation des données . . . 30
  - 6.1.1 Homogénéisation de la taille des groupes . . . 31
  - 6.1.2 Valeurs erronées ou manquantes . . . 31

7 Description de la base de données **32**

8 Discussion et perspectives **37**
- 8.1 Difficultés rencontrées . . . 37
  - 8.1.1 Accès aux données et autorisations d'utilisation . . . 37
  - 8.1.2 Contraintes de ressources limitées . . . 38
- 8.2 Limites de l'harmonisation . . . 38
  - 8.2.1 Hétérogénéité des protocoles . . . 38
  - 8.2.2 Hétérogénéité des échelles spatiales . . . 38

8.2.3 Qualité inégale des données 39
8.2.4 Perte d'information 39
8.3 Perspectives 39

9 Conclusion **39**

Références **40**

10 Annexes **41**
10.1 Nomenclature des jeux de données de suivis pour les parcs éoliens offshore 41
10.1.1 Abréviation des noms de parc éolien 41
10.1.2 Abréviation de la phases de projet 41
10.1.3 Abréviation de la méthode d'observation 41
10.2 Extraits de codes 42

1 Introduction

1.1 Le projet MODELCET ATL

Le projet MODELCET se décline en deux volets. Un volet "Manche et Atlantique" (que l'on nommera MODELCET ATL) et un volet Méditerranée (que l'on nommera MODELCET MED). Le premier a pour aire d'étude la façade Manche-Atlantique et le second la façade Méditerranéenne. Dans les deux cas, l'objectif est la conception d'un modèle prédictif de distribution et d'abondance des cétacés et le porteur de projet est l'Office Français de la Biodiversité. Il s'agit de valoriser les données existantes permettant d'évaluer la répartition spatio-temporelle des cétacés à l'échelle de la façade étudiée, dans l'objectif d'informer et d'orienter les politiques publiques de préservation de ces espèces et de leurs habitats naturels. Il s'agira notamment de produire des cartes adéquates pour répondre aux besoins de la planification spatiale à l'échelle des différentes sous-régions marines, dans le contexte du développement des énergies marines renouvelables. Le modèle prédictif devra donc être intégratif et capable de traiter simultanément plusieurs sources de données collectées selon des protocoles potentiellement différents et à des échelles spatio-temporelles variées.

Le projet MODELCET ATL a fait l'objet d'un appel d'offre en 2023 porté par l'Office Français de la Biodiversité. Ce marché public a été remporté par un consortium composé de Matthieu Authier, chercheur de l'Observatoire Pelagis (mandataire), Floriane Plard, chercheuse indépendante en statistique et écologie quantitative (co-traitante) et d'un bureau d'étude en ingénierie écologique, Biotope SAS (co-traitant). Le projet a débuté en octobre 2023 et s'achèvera en avril 2027.

1.2 Typologie des données

La distribution des cétacés est classiquement estimée à partir de suivis nautiques/aériens visuels et "digitaux". Ces données sont dites protocolées, c'est-à-dire qu'elles sont collectées en respectant une suite d'étapes et de méthodologies strictement définies visant à quantifier l'effort de prospection et à minimiser les biais de détection des animaux. Ces données sont généralement collectées par des professionnels (scientifiques et experts naturalistes). Elles sont relativement coûteuses à acquérir mais sont très riches en informations. Notamment, l'échantillonnage est connu, représentatif de la zone étudiée et stable dans le temps.

Il existe également les données dites "de présence seulement" (aussi appelées données "opportunistes" ou données "d'occurrence"). Ces données sont généralement collectées par des bénévoles et archivées sur des plateformes en ligne. Elles sont peu coûteuses, abondantes et de plus en plus nombreuses. En revanche, elles sont de moindre qualité que les données protocolées. La fiabilité de l'identification taxonomique est hétérogène. L'échantillonnage est "spontané", le plus souvent non documenté, et il est généralement hétérogène dans l'espace et le temps. Pour ce type d'échantillonnage, il est souvent difficile de savoir si une observation correspond à une rencontre fortuite de l'observateur et de l'espèce cible, ou si le premier était activement à la recherche du second.

Les données protocolées constituent la matière première la plus importante pour l'élaboration d'un modèle prédictif fiable tel que cela est attendu dans le cadre du projet MODELCET ATL. Les données de présence seulement pourraient être utilisées pour améliorer la précision des estimations et la confiance dans les prédictions du modèle. Toutefois, l'intégration de ces données reste délicate, voire controversée, car elle peut introduire des biais significatifs dans les estimations (voir UO 1.2.1 "Description des méthodes de couplage et de modélisation"). Dans ce contexte, les données de présence seulement pourraient également être valorisées lors des étapes de validation du modèle.

Dans ce contexte, la première phase du projet MODELCET ATL est de produire une base de données réunissant les jeux de données protocolées et de présence seulement dans la quantité la plus importante possible et dans un format exploitable pour le travail de modélisation. La production de la base de données nécessite la réalisation des tâches suivantes (Figure 1) :

1. **Inventaire** des sources de données existantes.
2. **Hiérarchisation** des sources de données afin de sélectionner les plus importantes et pertinentes pour le projet
3. **Collecte** des jeux de données sélectionnés et des autorisations d'utilisation.
4. **Harmonisation** des différents jeux de données sélectionnés.
5. **Assemblage** et mise à disposition de la base de données.

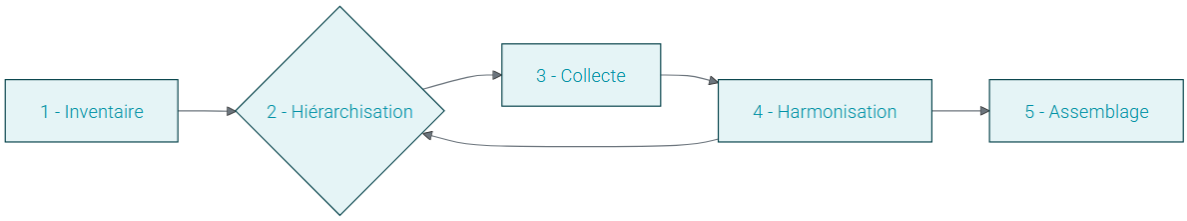


FIGURE 1 – Logigramme de la construction de la base de données

1.3 Objectifs du document

Le présent document s'inscrit dans la première phase du projet MODELCET ATL consistant à collecter et à harmoniser les données disponibles concernant les cétacés et pertinentes pour la conception du modèle prédictif. Au sein du consortium, cette mission est assurée principalement par Biotope SAS, en partenariat avec La Rochelle Université (Matthieu Authier et Marion Pillet) et Floriane Plard. Ces derniers participant à la collecte des données et à la validation scientifique de la base de données finale.

L'objectif du livrable "UO 1.1.1 - Description des méthodes de collecte, compilation et d'harmonisation des données" est d'apporter toutes les informations permettant de comprendre les étapes de travail et les choix qui ont conduit à la base de données finale.

2 Inventaire des données

2.1 Démarche générale

Dans un premier temps, l'objectif est de constituer une liste la plus exhaustive possible des sources de données existantes avec, pour chaque source :

- L'institution dépositaire des données, avec une personne contact pour la collecte et les éventuelles questions techniques sur les données.
- L'institution propriétaire des données, avec une personne contact pour établir les autorisations nécessaires à l'utilisation des données.
- Les informations concernant le protocole et le matériel utilisé
- Les informations concernant l'effort d'échantillonnage, telles que les années et zones géographiques échantillonnées.

Les informations concernant le protocole et l'effort d'échantillonnage sont des prérequis au travail de hiérarchisation des sources de données qui doit aboutir à la sélection des jeux de données les plus pertinents pour le projet.

Un inventaire des sources de données à considérer, et des organismes à consulter a été établi sur la base des connaissances du consortium. Cette liste a été soumise à l'OFB pour avis et des compléments ont pu être apportés. La Table 1 décrit la liste des personnes consultées pour la réalisation de l'inventaire des sources de données. En outre, les sources exploitées par certains travaux de référence (Savouré-Soubelet et al. 2016) ont été inspectées également pour identifier des sources supplémentaires.

TABLE 1 – Personnes consultées pour la réalisation de l'inventaire des sources de données

Organisme	Personnes
Biotope	Marjolaine Caillat, Pauline De Rock (Service Énergies Marines Renouvelables) Yves Le Bras, Cindy Gidoin (Equipe biostatistique)
Pelagis	Matthieu Authier
Floriane Plard	Floriane Plard
OFB	Guillaume Paquignon, Sophie Poncet, Morgane Remaud, Guilhem Autret

Cet inventaire des sources de données a été complété et corrigé tout au long du travail d'harmonisation, au fur et à mesure de l'arrivée de nouvelles informations. Certaines sources se sont avérées être des doublons et ont été supprimées de l'inventaire.

La caractérisation du protocole est capitale pour une collecte et une harmonisation correctes et efficaces des jeux de données et de leurs métadonnées. Quatre critères ont été employés pour caractériser les protocoles :

- **Type des données** : protocolées (avec effort) ou présence seulement (sans effort).
- **Plateforme d'observation** : nautique ou aérienne.
- **Méthode d'observation** : visuelle ou "digitale".
- **Protocole scientifique** : estimation de l'abondance via "distance sampling" ou via "strip-transect".

Ces critères peuvent être mutuellement exclusifs, la Figure 2 représente de façon visuelle ces catégories et leur imbrication.

> **i** Note 1 : Cas des suivis acoustiques passifs
>
> Toutes les catégories présentées sur la Figure 2 ont été considérées pour l'intégration à la base de données MODELCET ATL. Néanmoins cela exclu les données de **suivis acoustiques passifs**. En effet, ces données sont considérablement plus volumineuses et soulèvent des problématiques supplémentaires dans le travail de modélisation à suivre [1], si bien que le consortium, en accord avec l'OFB, a décidé de ne pas élargir le cadre de ce projet à ce type de suivi.

L'une des principales difficultés dans la création de la base de données MODELCET ATL concerne la gestion des métadonnées, le suivi continue des informations à rechercher/ajouter/rectifier, pour une centaine de sources en parallèle et par différentes personnes/organisations simultanément. Pour faire face à ces défis et dans l'optique d'offrir transparence et reproductibilité au projet, nous nous sommes appuyés sur l'utilisation d'outils de versionnement, git pour le code et DVC pour les données et chaînes de traitement. L'ensemble du code et des métadonnées produits dans le cadre de cette mission est hébergé sur la plateforme gitlab de l'OFB. Ces outils permettent également d'assurer, auprès de l'OFB et du reste du consortium, une livraison continue, détaillée et sécurisée de tous les fichiers de code, de configuration, de métadonnées et de toutes les modifications apportées aux données au cours du travail d'harmonisation.

2.2 Inventaire des données protocolées

Comme présenté sur la Figure 2, nous nous sommes intéressés aux données protocolées issues de :

1. Sans rentrer dans les détails techniques, notons simplement que 1°) les données acoustiques ne quantifient pas un nombre d'individus mais un nombre de détections acoustiques, 2°) ce processus de détection évolue de façon importante et complexe dans le temps et l'espace en fonction de facteurs comme le trafic maritime ou les conditions météorologiques.

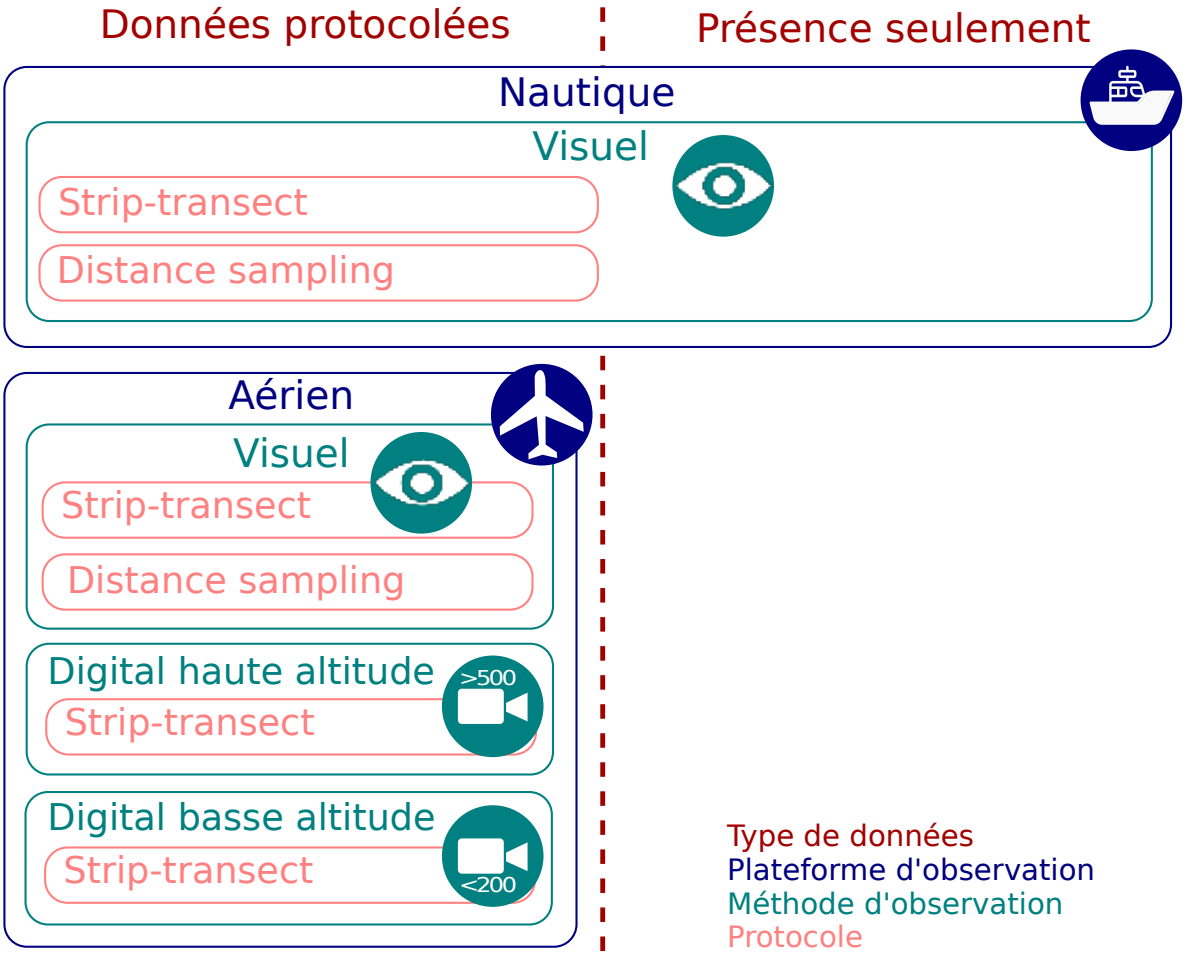


FIGURE 2 – Catégorisation des jeux de données en fonction du type de données (en rouge), de la plateforme d'observation (en bleu), de la méthode d'observation (vert) et du protocole d'estimation de l'abondance. Les plateformes d'observation, les méthodes d'observation et les protocoles pouvant être employés dans le cadre de données protocolées sont représentés à gauche de la ligne verticale pointillées. Dans le cadre des données de type présence seulement, ces éléments sont représentés à droite de la ligne verticale pointillées. Par exemple, les données de type présence seulement sont uniquement collectées lors de suivis nautiques avec une méthode visuel et sans protocole.

— Suivis nautiques
— Suivis aériens avec des observateurs embarqués (ci-après suivis aériens visuels)
— Suivis aériens digitaux (ci-après suivis digitaux)

Ces différentes plateformes d'observation (bateau ou avion) et les différentes méthodes d'observation (visuelle ou digitale) ont chacune des particularités et chaque jeu de données peut lui-même se distinguer au sein d'une même plateforme et d'une même méthode d'observation car ces dernières ne sont pas standardisées de façon immuable et intemporelle.

Nous avons inventorié 108 jeux de données protocolées contenant des observations de mammifères marins sur la façade Manche et Atlantique (Table 2). Il est important de noter qu'il s'agit de la liste des sources *potentielles* au moment de l'inventaire, sources dont l'existence restait alors à confirmer et les métadonnées à définir.

TABLE 2 – Liste des jeux de données protocolées inventoriées. Pour des raisons de lisibilités les noms abrégés des jeux de données sont utilisés. Consulter la Section 10.1 pour plus d'information sur ces noms. Les champs laissés vides indiquent des métadonnées que nous n'avons pas pu obtenir.

Façades	Nom	Plateforme	Méthode	Années	Propriétaire	Fournisseur	Intégré
Toutes	SAMM-I	Aérienne	Visuel	2011-2012 (2)	Publique	Pelagis	Oui
Toutes	MiniSAMM	Aérienne	Visuel	2014 (1)	Publique	Pelagis	Oui
Toutes	SAMM-II	Aérienne	Visuel	2021 (1)	Publique	Pelagis	Oui
Toutes	SCANS-IV_plane	Aérienne	Visuel	2022 (1)	Publique	Pelagis	Oui
Toutes	Orca	Nautique	Visuel		Orca	Orca	Oui
Toutes	OWFSOMM_boat	Nautique	Visuel	2021-2023 (3)	Publique	FEM	Non
Manche	JONATHAN-Sept-Îles	Aérienne	Digital	2014 (1)	Publique	OFB	Non
Manche	JONATHAN	Aérienne	Digital	2016 (1)	Publique	OFB	Oui
Manche	STB_EDR_digital	Aérienne	Digital	2020-2021 (2)	AILES-MARINES	Biotope	Oui
Manche	STB_Con01	Aérienne	Digital	2021 (1)	AILES-MARINES	Biotope	Oui
Manche	FEC_EDR_digital	Aérienne	Digital	2021 (1)		FEM	Non
Manche	COUR_EDR_digital2	Aérienne	Digital	2021 (1)		FEM	Non
Manche	COUR_EDR_digital	Aérienne	Digital	2021 (1)		FEM	Non
Manche	CEM_EI_digital	Aérienne	Digital	2022 (1)	EDF	Biotope	Oui
Manche	CEM_EI_digital_hytech	Aérienne	Digital	2022 (1)	EDF,PUBLIC	OFB	Non
Manche	STB_Con02	Aérienne	Digital	2022-2023 (2)	AILES-MARINES	Biotope	Oui
Manche	FEC_Con_digital	Aérienne	Digital	2022-2023 (2)	EDF	Biotope	Oui
Manche	TRE_EDR_digital	Aérienne	Digital	2022-2023 (2)	ENGIE	Biotope	Oui
Manche	TRE_EDR_digital2	Aérienne	Digital	2022-2023 (2)	ENGIE	FEM	Non
Manche	STB_Con03	Aérienne	Digital	2023 (1)	AILES-MARINES	Biotope	Oui
Manche	MMN	Aérienne	Digital	2024 (1)	Publique	Biotope	Oui
Manche	TRE_EI_plane1	Aérienne	Visuel	2007-2008 (2)	ENGIE	Biotope	Non
Manche	FEC_EI_plane1	Aérienne	Visuel	2008-2009 (2)	EDF	Biotope	Non
Manche	Grand-Large	Aérienne	Visuel	2009-2010 (2)	COMPANY	Biotope	Oui
Manche	TRE_EI_plane2	Aérienne	Visuel	2010-2011 (2)	ENGIE	Biotope	Oui
Manche	B2B_EI_plane	Aérienne	Visuel	2010-2011 (2)	EDF	Biotope	Oui
Manche	FEC_EI_plane2	Aérienne	Visuel	2012-2013 (2)	EDF	Biotope	Oui
Manche	STB_EI_plane	Aérienne	Visuel	2013-2014 (2)	AILES-MARINES	INVIVO	Non
Manche	COUR_EI_plane	Aérienne	Visuel	2014 (1)	EDF		Non
Manche	TRE_EI_plane3	Aérienne	Visuel	2014-2015 (2)	ENGIE	Biotope	Oui
Manche	DunkRisk_plane	Aérienne	Visuel	2017-2018 (2)	Publique	Pelagis	Oui
Manche	FEC_EDR_plane	Aérienne	Visuel	2019-2021 (3)	EDF	Biotope	Oui
Manche	COUR_EDR_plane	Aérienne	Visuel	2020-2021 (2)	EDF		Non
Manche	DUN_EI_plane	Aérienne	Visuel	2020-2021 (2)	EDF	Biotope	Oui
Manche	ZPS-BretagneNord	Aérienne	Visuel	2020-2021 (2)	Publique	OFB,SETEC	Oui
Manche	CEM_EI_plane	Aérienne	Visuel	2021-2023 (3)	Publique	Biotope	Oui

TABLE 2 – Liste des jeux de données protocolées inventoriées. Pour des raisons de lisibilités les noms abrégés des jeux de données sont utilisés. Consulter la Section 10.1 pour plus d'information sur ces noms. Les champs laissés vides indiquent des métadonnées que nous n'avons pas pu obtenir.

Façades	Nom	Plateforme	Méthode	Années	Propriétaire	Fournisseur	Intégré
Manche	MAMO (Estuaires Picards)	Aérienne	Visuel	2023 (1)		Pelagis	Oui
Manche	Bernard-Bril (sorties en ferry Dunkerque-Douvres)	Nautique	Visuel		Bernard-Bril (sorties en ferry Dunkerque-Douvres)	Bernard-Bril (sorties en ferry Dunkerque-Douvres)	Non
Manche	IBTS	Nautique	Visuel	2007-2023 (17)	IFREMER	Pelagis	Oui
Manche	COUR_EI_boat	Nautique	Visuel	2008-2009 (2)	EDF	COUR_EI_boat	Non
Manche	FEC_EI_boat	Nautique	Visuel	2008-2009,2012-2013 (4)	EDF	ASSO,BIOTOPE	Non
Manche	TRE_EI_boat1	Nautique	Visuel	2010-2011 (2)	ENGIE	Biotope	Non
Manche	Hauts-de-France-ZNIEFF	Nautique	Visuel	2011-2012 (2)	Publique	Biotope	Non
Manche	STB_EI_boat	Nautique	Visuel	2012-2014 (3)	AILES-MARINES	INVIVO	Non
Manche	CAMANOC	Nautique	Visuel	2014 (1)	IFREMER	Pelagis	Oui
Manche	TRE_EI_boat2	Nautique	Visuel	2014-2015 (2)	ENGIE	Biotope	Non
Manche	CGFS	Nautique	Visuel	2016-2023 (8)	IFREMER	Pelagis	Oui
Manche	LEDKON	Nautique	Visuel	2017-2018 (2)	Publique	Pelagis	Non
Manche	FEC_EDR_boat	Nautique	Visuel	2019-2020 (2)	EDF	Biotope	Oui
Manche	DUN_EI_boat	Nautique	Visuel	2020 (1)	EDF	Biotope	Oui
Manche	COUR_EDR_boat	Nautique	Visuel	2020-2021 (2)	EDF		Non
Manche	STB_EDR_boat	Nautique	Visuel	2020-2021 (2)	AILES-MARINES	Biotope	Non
Manche	CEM_EI_boat	Nautique	Visuel	2021-2023 (3)	EDF	Biotope	Oui
Manche	FEC_Con_boat	Nautique	Visuel	2022-2023 (2)	EDF	Biotope	Oui
Manche	TRE_EDR_boat	Nautique	Visuel	2022-2023 (2)	ENGIE	Biotope	Oui
Atl.	Yeu-HiDef	Aérienne	Digital	2016-2017 (2)		Biotope	Non
Atl.	SNA_Con_digital	Aérienne	Digital	2021-2022 (2)	EDF	Biotope	Oui
Atl.	AO5_digital2 (Bretagne Sud)	Aérienne	Digital	2022-2023 (2)	Publique	FEM	Non
Atl.	AO5_digital (Bretagne Sud)	Aérienne	Digital	2022-2023 (2)	Publique	FEM	Non
Atl.	NOY_EDR	Aérienne	Digital	2022-2023 (2)	ENGIE	Biotope	Oui
Atl.	BOY_digital	Aérienne	Digital	2023-2024 (2)		Biotope	Oui
Atl.	MARPRO_plane	Aérienne	Visuel	2011-2015 (5)	Publique	RESEARCH	Non
Atl.	NOY_EI_plane	Aérienne	Visuel	2014-2016 (3)	ENGIE	BIOTOPE,BV	Oui
Atl.	SPEE-I	Aérienne	Visuel	2019 (1)	Publique	Pelagis	Oui
Atl.	SPEE-II	Aérienne	Visuel	2020 (1)	Publique	Pelagis	Oui
Atl.	CAPECET-I	Aérienne	Visuel	2020 (1)	Publique	Pelagis	Oui
Atl.	SPEE-III	Aérienne	Visuel	2021-2022 (2)	Publique	Pelagis	Oui
Atl.	ZPS-BretagneSud	Aérienne	Visuel	2021-2022 (2)	Publique	OFB	Oui
Atl.	PNMI	Aérienne	Visuel	2021-2022 (2)	Publique	OFB,SETEC	Oui
Atl.	AO5_plane (Bretagne Sud)	Aérienne	Visuel	2022-2023 (2)		SETEC	Non

TABLE 2 – Liste des jeux de données protocolées inventoriées. Pour des raisons de lisibilités les noms abrégés des jeux de données sont utilisés. Consulter la Section 10.1 pour plus d'information sur ces noms. Les champs laissés vides indiquent des métadonnées que nous n'avons pas pu obtenir.

Façades	Nom	Plateforme	Méthode	Années	Propriétaire	Fournisseur	Intégré
Atl.	CAPECET-II	Aérienne	Visuel	2023 (1)	Publique	Pelagis	Oui
Atl.	AO7_EI	Aérienne	Visuel	2023-2024 (2)			Non
Atl.	PIL (Projet pilote flottant de La Rochelle & Belle-île)	Nautique	Visuel			SETEC	Non
Atl.	ERMMA (Centre de la Mer de Biarritz)	Nautique	Visuel		Publique	OFB	Non
Atl.	PELACUS	Nautique	Visuel	2007-2023 (17)	Publique	RESEARCH	Oui
Atl.	BOBGEO	Nautique	Visuel	2009 (1)		Pelagis	Oui
Atl.	EVHOE	Nautique	Visuel	2009-2014,2016-2023 (14)	IFREMER	Pelagis	Non
Atl.	MARPRO_boat	Nautique	Visuel	2011-2015 (5)	Publique	RESEARCH	Non
Atl.	JUVENA	Nautique	Visuel	2012-2021 (10)	Publique	RESEARCH	Oui
Atl.	SNA_Diag	Nautique	Visuel	2013 (1)	EDF	SNA_Diag	Non
Atl.	SNA_EI	Nautique	Visuel	2013-2015 (3)	EDF	Biotope	Non
Atl.	NOY_EI_boat	Nautique	Visuel	2014-2016 (3)	ENGIE	Biotope	Non
Atl.	ZPS-BOY-Periscope	Nautique	Visuel	2015-2017 (3)	ZPS-BOY-Periscope	OFB	Oui
Atl.	TRIENAL	Nautique	Visuel	2016,2022 (2)	Publique	RESEARCH	Oui
Atl.	BIOMAN	Nautique	Visuel	2016-2023 (8)	Publique	RESEARCH	Oui
Atl.	SNA_EDR	Nautique	Visuel	2019-2020 (2)	EDF	Biotope	Oui
Atl.	IBERAS	Nautique	Visuel	2019-2023 (5)	Publique	RESEARCH	Oui
Atl.	PELGAS	Nautique	Visuel	2019-2024 (6)	IFREMER	Pelagis	Non
Atl.	SNA_Con_boat	Nautique	Visuel	2021-2022 (2)	EDF	Biotope	Non
Atl.	SEPANSO	Nautique	Visuel	2023 (1)	Publique	ASSO,OFB	Oui
Atl.	PELAGO	Nautique	Visuel	2023 (1)	Publique	RESEARCH	Oui
	Hytech	Aérienne	Digital		Publique	OFB	Non
	Picardie-Nature	Aérienne	Visuel		Picardie-Nature	Picardie-Nature	Non
	CMNF	Aérienne	Visuel				Non
	SCANS-II_plane	Aérienne	Visuel	2005 (1)	Publique	Pelagis	Oui
	ObSERVE	Aérienne	Visuel	2015-2017 (3)	Publique	Pelagis	Oui
	SCANS-III_plane	Aérienne	Visuel	2016 (1)	Publique	Pelagis	Oui
	ObSERVE2	Aérienne	Visuel	2021-2023 (3)		Pelagis	Oui
	Bretagne-Vivante	Nautique	Visuel		Bretagne-Vivante	Bretagne-Vivante	Non
	GONm	Nautique	Visuel		GONm	GONm	Non
	SCANS2_boat	Nautique	Visuel	2005 (1)		Pelagis	Oui
	CODA	Nautique	Visuel	2007 (1)	Publique	Pelagis	Oui
	BOBECO (projet européen CoralFISH)	Nautique	Visuel	2011 (1)	Publique	Pelagis	Oui
	SCANS3_boat	Nautique	Visuel	2016 (1)	Publique	Pelagis	Oui

TABLE 2 – Liste des jeux de données protocolées inventoriées. Pour des raisons de lisibilités les noms abrégés des jeux de données sont utilisés. Consulter la Section 10.1 pour plus d'information sur ces noms. Les champs laissés vides indiquent des métadonnées que nous n'avons pas pu obtenir.

Façades	Nom	Plateforme	Méthode	Années	Propriétaire	Fournisseur	Intégré
	SCANS-IV_boat	Nautique	Visuel	2022 (1)	Publique	Pelagis	Non

2.3 Inventaire des données de présence seulement

Pour ces données, nous n'avons pas cherché à faire un inventaire exhaustif ou quasi-exhaustif des sites ou des propriétaires. Pour ce type de données, l'exhaustivité n'est ni un objectif atteignable (les sources sont trop nombreuses), ni un objectif pertinent car un petit nombre de plateformes en ligne réunissent de vastes ensembles de données d'occurrences. Ces bases de données, telles que le GBIF [2] ou iNaturalist, hébergent les données d'un grand nombre de contributeurs particuliers et d'associations (exemple : AI-Lark, APECS) et participent activement à la collecte des données de contributeurs importants tels que les entreprises (exemple : Faune Océan), les acteurs publics (exemple : OFB) et les institutions de recherche (exemple : CNRS, MNHN).

Nous avons inventorié 13 plateformes en ligne, entreprises ou associations possédant ce type de données en Manche/Atlantique (Table 3).

2. Global Biodiversity Information Facility

TABLE 3 – Listes des plateformes en ligne collectant des données de présence seulement.

Façades	Nom	Plateforme	Méthode	Années	Propriétaire	Fournisseur	Intégré
Toutes	ObsEnMer	Nautique	Visuel		ASSO,PUBLIC	ObsEnMer	Oui
Toutes	GBIF	Nautique	Visuel		ASSO,PUBLIC	GBIF	Oui
Toutes	iNaturalist	Nautique	Visuel		ASSO,PUBLIC	iNaturalist	Oui
Manche	OCEAMM (ancienne association Hauts de France)	Nautique	Visuel		OCEAMM (ancienne association Hauts de France)	OCEAMM (ancienne association Hauts de France)	Non
	observation.org	Nautique	Visuel		ASSO,PUBLIC	observation.org	Non
Toutes	OpenObs (INPN)	Nautique	Visuel		Publique	MNHN	Non
Manche	GMN (Groupe Mammalogique Normand)	Nautique	Visuel		GMN (Groupe Mammalogique Normand)	GMN (Groupe Mammalogique Normand)	Non
	LPO	Nautique	Visuel		LPO	LPO	Non
	Al-Lark	Nautique	Visuel		Al-Lark	Al-Lark	Non
	SiRF	Nautique	Visuel		SiRF	SiRF	Non
	Faune-Océan	Nautique	Visuel		COMPANY	COMPANY	Non
	Orca-PO	Nautique	Visuel		Orca-PO	Orca-PO	Non
	APECS (association requins)	Nautique	Visuel		APECS (association requins)	APECS (association requins)	Non

3 Hiérarchisation et sélection des jeux de données

3.1 Démarche générale

L'inventaire des données a permis de recenser de nombreuses sources de données. Face à cette multitude, le temps allouable à cette mission étant limité, il a fallu établir une hiérarchie dans la pertinence des différentes sources de données à contribuer qualitativement et quantitativement au projet MODELCET ATL.

Cette hiérarchisation a été établie en fonction des contraintes et des objectifs de modélisation du projet MODELCET ATL. C'est pourquoi les critères de hiérarchisation et de sélection à prendre en compte ont été établis en concertation avec l'OFB, Pelagis et Floriane Plard. Il convient de noter que le fait qu'une source de données ait pu être intégrée à la base de données finale ne garantit pas son utilisation lors de la modélisation. En effet, le modèle peut encore n'utiliser qu'une partie des données disponibles dans la base de données finale, et ce pour des raisons pratiques, techniques ou scientifiques dont l'émergence fait partie intégrante du processus de modélisation.

Ces critères (voir Table 4) peuvent être classés en trois familles : **échantillonnage**, **qualité** et **disponibilité**. Pour mieux comprendre ces critères, il est intéressant de décliner l'objectif principal du projet en **objectifs opérationnels** plus directement liés aux contraintes de la modélisation (Note 2).

> **i** Note 2 : Objectifs opérationnels
>
> 1. Permettre des prédictions de densité (c'est à dire un nombre d'animaux par unité de surface) au sein de l'ensemble de la zone d'étude ;
> 2. Obtenir des prédictions représentatives pour les prochaines années ;
> 3. Appréhender la saisonnalité inter-annuelle ;
> 4. Appréhender la saisonnalité intra-annuelle ;
> 5. Éviter les biais lors de l'estimation par modélisation (voir UO 1.2.1 "Description des méthodes de couplage et de modélisation") ;
> 6. Obtenir une base de données finale la plus volumineuse possible dans le temps imparti ;
> 7. Assurer une procédure d'intégration des jeux de données standardisée et documentée ;
> 8. Assurer de l'intégration des suivis qui ont été financés par l'état afin de valoriser au maximum les données publiques.

Les métadonnées et les données sont collectées et traitées de manière progressive. Leur disponibilité, ainsi que les incertitudes liées à leur collecte et à leur utilisation, varient considérablement selon les sources. Dans ce contexte, l'évaluation des critères a été conduite de façon qualitative.

La hiérarchisation constitue évidemment la base pour la sélection des sources de données. Il existe quelques exceptions de jeux de données que l'on jugerai au final moins prioritaires (e.g. certaines suivis nautiques, Figure 3) pourtant sélectionnés. Cela s'explique par leur disponibilité qui permettait de faire progresser le projet aux moments où l'équipe disposait de conditions favorables pour traiter des données (contraintes de travail exogènes).

Enfin, certains critères dépendent des jeux de données déjà intégrés ou de ceux qui ne pourront pas l'être. Les priorités évoluent donc au fur et à mesure de l'enrichissement de la base de données. La hiérarchisation des données s'effectue donc selon une approche itérative, dans laquelle l'importance relative des sources est régulièrement réévaluée.

TABLE 4 – Les quatre grandes familles de critères (échantillonnage spatial, échantillonnage temporel, qualité des données et disponibilité des données) ainsi que les critères de hiérarchisation des sources associés à chacune d'elles. Les objectifs opérationnels sont précisés entre parenthèses dans la deuxième colonne du tableau, les numéros indiqués faisant référence à ceux de la Note 2.

Famille	Critère (objectifs opérationnels)	Description
Échantillonnage spatial	Surface totale échantillonnée (1, 6)	Compte tenu de l'échelle spatiale de notre zone d'étude (façades maritimes entières), les sources couvrant de vastes secteurs sont favorisées.
	Recouvrement avec la zone d'étude (1, 5)	On recherche un maximum d'adéquation entre le secteur échantillonné et la zone d'étude. Il peut s'agir du recouvrement géographique, mais dans une moindre mesure d'une similarité dans les conditions environnementales.
	Complémentarité de l'échantillonnage (2, 5)	On cherche à disposer d'un effort de prospection relativement équilibré dans la zone d'étude (e.g. campagne ObSERVE), et non sur-concentré dans le secteur côtier et les parcs éoliens offshore
Échantillonnage temporel	Age des données	Les données collectées récemment sont plus à même de fournir un modèle dont les prédictions seront valables les prochaines années.
	Nombre d'années échantillonnées (3)	Plus la série temporelle est longue et mieux elle permettra d'appréhender la tendance de fond et la variabilité inter-annuelle.
	Passages répétés au cours de l'année (4)	Plus la fréquence de suivi est importante et mieux on pourra caractériser les effets saisonniers.
Qualité des données	Fiabilité de la probabilité de détection (5)	Le protocole distance-sampling, qui permet d'estimer la probabilité de détection, a plus de valeur que le strip-transect qui suppose la détection parfaite. En outre, cette hypothèse du strip-transect apparaît plus crédible dans le cadre des suivis digitaux que des suivis visuels.
	Conditions d'observation (5)	Les conditions d'observation associées aux segments d'effort sont disponibles. Cela permettra d'améliorer la prise en compte de cette source d'hétérogénéité.
	Qualité d' identification taxonomique (5)	Le taux d'identification partiel est faible (<10 %) et les étiquettes taxonomiques des espèces sont fiables.
	Représentativité (2, 5)	On évitera d'utiliser les données qui pourraient être soumises à des biais particuliers. Par exemple l'utilisation d'un sonar durant la prospection ou la réalisation de travaux de construction à proximité.
Disponibilité des données	Conditions d'utilisation (8)	Données publiques ou ayant une licence d'utilisation permissive (e.g licence ETALAB).
	Facilité d'accès aux données (6)	Un accès facile aux données via un site internet, une base de données ou un contact proche, facilitent grandement l'utilisation d'une source.
	Préparabilité des données (6)	Les données non bancarisées de façon numérique (format papier), non consolidées par leur producteur, incomplètes ou mal organisées compliquent considérablement l'intégration. On privilégie les données avec un bon compromis entre la difficulté ou la quantité de travail à accomplir et l'intérêt du jeu de données pour les enjeux de modélisation précités.

3.2 Données protocolées

Pour les données protocolées, nous avons recensé plus d'une centaine de sources. Toutefois, les informations disponibles variaient fortement d'une source à l'autre et se révélaient parfois incomplètes. Dans ce contexte, la hiérarchisation a été réalisée de manière qualitative, en nous appuyant à la fois sur notre connaissance des suivis et sur notre expérience en matière de modélisation. Dans la suite

du document, nous présentons de façon macroscopique les caractéristiques des grandes familles de critères pour trois types de données : les données de suivis des parcs éoliens offshore, les données fournies par Pelagis et les données OFB.

3.2.1 Données de suivis des parcs éoliens offshore

Les suivis des parcs éoliens présentent de nombreuses caractéristiques communes (Table 5).

TABLE 5 – Caractéristiques des suivis des parcs éoliens offshore

Famille de critères	Caractéristiques
Échantillonnage spatial	L'emplacement de leur zone d'étude étant dicté par la position des parcs éoliens, les secteurs couverts sont côtiers. Ils cumulent un suivi visuel nautique sur une zone d'étude dite *rapprochée* (resserrée autour le l'emprise spatiale du parc) avec un suivi aérien (visuel ou digital) au sein d'une zone d'étude dite *étendue* (qui s'éloigne de plusieurs dizaines de kilomètres de la périphérie du parc). — Il existe une forte redondance entre ces suivis. — Par rapport aux autres sources de données on peut qualifier l'échantillonnage des suivi aériens de modérément étendus et ceux des suivis nautique de faiblement étendus. La Figure 3 illustre cette différence d'échelle spatiale typiquement observable entre l'échantillonnage des suivis aériens et nautiques.
Échantillonnage temporel	Les années des suivis sont fonction du calendrier des projets éoliens, les suivis sont donc plutôt récents. Largement répétés dans le temps, plusieurs années et pour chaque année suivis plusieurs fois dans l'année. Les années de suivi sont les mêmes pour les suivis aériens et nautiques. Il existe une forte redondance ici encore.
Qualité des données	Suivis toujours effectués par des naturalistes professionnels et en respectant les protocoles standards.
Disponibilité des données	Ils sont facilement mobilisables et relativement homogènes en terme de structure étant donné que le même organisme, Biotope, était dans de nombreux cas en charge de la réalisation de ces suivis.

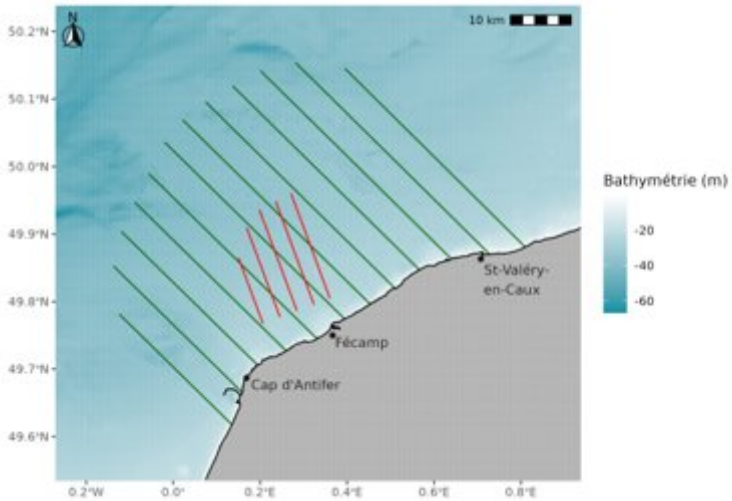


FIGURE 3 – Exemple d'échantillonnage spatial d'un suivi aérien (transects verts) et d'un suivi nautique (transects rouges) dans le cadre d'un projet éolien (Fécamp).

En considérant l'ensemble de ces propriétés, les suivis aériens des parcs éoliens offshore sont ciblés comme prioritaires. Compte tenu de l'intégration des suivis aériens et de la forte similarité de l'échan-

tillonnage spatio-temporel avec les suivis nautiques, ces derniers – dont l'échantillonnage spatial est peu étendu comparativement – sont jugés non prioritaires.

Au sein des données aériennes visuelles, les suivis pour lesquels un protocole distance-sampling avaient été mis en place sont prioritaires par rapport aux suivis utilisant le protocole strip-transect. Les suivis aériens digitaux sont basés sur le protocole strip-transect mais on considère qu'il disposent d'une probabilité de détection élevée et relativement similaire d'un suivi à l'autre. Surtout, ils sont pour la plupart très accessibles et très facilement "préparables" car déjà largement harmonisées. Les suivis aériens digitaux sont donc jugés aussi prioritaires que les suivis aériens visuels distance-sampling.

Bien que non prioritaires par rapport aux données aériennes, quelques données de suivis nautiques ont été intégrées à la base de données finale. Au sein de cette catégorie, les données récentes sont prioritaires car elles ont l'avantage d'être de meilleure qualité (avec des traces GPS) et plus standardisées (utilisation d'un outil tablette à partir de 2022).

3.2.2 Données fournies par Pelagis

Les données fournies par Pelagis sont issues de multiples sources. Elles ont été collectées dans des contextes variés (campagnes halieutiques, programmes de monitoring internationaux, nationaux ou régionaux). Néanmoins, ces sources de données présentent de nombreuses caractéristiques communes (Table 6).

TABLE 6 – Caractéristiques des suivis fournis par l'observatoire Pelagis

Famille de critères -	Caractéristiques
Échantillonnage spatial	Qu'il s'agisse de suivis nautiques ou aériens, les jeux de données couvrent des zones d'études étendues, incluant dans de nombreux cas des secteurs hauturiers.
Échantillonnage temporel	Certains suivis sont répétés depuis de nombreuses années. C'est le cas pour différents suivis menés dans le cadre de campagnes scientifiques halieutiques (`IBTS`, `EVHOE`, `CGFS` et `PELGAS`). Le jeu de données `PELGAS` a en plus l'intérêt d'être particulièrement récent puisque les données vont de 2019 à 2024.
Qualité des données	Les suivis sont toujours effectués par des naturalistes professionnels et selon des protocoles conformes à la méthodologie du distance sampling. Les protocoles sont particulièrement homogènes d'un suivi à l'autre, avec le protocole MEGASCOPE (Dorémus et Van Canneyt 2025) pour les suivis nautiques et l'utilisation de SAMMOA (Observatoire Pelagis 2025) pour les suivis aériens.
Disponibilité des données	Ils sont facilement mobilisables étant donné que Pelagis était, soit en charge de la réalisation de ces suivis, soit associé à la valorisation scientifique des données. Par ailleurs, ces jeux de données ont fait l'objet d'un travail de nettoyage et d'un archivage dans une base de données (voir Observatoire Pelagis (2021)), ils sont donc particulièrement homogènes en terme de structure.

Parmi les 28 jeux de données que Pelagis pouvait potentiellement fournir (voir Table 2, fournisseur Pelagis), nous avons estimé que 16 jeux de données étaient prioritaires pour les raisons détaillées dans la Table 6. Les 16 jeux de données prioritaires ont tous pu être intégrés.

Le jeu de données `CAMANOC` a été jugé moins prioritaire car il n'y a eu qu'une année de suivi (2014). Le jeu de données `IBTS` également, l'utilisation d'un sonar durant l'observation pouvant constituer une source de biais pour certaines espèces. Cependant, ces deux sources ont tout de même été traitées car elles ont bénéficié du fait qu'elles étaient contenues dans la même base de données (rassemblant `CAMANOC`, `IBTS` et `CGFS`). Ajouter ce jeu de données ne demandait pas d'effort important ou de travail supplémentaire et répondait donc à l'enjeu 6 (Note 2).

Nous avons considéré que les données protocolées de l'AZTI (Centro de Investigación Marina y Alimentaria) et l'IEO (Instituto Español de Oceanografía) étaient moins prioritaires que les données provenant des instituts français (voir Table 2, fournisseur research). Au sein de cette catégorie (AZTI et IEO), tous les jeux de données avaient le même niveau de priorité. Par exemple, le suivi le plus éloigné de la zone d'étude était aussi le plus récent (`PELAGO` de l'IEO 2023). De plus, différents jeux pouvant être livrés en même temps, avec le même format et donc nécessitant les mêmes traitements de données, il aurait été dommage de ne pas les intégrer sous prétexte que le niveau de priorité n'était pas maximal (enjeu 6, Note 2).

3.2.3 Données OFB

Les jeux de données pouvant potentiellement être fournis directement par l'OFB sont au nombre de 10.

TABLE 7 – Caractéristiques des suivis fournis par l'Office Français de la Biodiversité

Famille de critères	Caractéristiques
Échantillonnage spatial	Les aires d'étude ayant des étendues comparables à celles des suivis de parcs éoliens offshore.
Échantillonnage temporel	Jeux de données plutôt récents (2014-2022). Pas de suivis sur de longues périodes (au maximum 3 ans).
Qualité des données	Suivis toujours effectués par des naturalistes professionnels et en respectant les protocoles standards.
Disponibilité des données	Moyenne car différents producteurs de données ce qui implique différents interlocuteurs et des données structurée de façon différentes.

Outre les caractéristiques listées dans la Table 7, on notera qu'il s'agit de suivis financés par l'État français. À ce titre, ils bénéficient de conditions d'utilisation permissives (e.g. licence ETALAB) et permettent de répondre à l'enjeu 8 (Note 2).

Sur les 10 jeux de données inventoriés, 6 ont pu être intégrés la base de données finale. Malgré les efforts, certains jeux de données n'ont pas pu être récupérés. C'est le cas notamment de `JONATHAN-Sept-Îles` et `ERMMA` (Centre de la Mer de Biarritz). D'autres, comme `CEM_EI_digital_hytech`, malgré qu'ils aient été en notre possession, n'ont pas été intégrés car ils étaient redondants spatialement et temporellement avec d'autres jeux de données déjà intégrés (`CEM_EI_digital`, `CEM_EI_plane` et `CEM_EI_boat`) et parce qu'ils étaient moins intéressants du point de vue des enjeux sur la saisonnalité (une seule année contre 3 pour `CEM_EI_plane` et `CEM_EI_boat`).

3.3 Données de présence seulement

Pour les données de présence seulement, le travail de hiérarchisation s'est limité à prioriser les sources de données les plus volumineuses, les plus faciles à mobiliser et les plus pertinentes par rapport aux mammifères marins, puis à écarter les sources présentant des risques importants de redondance. En effet, l'une des difficultés pour ce type de données est que les plateformes qui les hébergent peuvent échanger des données entre elles. Par exemple, le GBIF intègre les données d'iNaturalist. De plus, il est possible que les données fournies dans ces plateformes soient soumises à plusieurs licences d'utilisation. Pour la collecte de ces données, l'enjeu n'était donc pas de multiplier les sources possibles, car il y avait un fort risque d'introduire des doublons et l'investissement en temps pour se renseigner auprès des fournisseurs sur les licences d'utilisation aurait été trop important. Ainsi, trois principales sources ont été identifiées : **Global Biodiversity Information Facility (GBIF), iNaturalist et OBSenMER**.

3.3.1 Global Biodiversity Information Facility

> "Le GBIF – Global Biodiversity Information Facility (= Système mondial d'information sur la biodiversité) – est un réseau international et une infrastructure de données financés par les gouvernements mondiaux ayant pour but de fournir à tous et partout un accès libre aux données sur toutes les formes de vie sur Terre."
> — gbif.org

Il s'agit de la plateforme mondiale de référence pour les données de biodiversité et, de loin, la source la plus volumineuse. Elle est régulièrement alimentée par les données de nombreuses sources publiques et privées et dispose d'une interface de programmation (API) qui peut être interrogée directement depuis R avec les packages `{rgbif}` (Chamberlain et al. 2025) ou `{spocc}` (Owens, Barve, et Chamberlain 2024). La principale difficulté posée par cette plateforme est l'hétérogénéité de la qualité des données compte tenu de la diversité des sources qu'elle intègre.

3.3.2 iNaturalist

> "iNaturalist regroupe plusieurs fonctionnalités, mais il s'agit essentiellement d'un réseau social en ligne où les utilisateurs partagent des informations sur la biodiversité pour s'entraider et en apprendre davantage sur la nature. C'est également un système participatif d'identification des espèces et un outil d'enregistrement des occurrences d'organismes."
> — inaturalist.org

Cette plateforme américaine de science citoyenne bénéficie d'un large réseau de contributeurs dans le monde. Elle fournie des données ouvertes et dispose d'une API qui peut être interrogée directement depuis R avec les packages `{rinat}` (Barve et Hart 2022) ou `{spocc}` (Owens, Barve, et Chamberlain 2024). Deux caractéristiques intéressantes de cette plateforme :

1. L'existence d'une intelligence artificielle qui aide à déterminer les espèces
2. La présence d'un label de "qualité recherche".

Le label "qualité recherche" est une information très utile car elle permet de permet de réaliser un tri efficace des observations susceptibles d'être pertinentes dans le cadre du projet MODELCET ATL. En effet, ce label indique que l'observation est jugée suffisamment précise et vérifiée par la communauté pour être utilisée dans des projets scientifiques. Pour obtenir ce statut, une observation doit être explicitement marquée comme "sauvage", accompagnée d'une preuve visuelle ou sonore, avoir une date et des coordonnées géographiques précises, être identifiée au niveau de l'espèce (ou plus précis) et cette identification doit être confirmée par au moins deux utilisateurs différents. Les données de qualité recherche sont transmises au GBIF, il y a donc une forte redondance et seules les données les plus récentes sont inédites.

3.3.3 OBSenMER

> "OBSenMER est une plateforme collaborative qui facilite la saisie et l'analyse des observations en mer. Elle concerne tous les types d'observations : mammifères marins, tortues marines, poissons, oiseaux, mais aussi les activités humaines, telles que la plaisance, la pêche, la pollution, etc.
> OBSenMER s'adresse aussi bien au grand public qu'aux structures naturalistes professionnelles, et collecte aussi bien les données de sciences participatives que les données issues d'un protocole expert."
> — obsenmer.org

Bien que moins volumineuse que le GBIF ou iNaturalist, cette plateforme accueille une quantité importante de données sur le territoire français, toutes libres d'utilisation. Par ailleurs, c'est une plateforme soutenue financièrement par l'État français, et à laquelle les agents des réserves naturelles

marines contribuent. Une autre particularité intéressante de cette plateforme est qu'elle dispose de données "semi-protocolées" (observations "avec effort" et observations "expertes", GECC (2025)). Il s'agit de données pour lesquelles l'effort est connu, sans pour autant que cet effort soit strictement encadré. Par exemple, l'effort n'est pas nécessairement constitué de transects linéaires, ne suit pas forcément un plan d'échantillonnage planifié ou réfléchi, et l'utilisation de protocoles standards tels que le distance sampling ou le strip-transect n'est pas exigée.

OBSenMER ne dispose pas d'une API mais ces données sont librement téléchargeables depuis la plateforme après y avoir créé un compte pour s'authentifier. Les administrateurs de la plateforme ont tout de même été contactés par écrit afin d'être informés du projet MODELCET et de confirmer leur accord quant à l'utilisation de ces données.

3.3.4 Sources non sélectionnées

OpenObs

> OpenObs permet de visualiser et de télécharger les données d'observation sur les espèces disponibles dans l'Inventaire national du patrimoine naturel, plateforme nationale du SINP (Système d'Information de l'iNventaire du Patrimoine naturel).

Les données présentes dans OpenObs sont également publiées dans GBIF, à l'exception les données sensibles, ou privées qui sont des données ayant une diffusion restreinte. Par ailleurs, si OpenObs dispose d'une API, il n'existe à ce jour aucun package R qui permet d'interagir avec directement. Cette base de donnée n'a donc pas été sélectionnée mais la plupart de ses données ont été collectées indirectement via le GBIF.

ORCA-PO

Concernant les données de l'association caritative britannique ORCA, nous avons décidé de nous concentrer sur les données issues du programme "OceanWatchers". Ces données sont collectées par les passagers et équipages sur les navires de croisière et les ferries qui traversent la Manche et le golfe de Gascogne. Cette méthode de collecte permet de disposer d'informations sur l'effort d'échantillonnage ainsi que d'une meilleure couverture des secteurs hauturiers. Ces données ont été classées comme protocolées. Cette source inclue les données collectées par l'association `OCEAMM` (Observatoire pour la Conservation et l'Étude des Animaux et Milieux Marins).

Autres sources

Les sources `Faune-Océan`, `GMN` (Groupe Mammalogique Normand), `LPO` (Ligue de Protection des Oiseaux), `SiRF` (Système d'Information Régional sur la Faune – Groupe Ornithologique et Naturaliste), `Al-Lark` et `APECS` (Association Pour l'Étude et la Conservation des Sélaciens) ont été contactées. Certaines ont répondu favorablement, d'autres n'ont pas répondu aux sollicitations. Finalement, nous avons estimé que le compromis entre temps de travail et la valeur ajouté par ces sources n'était pas assez favorable compte tenu de nos objectifs et du volume de données, d'importance secondaire pour ces sources.

4 Collecte des données

4.1 Procédure de demande

Pour les sources de données externes au consortium, les organismes propriétaires et dépositaires de données ont été contactés. La prise de contact a été faite via un courrier type présentant le projet MODELCET ATL, le consortium et les grandes lignes concernant l'utilisation des données qui est envisagée.

Objet : Demande de données pour le projet MODELCET - Façade Manche Atlantique
Madame, Monsieur,
Nous nous permettons de solliciter votre assistance dans le cadre du projet MODELCET, qui vise à concevoir un modèle de distribution et d'abondance des cétacés sur la façade Manche Atlantique afin d'améliorer la prise de décisions et la mise en œuvre des politiques publiques de préservation de ces espèces.
Le projet MODELCET (Façade Manche Atlantique) a fait l'objet d'un appel d'offre en 2023 porté par l'**OFB**. Ce marché public a été remporté par un consortium composé d'un chercheur de l'**observatoire Pelagis** (mandataire), Matthieu Authier, d'une chercheuse indépendante en statistique et écologie quantitative, **Floriane Plard** (cotraitante) et d'un bureau d'étude en ingénierie écologique, **Biotope SAS** (co-traitant). Le projet a officiellement commencé en février 2024 et s'achèvera en début d'année 2027. L'accès à un maximum de données d'observations de cétacés sera essentiel afin d'améliorer la précision et la fiabilité des prédictions du modèle. Par ailleurs, le projet a également pour but d'étudier la pertinence et les avantages d'intégrer diverses sources de données.
Dans ce contexte, nous souhaitons ouvrir avec vous un échange concernant le partage des données d'observations de cétacés dans la région Manche Atlantique qui seraient en votre possession. Si de telles données existent, le but de cet échange serait de convenir des conditions dans lesquelles les données pourraient nous être transmises et être potentiellement utilisées par le projet MODELCET.
Ainsi, dans un premier temps, nous souhaiterions vous poser quelques questions telles que :

- Comment les données ont-elles été collectées : ont-elles été collectées de manière protocolée et sont-elles accompagnées d'un effort d'échantillonnage connu ?
- Pouvez-vous décrire le protocole utilisé pour collecter les données ?
- Les observations ont-elles été collectées ou vérifiées par des experts qualifiés ?
- Quelle(s) année(s) les données ont-elles été collectées ?
- Quelles zones géographiques et périodes de l'année les données couvrent-elles ?
- Quel est l'ordre de grandeur du nombre d'observations ?
- Combien d'observateurs sont présents lors de la collecte des données ? Est-ce systématique ?

Dans le cadre de MODELCET, le consortium mandaté s'engage à maintenir une grande transparence quant aux sources des données qui seront utilisées et aux choix méthodologiques. Les méthodes développées et les résultats obtenus seront accessibles au grand public. Les résultats seront notamment diffusés par le biais d'une publication scientifique. Nous croyons fermement que cette transparence est essentielle pour garantir la robustesse de nos résultats, assurer leur validité scientifique et faciliter leur utilisation pour la préservation des espèces marines. Cette politique de transparence ne s'applique pas nécessairement aux données brutes qui nous seraient fournies. En effet, ces dernières si vous nous laissez la possibilité de les utiliser, resteraient sous votre propriété et nous nous engageons à respecter strictement vos exigences en matière de confidentialité et de sécurité, ainsi que les réglementations en vigueur concernant la protection des informations sensibles. Par ailleurs, les données transmises seront exclusivement utilisées dans le cadre du projet MODELCET.
Nous vous serions très reconnaissants de bien vouloir répondre à cette sollicitation dans de brefs délais afin de nous aider à garantir la progression efficace du projet. Vous pouvez nous contacter à l'adresse électronique `courriel` ou par téléphone au `numéro de téléphone`.
Dans l'attente de votre réponse, nous vous prions d'agréer, Madame, Monsieur, l'expression de nos salutations distinguées.

Quand cela a été nécessaire, les contacts ont été relancés et des contacts alternatifs recherchés. L'intégration des jeux de données ayant reçu un retour négatif ou bien pour lesquels les demandes sont restées sans réponse après deux relances, a été abandonnée.

Par ailleurs, une période de plusieurs mois doit être réservée pour l'harmonisation des données et la

consolidation de la base de données finale. Nous avons donc dû établir une date butoir. Initialement, nous nous limitions aux données collectées sur le terrain avant 2024. Compte tenu du temps nécessaire à la collecte des données et du retard que cela a pu causer au projet, nous avant décidé d'étendre l'harmonisation aux données collectées en 2024 tant que celles-ci peuvent être livrées le 31 décembre 2024. Finalement les toutes dernières données que nous avons réussi à intégrer à la base ont été reçues le 3 mars (certaines sources de Pelagis sujettes à des problèmes techniques et devant faire l'objet de vérifications internes).

4.2 Conditions d'utilisation

Pour les sources dont les données brutes étaient déjà en la possession du consortium, la phase de collecte s'est limitée à obtenir une autorisation explicite et écrite d'utiliser les données dans le cadre du projet MODELCET ATL. Dans chaque cas, il a été expliqué que les méthodes développées et les résultats obtenus dans le cadre du projet MODELCET ATL (les "données résultats") seront accessibles au grand public, mais que cette contrainte ne s'applique pas forcément aux données brutes.

Certains jeux de données ont dû faire l'objet d'une "convention de données" précisant les conditions d'utilisation. C'est le cas par exemple pour le jeu de données `Grand-Large`. Il s'agit néanmoins d'une procédure chronophage et faisant appel à des compétences juridiques. En absence de démarche de la part du propriétaire des données pour préciser en détail les termes de l'utilisation des données, une licence d'utilisation par défaut conservatrice (nommée "MODELCET-restricted") a été utilisée :

> Data utilization granted to the MODELCET consortium for the duration of the MODELCET project. These data may not be used or redistributed without notice and explicit authorization of the owner.

L'objectif est de s'assurer que chaque source est accompagnée, dès le départ, d'un cadre juridique assurant à la fois la protection des droits du propriétaire et la sécurité d'usage pour les membres du consortium. Néanmoins, compte tenu de la nature générique et restrictive de ces conditions d'utilisation, cette licence constitue un point de départ et les termes sont susceptibles d'être révisés au cours de la vie du projet.

5 Harmonisation des données

L'objectif du travail d'harmonisation de jeux de données est de garantir que les données de différents types (protocolé et présence seulement), différentes plateformes (bateau, avion), différentes méthodes d'observation (visuel et digital) et différents protocoles (strip-transect, distance-sampling) puissent être combinées et analysées dans un même modèle, sans dégrader la qualité ou compromettre la validité des données. Cela implique des travaux interdépendants à l'échelle de chaque jeu de données et à l'échelle de l'ensemble des données, où chaque changement concernant la structure des données ou la nomenclature doit être propagé au reste des données :

- Au sein d'un jeu de données, uniformiser et vérifier la cohérence des informations entre données d'effort, observations et métadonnées. Par exemple, les termes communs (tels que le nom du jeu de données), les unités, les noms de colonne…
- Entre les jeux de données, uniformiser les formats et l'organisation. Par exemple s'assurer que les données essentielles sont présentes dans chaque jeu de données, que les couches effort/observation soient nommées de la même façon, que leur structure soit identique, que les noms et les types et les unités des variables soient identiques…

5.1 Flux de travail pour l'ajout d'un nouveau jeu de données

Ce que l'on rattache concrètement à la notion de "donnée" varie en fonction qu'il s'agisse de données protocolées ou de présence seulement. En effet, si pour les données de présence seulement, nous pouvons nous contenter de réunir une quantité d'information sommaire autour de l'occurrence (à minima la localisation et la date), les données protocolées doivent être collectées avec une mesure de l'effort d'échantillonnage (surface ou longueur de transect prospectée) que l'on essaiera de renseigner avec la résolution spatiale la plus fine possible.

Du fait de ces différences, les deux types de données (protocolées et présence seulement) ont fait l'objet de procédures de collecte et de nettoyage distinctes.

Une fois les niveaux de priorités fixés, un flux de travail a été défini pour tout jeu de données que l'on souhaite intégrer. L'ajout d'un jeu de données nécessitait les étapes de travail suivantes :

1. Création d'une branche nommée de façon standardisée `feat-data-dataset`. Chaque jeu de données porte un nom qui est utilisé comme un identifiant tout au long du travail (voir Note 3, `dataset` dans l'exemple précédent).
2. Ajout des données brutes au versionnement des données avec DVC. Cette étape assure la sauvegarde des données brutes telles qu'elles ont été reçues sur un serveur.
3. Création d'un fichier de métadonnées (voir Section 5.3.1).
4. Si nécessaire, une étape de pré-traitement des données pour les faire correspondre à un fichier de spécifications pré-existant (voir exemple de fichier dans le Listing 3), pour corriger des erreurs ou ajouter des informations. Cette étape est parfois nécessaire pour s'épargner d'importantes complications dans le code commun à toutes les sources de données. Toutefois, elle est évitée le plus possible afin de limiter au maximum le nombre de scripts et copies de code difficile à maintenir.
5. Création d'un fichier de spécifications pour la prise en charge des données par le code (voir Section 5.3.2).
6. Mise à jour de la liste des synonymes taxonomiques (voir Section 5.3.3), si besoin.
7. Ajout du jeu de données à la liste des jeux de données à traiter (`params.yaml`, section `datasets`)
8. Mise à jour de la chaîne de traitement DVC
9. Renseignement du fichier `NEWS.md` qui contient un résumé clair, non technique et à destination de l'utilisateur final des évolutions du projet. Dans ce contexte, il sera spécifié l'ajout du nouveau jeu de données dans la catégorie appropriée.
10. Demande de fusion de la branche de travail `feat-data-dataset` vers la branche `dev` pour procéder à un premier contrôle qualité interne.

Les étapes 1, 2, 3 7, 8, 9 et 10 sont obligatoires alors que les étapes 4 à 6 sont fonction des caractéristiques du jeu de données concerné et des précédents déjà intégrés. L'ensemble de ce processus est documenté dans l'historique Git du projet (voir Figure 4), ce qui permet d'assurer traçabilité, reproductibilité et transparence dans le cycle de vie des données.

> **Note 3 : Nomenclature des jeux de données**
>
> A chaque jeu de données est associé un nom unique et concis.
> Ce nom est notamment utilisé dans le jeu de données lui-même (colonne `SURVEY_ID`), dans les noms de fichier d'entrée et de sortie et dans le fichier d'inventaire des sources.
> Pour les données de suivis des parcs éoliens offshore, la nomenclature est composée des abréviations suivantes (séparées par un underscore, _) :
> — L'abréviation du nom du parc éolien (par exemple `CEM` pour Centre Manche), l'abréviation de l'étape du projet (par exemple `EI` pour état initial) et l'abréviation de la

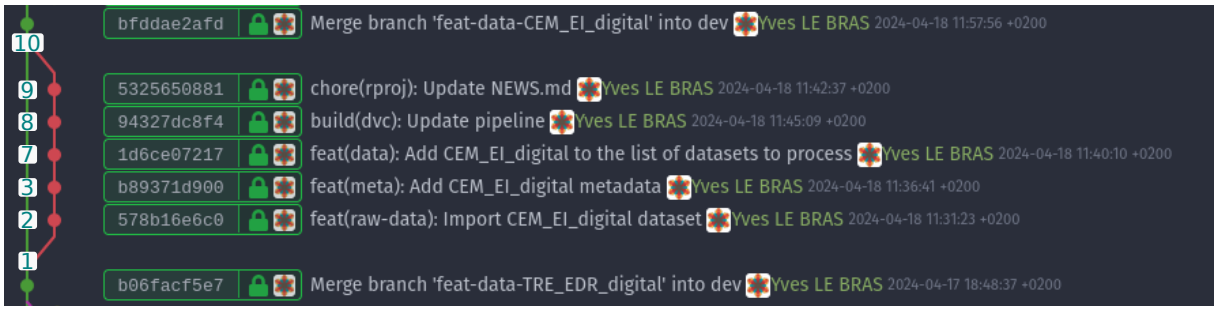


FIGURE 4 – Illustration du flux de travail tel qu'il se présente concrètement dans l'historique git du projet, ici concernant l'intégration du jeu de données `CEM_EI_digital`. L'ensemble de modifications relatives à l'intégration de ce jeu de données peuvent être consultées (et au besoin supprimées) au niveau du commit de fusion `bfddae2afd`.

méthode d'observation (par exemple digital pour les données de suivis aériens digitaux haute altitude). Nous indiquons en annexe la liste de ces abréviations.

— Il peut arriver que ce nom soit suivi d'un numéro (par exemple, `TRE_EI_plane2` et `TRE_EI_plane3`). Cela permet de distinguer des jeux de données qui ont été produits pour le même parc, à la même phase du projet et avec la même méthode d'observation mais qui se distinguent dans la façon de mettre en place le protocole, les outils utilisés ou les périodes de temps échantillonnées. Il faut alors se référer aux informations fournies dans le fichier de métadonnées pour déterminer ce qui distingue les deux jeux de données. Par exemple, dans le cas du parc de Dieppe - Le Tréport, les suivis aériens visuels pour l'état initial ont eu lieu en deux temps, une session en 2010-2011 et une autre en 2014-2015.

5.2 Extraction des données d'effort et des observations

Comme indiqué précédemment dans ce document, nous avons créé un ensemble de fonctions et de scripts R permettant d'effectuer tous les traitements nécessaires pour passer des données brutes (disponibles dans le `data/cetaceans/raw`) aux données formatées (disponibles dans le dossier `data/cetaceans/prepared`).

Ces traitements dépendent des types de données (protocolées ou présence seulement), des plateformes d'observation (bateau ou avion) et des protocoles employés (distance sampling ou strip-transect). En fonction de ces paramètres, différentes fonctions vont prendre en charge et formater les jeux de données. Ces fonctions permettent de prendre en compte les spécifications communes à 6 grands types de données :

— Données nautiques
— Données aériennes avec un protocole distance-sampling
— Données aériennes avec un protocole strip-transect
— Données aériennes digitales
— Données présence seulement (hormis OBSenMER)
— Données OBSenMER

Pour chaque jeu de données, l'appartenance à l'une de ces catégories est spécifiée dans la variable `survey_type` des fichiers de métadonnées.

5.2.1 Données d'effort

L'existence de ces données distingue les données "protocolées" des données de "présence seulement". Les données protocolées doivent être collectées avec une mesure de l'effort d'échantillonnage qui doit impérativement être harmonisée et intégrée à la base de données. Cette mesure de l'effort varie en fonction du protocole utilisé. Par exemple

- Pour les données de suivis digitaux, l'effort d'échantillonnage est une surface, celle correspondant à l'emprise de l'image qui a été analysée.
- Pour les autres méthodes d'observation, il s'agit de la longueur de transect parcouru. L'effort surfaçique est obtenu dans un second temps, en utilisant le paramètre de largeur de bande renseigné dans le fichier de métadonnées (strip-transect) ou à l'issue d'une analyse distance sampling.

Dans tous les cas, nous avons essayé de renseigner l'effort d'échantillonnage avec la résolution spatiale la plus fine possible, de manière à ce que le choix de la résolution spatiale par le modélisateur ne soit pas "artificiellement" contraint par les limites des jeux de données qui auraient été dégradés durant l'harmonisation.

Les données d'effort ne se limitent pas à connaître le lieu et l'horaire de l'échantillonnage. Elles comprennent également un ensemble d'informations contextuelles, les conditions d'observation, qui peuvent influencer fortement la détection et sont donc essentielles à la bonne interprétation des données dans les modèles (voir UO 1.2.1 "Description des méthodes de couplage et de modélisation"). Ainsi, en complément des mesures d'effort d'échantillonnage, des conditions d'observation sont également collectées lorsque disponibles. Ces facteurs sont nombreux et les informations disponibles divergent selon les sources de données. Notre effort d'harmonisation a porté surtout sur les variables les plus communément renseignées. Voici la liste des variables retenues :

- `ELEVATION` : *Hauteur de l'observateur ou du capteur au-dessus de la surface de la mer.* Que ce soit en avion ou en bateau, la hauteur de l'observateur joue un rôle important dans la détection visuelle. Cette variable influence le champ de vision et la distance de détection maximale.
- `GLARE_SEVERITY` : *Niveau d'éblouissement causé par la réverbération du soleil sur la surface de l'eau*. Typiquement exprimé de manière semi-quantitative (par exemple : faible, modéré, sévère), l'éblouissement réduit la visibilité et peut fortement affecter la probabilité de détection.
- `SEASTATE_BEAUFORT`, `SEASTATE_PETERSEN`, `SEASTATE_DOUGLAS` : *État de la mer codé selon les échelles de Beaufort*, *Petersen et Douglas.* L'état de la mer est facteur très influant sur la détectabilité des mammifères marins. Il existe différente échelles pour le qualifier. L'échelle de Beaufort va de 0 (calme plat) à 12 (ouragan) et c'est de loin la plus commune dans les données rencontrées. L'échelle de Petersen met davantage l'accent sur les caractéristiques visuelles de la houle plutôt que sur la vitesse du vent. Pour finir l'échelle de Douglas, utilisée plus rarement, distingue l'état de la mer du vent de celui de la houle.
- `TURBIDITY` : *La turbidité de l'eau*, correspondant à la concentration de particules en suspension qui réduit la pénétration de la lumière et la visibilité sous la surface. Ce paramètre peut se révéler important pour les observations aériennes et dans les zones où ce paramètre est très variable.
- `SUBJECTIVE` : *Appréciation qualitative globale par l'observateur*. Bien que cette variable soit la plus subjective, elle est très utile car c'est l'une des informations les plus communes et elle est intégrative de l'ensemble des facteurs influençant la détectabilité. Toutefois, selon l'objectif scientifique des suivis (e.g. centrés sur les mammifères marins ou sur les oiseaux marins), cette variable peut revêtir différentes interprétations.

5.2.2 Observations biologiques

C'est le socle commun à toutes les données, quelle que soit leur origine ou leur protocole. Les observations biologiques sont les détections d'individus ou de groupes d'animaux. Dans le cas des données protocolées, elles sont accompagnées d'informations permettant de les associer à une unité d'effort.

Les principales variables harmonisées dans les observations biologiques sont les suivantes :

- `TAX_ORIGIN` : *Groupe taxonomique observé, tel que renseigné dans les données brutes.* Ces données sont ensuite harmonisées et déclinées en différents niveaux (`TAX1`, `TAX2`, `TAX3` etc.) via l'utilisation d'un arbre des synonymes taxonomiques commun à tous les jeux de données. L'utilisation d'un arbre permet d'obtenir en sortie des données compatibles avec des représentations ou des modélisations à différentes échelles, en fonction des besoins de la rareté du groupe taxonomique considéré ou des confusions d'identifications possibles.
- `GROUP_SIZE` : *Nombre d'individus observés*. Cette variable est parfois une estimation (cas des grands groupes ou de l'observation aérienne), parfois un comptage exact.
- `DATE`, `TIME` : *Date et heure précises de l'observation*. Ces informations peuvent être utiles pour relier les observation aux unités d'effort et sont indispensables pour extraire les conditions environnementales "dynamiques" pertinentes.
- `LON` / `LAT` : *Coordonnées géographiques de l'observation*. Ces données sont toujours rapportées en WGS84 et exprimées en degrés décimaux (code EPSG 4326).
- `PERP_DIST` : *Distance orthogonale de l'observation par rapport au transect*. Cette variable est nécessaire pour estimer la probabilité de détection selon la méthode distance sampling. Elle est exprimée en mètres, mesurée perpendiculairement au transect suivi, et n'est disponible que pour les observations issues de protocoles adaptés à cette méthode.

D'autres informations importantes, telles que la plateforme d'observations, la méthode d'observation ou les modalités d'implémentation du protocole sont communes à toutes les observations d'un même jeu de données, et figurent plutôt dans un fichier de métadonnées (voir Section 5.3.1 et Listing 2).

5.2.3 Jointure entre les observations et l'effort d'échantillonnage

Plusieurs variables peuvent être communes entre le tableau des observations et celui de l'effort d'échantillonnage (e.g. la date, l'heure, le transect, etc.). Il était donc capital d'identifier les variables communes qu'il faut utiliser pour réaliser la jointure entre le tableau des observations et celui de l'effort d'échantillonnage.

> **Note**
>
> Chaque unité d'effort n'est pas nécessairement associée à une observation de mammifère marin. En revanche, chaque observation collectée dans le cadre d'un protocole doit pouvoir être associée à un effort d'échantillonnage.

- `LEG_ID` : *Variable de jointure des tableaux d'observations et d'effort.* Dans la base de données harmonisées, quatre colonnes commençant par `LEG_ID` sont disponibles : `LEG_ID`, `LEG_ID1`, `LEG_ID2` et `LEG_ID3`). Le nombre de variables nécessaire pour réaliser la jointure varie en fonction des bases de données. Si une seule variable est nécessaire, elle est disponible dans la colonne `LEG_ID` et les autres colonnes sont vides. Si plus d'une colonne est nécessaire pour réaliser la jointure, la colonne `LEG_ID` est vide et les colonnes `LEG_ID1`, `LEG_ID2` (voir même `LEG_ID3`) sont renseignées.

5.3 Données de configuration

5.3.1 Métadonnées

Nous avons également veillé à réunir le plus d'informations possible concernant les plateformes et les méthodes d'observations ainsi que les protocoles employés. Pour chaque jeu de données intégré, nous avons récupéré le rapport de la mission si existant, ou nous avons exploré les informations disponibles sur internet à travers les articles scientifiques ayant utilisé le jeu de données ou les sites officiels des propriétaires des données (exploitant éolien, association, etc). Chaque lien internet utilisé comme source d'information est référencé dans les métadonnées.

Les fichiers de métadonnées contiennent deux grands types d'informations :

1. Des informations générales, utiles au référencement des jeux de données : titre, description, sources des données, conditions d'utilisation, citation du document de référence, liens utiles, le type de suivi… Un exemple est présenté en annexe (voir Listing 1).
2. Des informations plus techniques concernant l'échantillonnage : détails du protocole, des dates et de la zone géographique échantillonnées… Un exemple est présenté en annexe (voir Listing 2).

> **Note 4 : Métadonnées et plan de gestion des données**
>
> Les métadonnées, en particulier celles présentées dans le Listing 1, entrent dans la constitution du plan de gestion des données. Le livrable UO 1.1.3 renvoies aux métadonnées consultables sur le dépôt Gitlab de l'OFB.

5.3.2 Spécification du format des données

Les différents jeux de données partagent des informations communes mais rarement sous le même nom ou sous le même format. Une tache récurrente lors de l'harmonisation des données consiste à renommer certaines variables de façon à ce que les mêmes informations portent le même nom dans la base de données finale. De même, certaines variables, comme la date ou l'heure, existent sous une diversité de formats qu'il convient d'harmoniser dans la base de données finale.

Ces taches sont réalisées automatiquement à partir d'un fichier de configuration, dit de "spécification du format des données", permettant de renseigner comment les variables et leur format doivent être harmonisés. L'utilisation d'un tel fichier permet de réutiliser le même fichier plusieurs fois quand plusieurs sources de données partagent les mêmes caractéristiques. Ces spécifications sont inscrites dans les fichiers yaml du dossier `data/cetaceans/metadata/data_specs`. Par exemple, le fichier `data/cetaceans/metadata/data_specs/bioconsult_digital.yaml` permet de spécifier tous les jeux de données aériens digitaux de haute altitude soit 8 jeux de données (`NOY_EDR`, `SNA_Con_digital`, `FEC_Con_digital`, `TRE_EDR_digital`, `CEM_EI_digital`, `BOY_digital_Bourgneuf`, `BOY_digital_Yeu` et `MMN`). Ajouter un jeu de données de ce type demande donc un temps de travail minime une fois la spécification écrite et la vérification qu'elle corresponde au jeu de données. À l'inverse, certains jeux de données ont demandé une attention particulière car ils ne correspondaient à aucune spécification déjà écrite et leurs spécifications n'ont été partagées par aucun autre jeu de données par la suite. C'est le cas par exemple des jeux de données `B2B_EI_plane`, `SNA_EDR` ou `ZPS_BOY_Periscope`.

Ces spécifications se divisent en deux catégories : celles concernant l'effort d'échantillonnage et celles concernant les observations. Certaines variables, communes aux deux parties, servent de base pour la jointure des données d'effort et d'observations (colonnes numérotées commençant par `LEG_ID`, cf. Section 5.2.3). Un exemple de fichier de spécification du format des données est consultable en annexe Listing 3.

5.3.3 Synonymes taxonomiques

L'arbre des synonymes taxonomiques (fichier `data/cetaceans/metadata/synonyms.yaml`) permet d'harmoniser les très nombreuses variations dans les codes des groupes taxonomiques présents dans les données brutes. Il assure la standardisation des noms d'espèces et réalise les regroupements taxonomiques. Grâce à une structure arborescente, il est possible de faire correspondre à chaque étiquette taxonomique rencontrée une succession de noms standardisés et nichés les uns dans les autres. Cette approche nous a semblé intéressante pour pouvoir s'adapter à différentes qualités de l'identification en conservant une information hiérarchique cohérente. En outre, il est utile de disposer d'étiquettes taxonomiques à différents niveaux emboîtés, par exemple on peut vouloir réaliser des cartes des données à différentes échelles spatiales, ou modéliser les données à différents niveaux taxonomiques selon la rareté des taxons, leur proximité écologique ou les risques de confusion dans l'identification. Notons toutefois que le regroupement des taxons à dire d'expert, n'est pas pertinent, peut introduire des biais lors de la modélisation (Plard et al. 2024). Il ne s'agit donc pas seulement d'une tâche de mise en forme des données mais également d'un choix de modélisation. A ce titre, ce fichier est susceptible d'évoluer même après la livraison de la base de données (UO 1.1.2 & 1.1.4).

Ce fichier est trop long pour être reproduit ici ou dans les annexes de ce document mais, comme tous les fichiers de configuration employés, il peut être consulté sur le dépôt gitlab de l'OFB.

6 Assemblage des données

Les objectifs de l'assemblage sont les suivants :

- Combiner ensemble les différentes sources et les différentes couches (observation, effort, métadonnées).
- Identifier et documenter les incohérences, c'est-à-dire détecter et décrire les contradictions ou valeurs aberrantes au sein des jeux de données.
- Vérifier les éventuels problèmes de jointure entre les observations et l'effort d'échantillonnage.

6.1 Concaténation des données

Les jeux de données harmonisés et les métadonnées indispensables à leur exploitation ont été assemblés dans un unique fichier appelé `modelcet_atl` au format GeoPackage. Ce fichier est composé de trois couches :

- une couche géospatiale avec l'effort d'échantillonnage appelée *Effort*.
- une couche geospatiale avec les observations appelée *Sightings*.
- une couche avec les métadonnées les plus importantes pour l'exploitation de la base de données, appelée *Metadata*.

L'assemblage a nécessité la création d'un ensemble de fonctions et de scripts R permettant d'effectuer tous les traitements nécessaires pour passer des données formatées (disponibles dans le dossier `data/cetaceans/prepared`) aux données assemblées (disponibles dans le dossier `data/cetaceans/combined`).

L'objectif étant que les jeux de données composant la base de données finale soient exploitables comme un tout et non individuellement, la phase d'assemblage a conduit à une simplification et donc à la perte de certaines variables. Les traitements de la phase d'assemblage ont notamment permis de **sélectionner les variables à conserver** et écarter les autres. Ils ont également permis d'**homogénéiser les types des variables**. Cette étape consiste à vérifier que les colonnes dans lesquelles on s'attend, par exemple, à trouver une valeur numérique, ne contiennent pas une chaîne de caractères, ou inversement (les types pouvant être rencontrés dans le cadre de la base de données MODELCET-ATL sont les suivants : numérique, caractère, booléen et date).

6.1.1 Homogénéisation de la taille des groupes

Une information capitale contenue dans le tableau des observations est la taille du groupe observé. Cette variable contenait généralement un nombre entier mais elle pouvait également contenir une valeur qui semble aberrante (9999 ou 0) ou un intervalle indiquant la limite basse et haute de la taille du groupe observé. Afin de ne pas dégrader l'information et laisser la décision de la manière la plus appropriée de gérer ces valeurs à l'utilisateur de la base de données, nous avons créé trois variables :

- `GROUP_SIZE` : *contient un nombre entier correspondant à la taille du groupe observé*. Cette variable correspond soit la valeur d'origine lorsqu'il s'agissait d'un nombre entier (même s'il s'agit d'une valeur aberrante), soit la médiane arrondie à la valeur paire la plus proche lorsqu'un intervalle était fourni. Lorsqu'il s'agit de la médiane ou que la valeur est jugée aberrante, on considère que l'information a été dégradée ou est de mauvaise qualité. La colonne `WARNING` est alors utilisée pour en informer l'utilisateur (cf. Section 6.1.2).
- `GROUP_SIZE_LOWER` : *indique la limite basse de l'intervalle définissant la taille du groupe observé dans les données d'origine.*
- `GROUP_SIZE_UPPER` : *indique la limite haute de l'intervalle définissant la taille du groupe observé dans les données d'origine.*

6.1.2 Valeurs erronées ou manquantes

Au cours de la constitution de la base de données, nous nous sommes rendus compte de la présence d'erreurs ou de données manquantes dans certains jeux de données que nous avions intégrés. Il pouvait s'agir de colonnes manquantes, d'incohérences dans les données de jointures, d'erreurs d'encodage, d'unités ambiguës...

Au lieu de supprimer les jeux de données, nous avons décidé de les conserver mais d'ajouter des variables dans la base de données finale qui permettront de repérer rapidement les erreurs, de les comprendre et, si cela s'avère nécessaire, de filtrer les données pour supprimer les lignes incriminées. Ces variables ont donc pour but de faciliter la validation et de dialogue avec les parties prenantes (autres membres du consortium, fournisseurs des données, OFB). Elles ont été ajoutées au tableau des observations et sont au nombre de trois :

- `WARNING` : *indique la présence/absence (TRUE/FALSE) d'erreur ou d'information manquante dans la ligne*. Cette information encodée par un booléen permet de filtrer rapidement les données en fonction de la présence/absence d'erreur. Avant de filtrer les observations, nous conseillons de vérifier si l'erreur ou la donnée manquante est de nature à dégrader l'utilisation que l'on entend faire de la base de données. Pour cela, il est possible de se référer aux deux autres variables.
- `WARNING_COLUMN` : *indique la ou les colonne.s concernée.s par l'information erronée ou manquante*. Cette variable qui prend la forme d'une chaîne de caractères n'est renseignée que lorsque `WARNING = TRUE`. Les valeurs possibles sont les suivantes (ou une combinaison de ces valeurs) : `TRANSECT_ID`, `GROUP_SIZE` et `LEG_ID*` (l'étoile de `LEG_ID*` permet de d'indiquer qu'au moins une colonne commençant par `LEG_ID` contient une information erronée ou manquante). Lorsqu'une erreur ou une donnée manquante est repérée dans les colonnes commençant par `LEG_ID`, la jointure entre les observations et l'effort d'échantillonnage ne peut pas se faire.
- `WARNING_DETAIL` : *apporte des informations complémentaires sur l'erreur ou la données manquante*. Cette variable qui prend la forme d'une chaîne de caractères n'est renseignée que lorsque `WARNING = TRUE`. Les valeurs possibles sont les suivantes (ou une combinaison de ces valeurs) :
  - *TRANSECT_ID value is unknown (NA)*
  - *GROUP_SIZE value is unknown (NA) or incoherent*

— *GROUP_SIZE is a rounded median*
— *At least one of the LEG_ID value is unmatched in the effort table (lead to join issues with effort dataset)*
— *At least one of the LEG_ID value is missing (lead to join issues with effort dataset)*

7 Description de la base de données

L'objectif est de présenter succintement la base de données qui constitue le livrable UO 1.1.4 du projet MODELCET ATL. Cette présentation donne une visualisation à un instant donné de la base de données. Il est possible qu'au cours du projet, cette base de données évolue. De plus, les observations de Pinnipèdes, bien que présentes dans la base de données protocolées, ne sont pas prises en compte dans la description des données car elles ne seront pas utilisées dans le cadre du projet MODELCET ATL.

La base de données réunit au total 65 jeux de données, 62 jeux de données protocolées et 3 jeux de données de présence seulement qui représentent la majeur partie des observations :

— 39207 observations proviennent des jeux de données protocolées (voir Figure 5 et Figure 6).

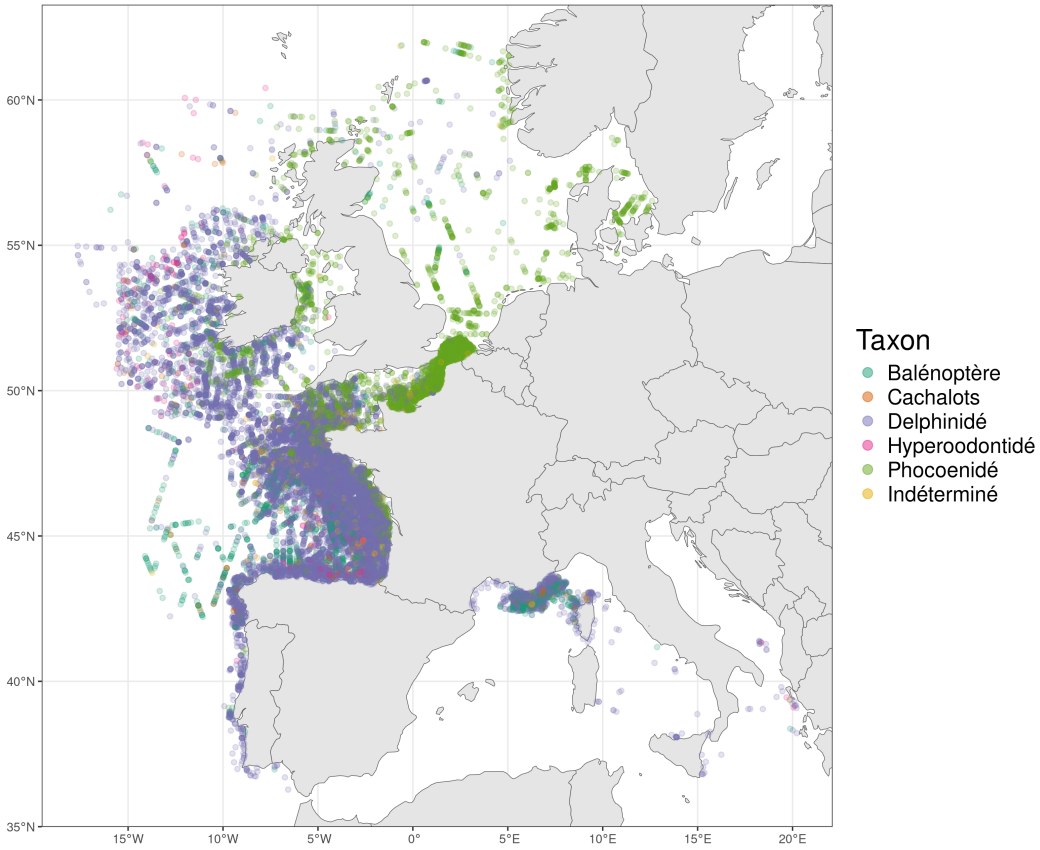


FIGURE 5 – Carte de l'ensemble des observations issues des 62 jeux de données protocolées rassemblés dans le cadre du livrable UO 1.1.4. Les différentes couleurs correspondent aux différents groupes taxonomiques observés.

— 258312 observations proviennent des jeux de données de présence seulement (voir Figure 7).

En terme d'unités d'effort, la base de données est constituée de :

— 117670 segments de transect mesurant entre 1.2 m et 186 km.
— 3045301 surface d'images mesurant entre 0.0002 et 1.6 km^2.

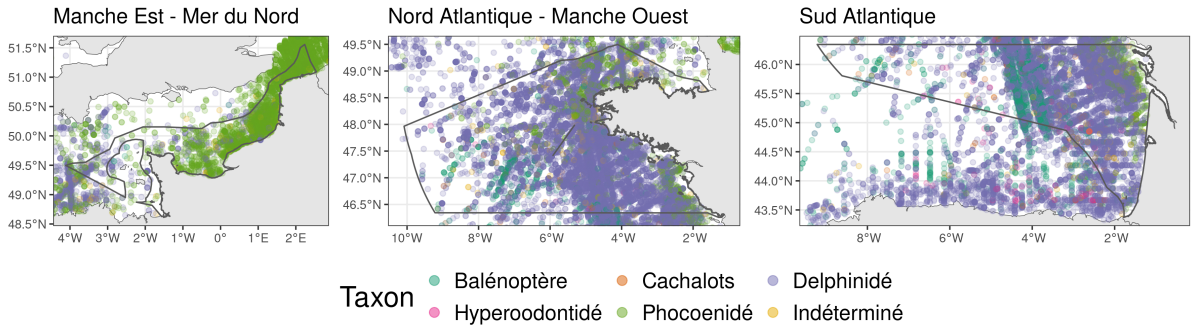


FIGURE 6 – Cartes à l'échelle des façades maritime des observations issues des 62 jeux de données protocolées rassemblés dans le cadre du livrable UO 1.1.4. Les différentes couleurs correspondent aux différents groupes taxonomiques observés.

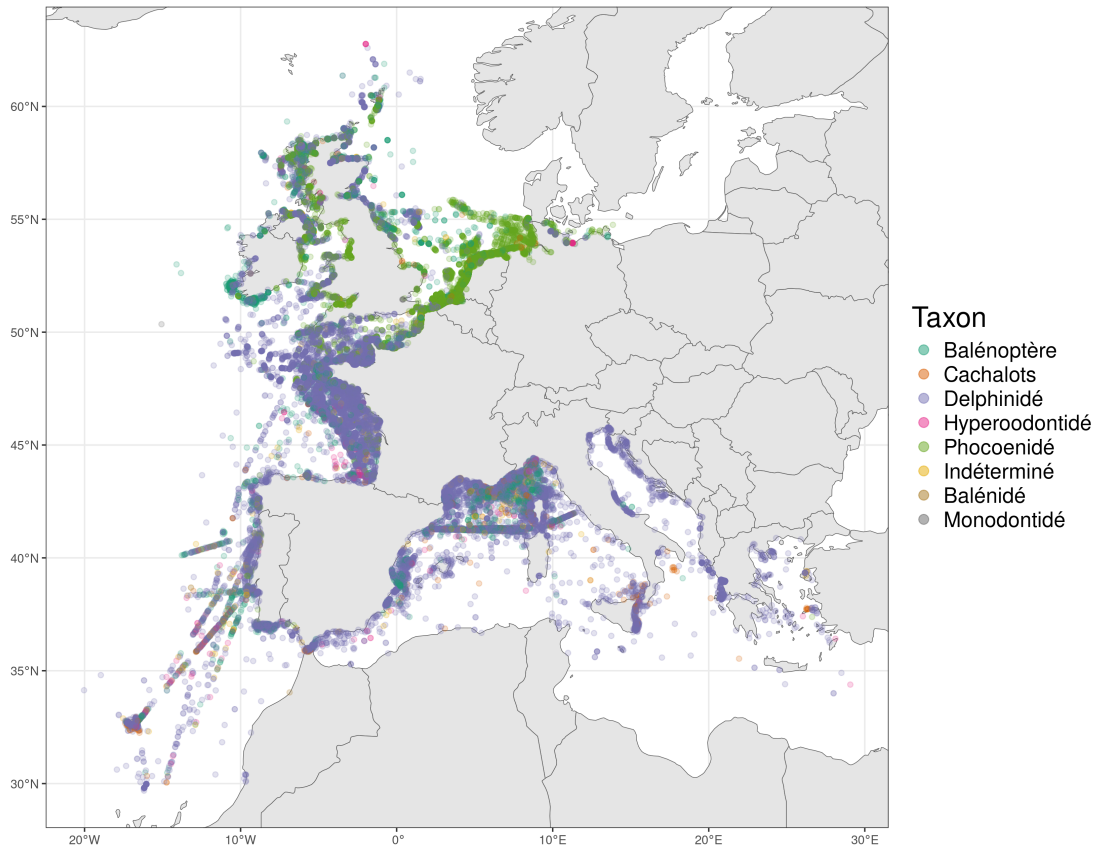


FIGURE 7 – Carte de l'ensemble des observations issues des 3 jeux de données de présence seulement rassemblés dans le cadre du livrable UO 1.1.4. Les différentes couleurs correspondent aux différents groupes taxonomiques observés.

Cette hétérogénéité de la segmentation nécessitera un travail complémentaire d'aggrégation pour les plus petits segments et surfaces et de re-segmentation pour les plus grands segments et surfaces qui sera entrepris par Pelagis avec l'aide de Floriane Plard et de Biotope.

Afin de représenter sur une même carte, les différentes mesures de l'effort (surface pour les données digitales et longueur de segment de transect parcouru pour les autres types de protocole), nous avons transformé la longueur de transect parcouru en surface échantillonnée. Pour cela, nous avons considéré que la largeur de la bande effectivement observée de part et d'autre du transect est de 200m.

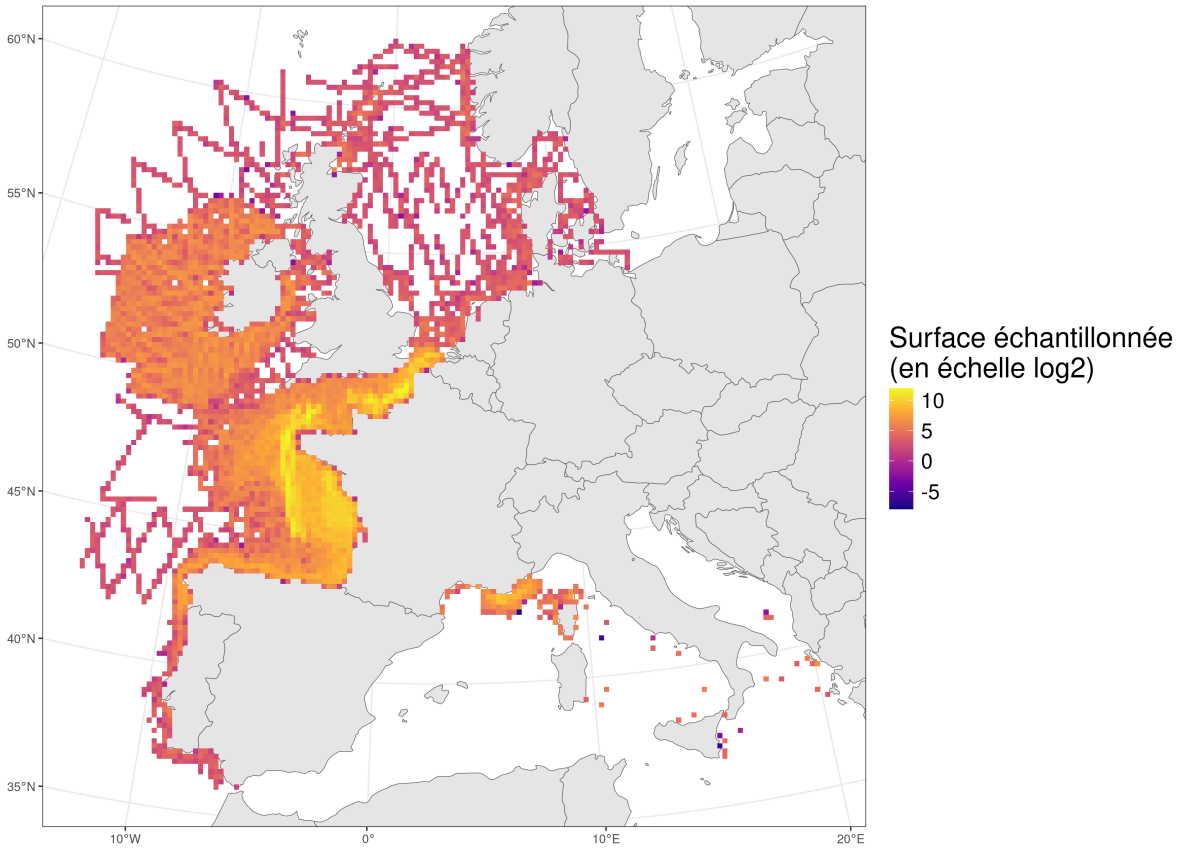


FIGURE 8 – Grille représentant la surface échantillonnée sur l'ensemble de la zone d'étude. Chaque carré de la grille représente une surface de 20 km^2. La couleur indique la surface échantillonnée en échelle log2. L'absence d'effort apparait en blanc.

La Figure 8 présente la surface échantillonnée par pixel de 20 km^2. La surface échantillonnée est calculée en km^2 et représentée en échelle log2 afin de mettre en évidence les différences relatives d'échantillonnage sur l'ensemble de la zone d'étude. Lorsque le pixel est de couleur jaune, la surface échantillonnée est de l'ordre de 1000 (10 en échelle log2) à 4000 km^2 (12 en échelle log2). L'absence d'échantillonnage apparaît en blanc. Les plus petites surfaces échantillonnées apparaissent en violet et ont des valeurs allant de 0.008 km^2 (-7 en échelle log2) à 0.032 km^2 (-5 en échelle log2).

Cette carte est également déclinée par grandes périodes de temps (voir Figure 9). Ces périodes sont constituées de 5 années hormis la dernière, la plus récente, qui s'étale sur 6 années mais inclue l'année 2024 qui n'a pu être prise en compte que partiellement car le travail de collecte de données s'est déroulée en 2024.

La répartition temporelle de l'effort d'échantillonnage est également représentée Figure 10

La Figure 11 permet d'avoir un aperçu de la diversité taxonomique de la base de données protocolées.

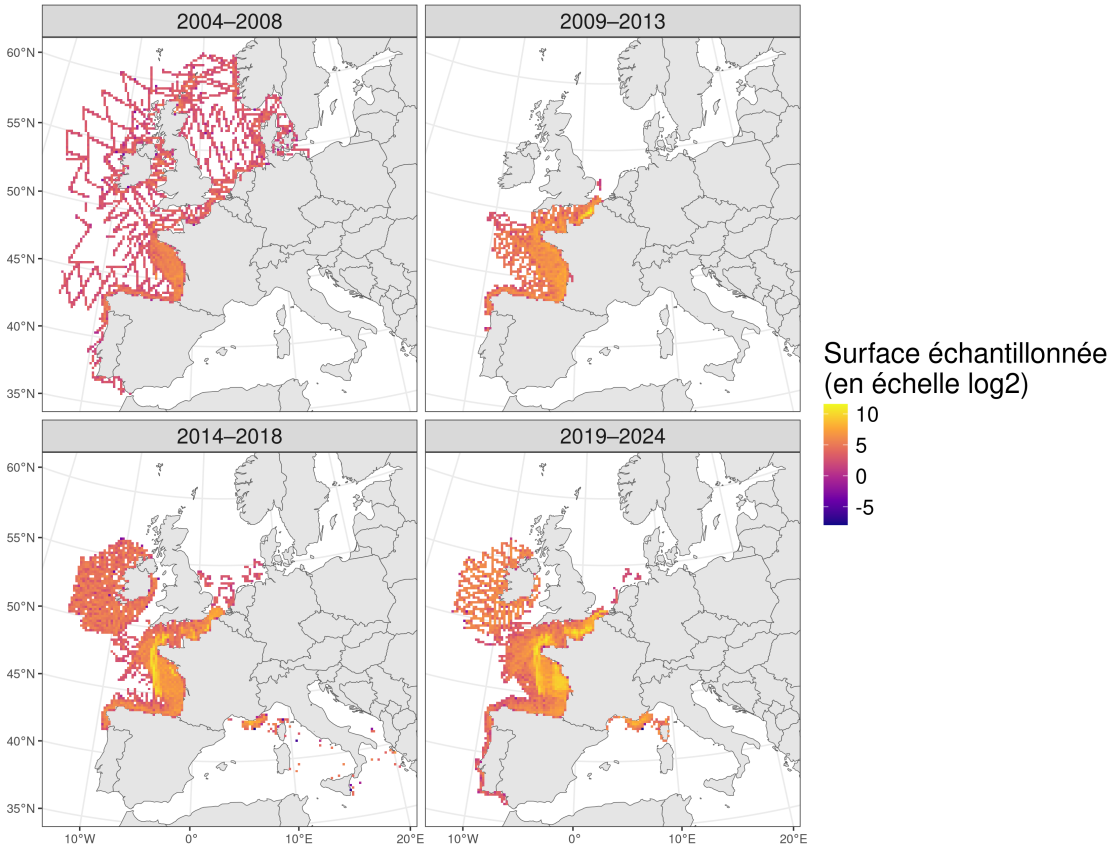


FIGURE 9 – Grille représentant la surface échantillonnée sur l'ensemble de la zone d'étude pour différentes périodes. Chaque carré de la grille représente une surface de 20 km^2. La couleur indique la surface échantillonnée en échelle log2. L'absence d'effort apparait en blanc.

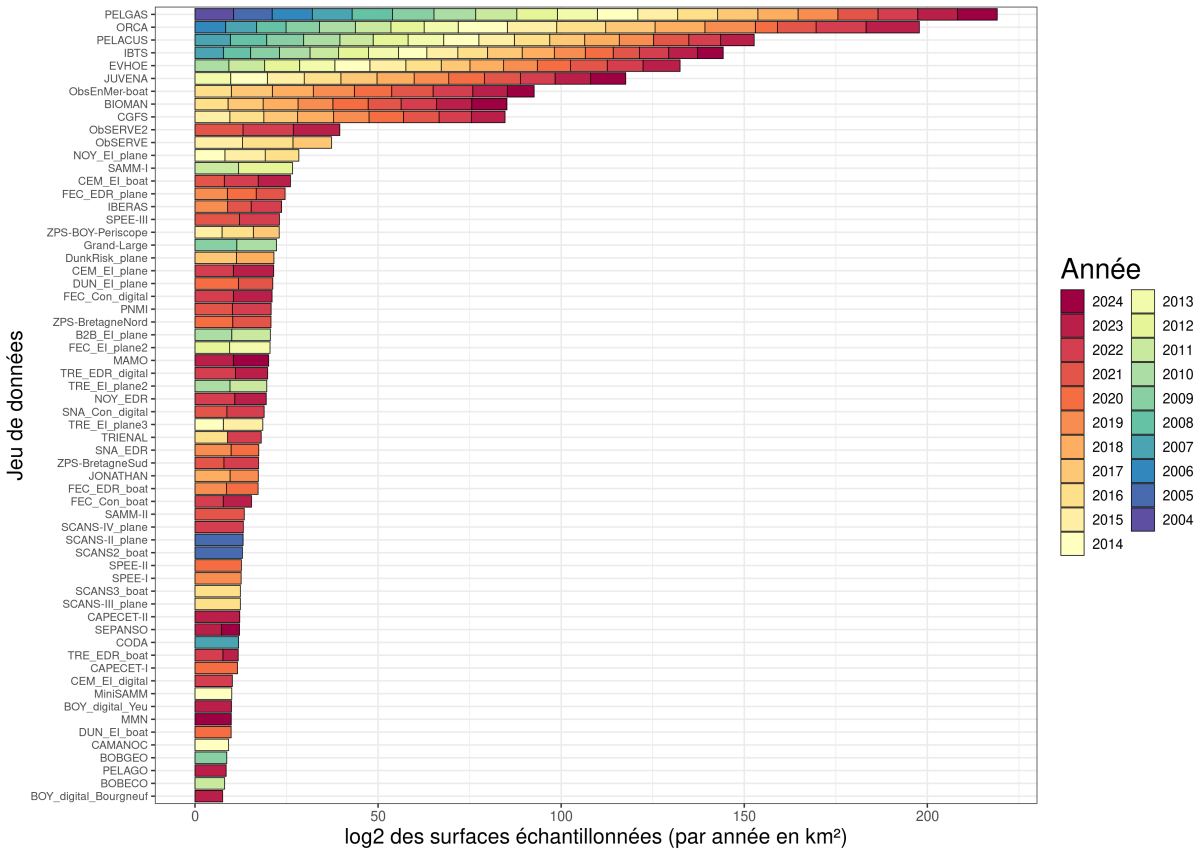


FIGURE 10 – Répartition temporelle de l'effort d'échantillonnage en surface échantillonnée (km²) par jeu de données et par année. Chaque ligne correspond à un jeu de données. Les jeux de données sont ordonnés selon la surface totale échantillonnée. Chaque ligne est constituée de plusieurs barres horizontales dont la longueur indique la surface échantillonnée en échelle log2 et les couleurs représentent la contribution de chaque année.

Le groupe taxonomique le plus diversifié est celui des Delphinidés avec 14 espèces pour les données de présence seulement et 11 espèces pour les données protocolées. Il s'agit du groupe taxonomique le plus représenté dans la base de données protocolées avec 22529 observations alors qu'il s'agit des Phocoenidés dans la base de données de présence seulement avec 183480 observations. Deux groupes taxonomiques sont présents dans la base de données de présence seulement alors qu'ils sont absents dans la base de données protocolées : les Balénidés (3 espèces et un total de 21 observations) et les Monodontidé (8 observations de Beluga).

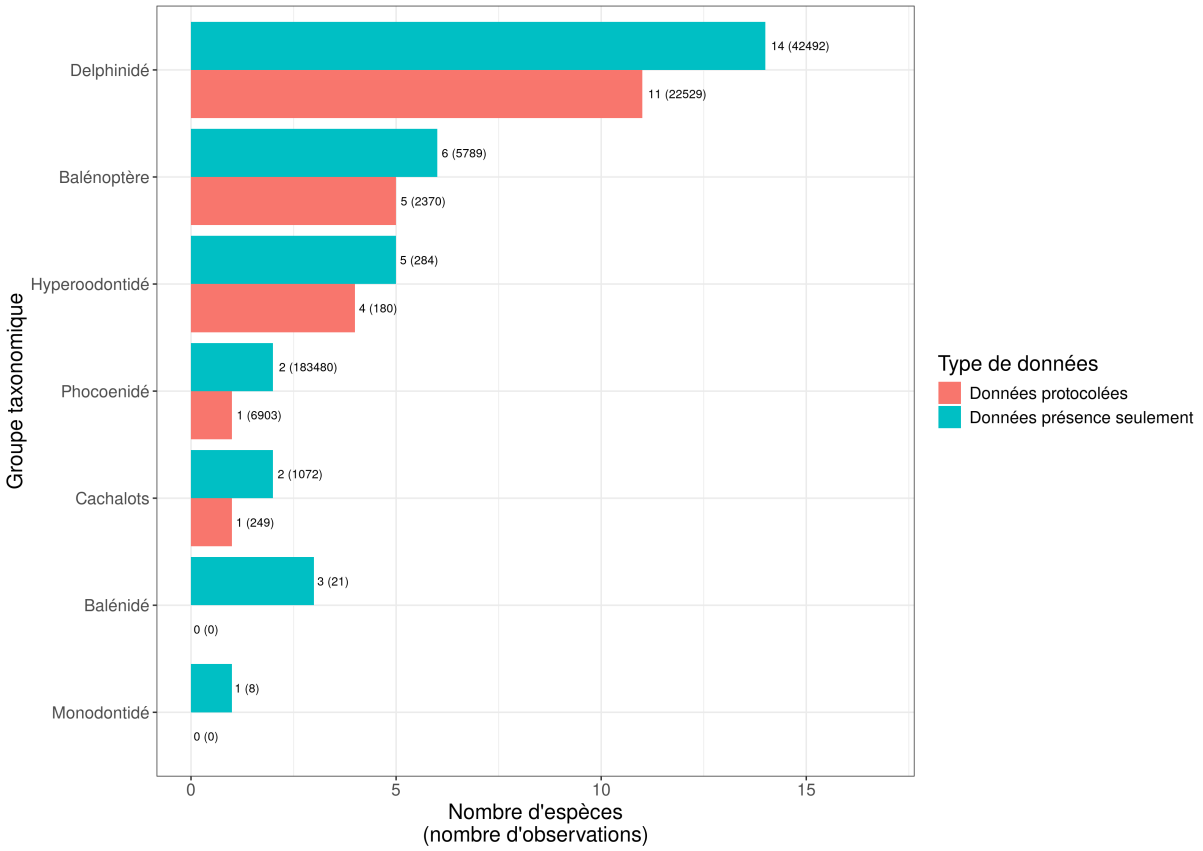


FIGURE 11 – Répartition du nombre d'espèces et du nombre d'observations par groupe taxonomique et pour chaque type de données (protocolées et présence seulement). Le nombre d'espèces est indiqué en chiffre devant chaque barre et le nombre d'observations est indiqué entre parenthèses. Les Pinnipèdes et les espèces indéterminées ont été écartés de cette analyse.

8 Discussion et perspectives

8.1 Difficultés rencontrées

Sur les 108 jeux de données protocolées inventoriées (Table 2), 46 jeux de données n'ont pas pu à ce jour être intégrés à la base de données finale. Dans cette partie nous revenons sur les principales difficultés rencontrées.

8.1.1 Accès aux données et autorisations d'utilisation

Lors de la phase de collecte, nous avons parfois été confrontés à des difficultés pou accéder aux données ou pour obtenir les autorisations nécessaires à leur utilisation dans le projet MODELCET ATL.

— L'implication de nombreuses structures – fournisseurs et propriétaires des données, partenaires,

sous-traitants – conduit à des échanges plus longs. Ceci a pu causer des retards, même dans certains cas où les données étaient d'origine publique (`AO7`).
— Données accessibles uniquement via l'élaboration d'une conventions de données (`STB`, `Grand_Large`), nécessitant des démarches juridiques ou administratives loin des compétences de notre équipe (et potentiellement aussi de nos interlocuteurs).
— Non-réponse ou absence de livraison. Certains contacts identifiés lors de l'inventaire des données, n'ont pas répondu à nos sollicitations malgré des relances. L'interprétation des non réponses est nécessairement floue : peut-être que la demande n'était pas pertinente car la personne contacté n'était pas la bonne ou bien car les données n'existent pas, manque de temps, pas de volonté de partager des données…

8.1.2 Contraintes de ressources limitées

Le processus d'harmonisation étant un travail complexe, sa mise en œuvre s'est heurtée à une limitation en temps. C'est pourquoi nous avons réalisé un travail de hiérarchisation (Section 3).

C'est également un effort de longue durée, ne serait-ce que parce qu'il faut collecter les données, ce qui implique de basculer sur d'autres projets. Dans ce contexte, le suivi des avancées, l'actualisation des informations (emails envoyés, contacts, protocoles, années et autres métadonnées) pour les plus de 100 sources en parallèle constitue un défit organisationnel.

8.2 Limites de l'harmonisation

8.2.1 Hétérogénéité des protocoles

Les jeux de données intègrent des méthodes très différentes (distance sampling nautique ou aérien, strip transect visuel ou digital, observations opportunistes). A chaque méthode et selon les différentes sources il existe des différences dans la manière de quantifier l'effort.

Le distance sampling permet d'estimer les largeurs de bande effectives (ESW - *efficient strip width*) pour que les différences dans les probabilités de détection puissent être prises en compte ensuite dans les modèles, mais la prise en compte des conditions d'observation selon la source peut-être problématique car les informations disponibles ou leur encodage est variable (par exemple les variables `SUBJECTIVE` ou `SEASTATE`, voir Section 5.2.1).

Dans certains protocoles, l'effort est connu au sens où l'on sait où, quand et comment a eu lieu l'échantillonnage, mais probabilité de détection n'est pas estimable à partir des données. C'est le cas du strip-transect qui est utilisé par de nombreuses sources. La modélisation de l'hétérogénéité de la probabilité de détection doit alors s'appuyer sur des hypothèses.

8.2.2 Hétérogénéité des échelles spatiales

Une autre limite importante est l'hétérogénéité des échelles spatiales des unités d'effort. Celles-ci varient fortement selon les protocoles et les moyens d'observation. Par exemple, les suivis digitaux réalisés à partir d'images aériennes produisent des unités d'effort très fines (une donnée d'effort par seconde), tandis que dans certains suivis réalisés par avion ou bateau une unité d'effort peut correspondre à un transect complet de plusieurs dizaines de kilomètres.

Nous avons fait le choix de garder la résolution la plus fine qui était disponible dans les données brutes. Cependant, la création d'une base de données enrichies avec les variables environnementales et prêtes pour la modélisation impliquera certainement de retravailler les données autour d'une résolution commune. Cette résolution doit être cohérente avec celle des données environnementales retenues, avec l'échelle des processus physiques et biologiques pertinents, ainsi qu'avec la résolution spatiale attendue pour les prédictions des modèles.

8.2.3 Qualité inégale des données

Dans certains cas, les données anciennes notamment, il nous a été impossible de reconstituer au mieux l'effort et les conditions d'observation associées. La raison principale était des données incomplètes ou organisées de façon sous-optimale, probablement suite à de mauvaises manipulations ou à des changements dans les protocoles, les logiciels de collecte ou le lieu de stockage des données.

Les données manquantes étaient systématiquement des données d'effort. Les situations que nous avons rencontrées couramment :

— Données où l'information sur l'effort et les conditions d'observation est conditionnée à la présence d'au moins une observation. Cela se produit quand les données d'effort ont été fusionnées avec les observations mais, durant ce processus, l'unité d'échantillonnage retenue est l'observation (et non le segment d'effort) si bien que les segments d'effort sans observation sont ignorés.
— Données où l'information sur l'effort est disponible avec une résolution spatiale grossière. Le manque d'information précise sur l'heure exacte des observations ou la présence des traces GPS ne permet pas de reconstituer facilement des données à fine échelle.

Une seconde source d'hétérogénéité dans la qualité des données est la disponibilité des métadonnées. Certaines sources de données sont très bien documentées (rapports de mission, publications scientifiques, référencement dans des catalogues de données), tandis que d'autres ne disposent d'aucune information publiquement disponible. Les métadonnées sont indispensables 1° pour hiérarchiser l'importance des sources de données et 2° pour créer une base de données de qualité. Dans ce contexte, et compte tenu des ressources limitées, les sources les mieux documentées sont avantagées.

8.2.4 Perte d'information

L'effort d'harmonisation des données s'accompagne d'une simplification. En effet, certaines variables spécifiques à un seul ou à quelques jeux de données ont été écartées. Une partie de la richesse initiale des données est donc sacrifiée pour que les données soient plus compréhensibles et plus exploitables. Notons toutefois que l'intégralité des informations est conservé dans les données brutes (versionnées avec DVC). Elles ne sont donc pas perdues pour le projet dans son ensemble et il est possible de venir récupérer toute information utile en cas de besoin.

8.3 Perspectives

La base de données constituée représente un socle robuste pour la suite du projet, néanmoins quelques tâches doivent encore être réalisée pour parachever la base de données en vue de la modélisation. Les prochaines étapes viseront à :

— associer à chaque segment d'effort des variables environnementales statiques et dynamiques,
— finaliser l'harmonisation de l'encodage des variables qualitatives (ex. conditions d'observation) et de la longeur des segments d'effort,
— valider l'arbre taxonomique,
— et, si nécessaire, enrichir la base avec des jeux de données complémentaires dès qu'ils seront techniquement disponibles.

9 Conclusion

Ce travail a permis de constituer une base de données sur les mammifères marins des façades maritimes françaises de l'Atlantique et de la Manche rassemblant et harmonisant un large éventail de jeux de données. Cette base de données est inédite par sa taille. Au total, 108 sources de données protocolées et 13 sources de présence ont été inventoriées, dont 68 intégrées dans la base harmonisée.

Cela représente 39 000 observations biologiques issues de sources de données protocolées (297 000 en incluant les données de présence seulement), pour un d'effort d'échantillonnage de $1,00 \times 10^6$ km de transects (approximativement $12,7 \times 10^3$ km^2 de couverture), couvrant la période 2004–2024, avec une forte densité d'observations récentes. Ces chiffres témoignent de la diversité des données compilées et du potentiel de la base de données en termes de modélisation.

Guidé par objectifs de MODELCET ATL et les contraintes de la modélisation, le processus de construction de la base de données s'est appuyé sur une démarche rigoureuse, transparente et reproductible, incluant inventaire, collecte, hiérarchisation et harmonisation des données (voir Figure 12), et en utilisant des outils libres, gratuits, adaptés et durables tels que R, git et DVC.

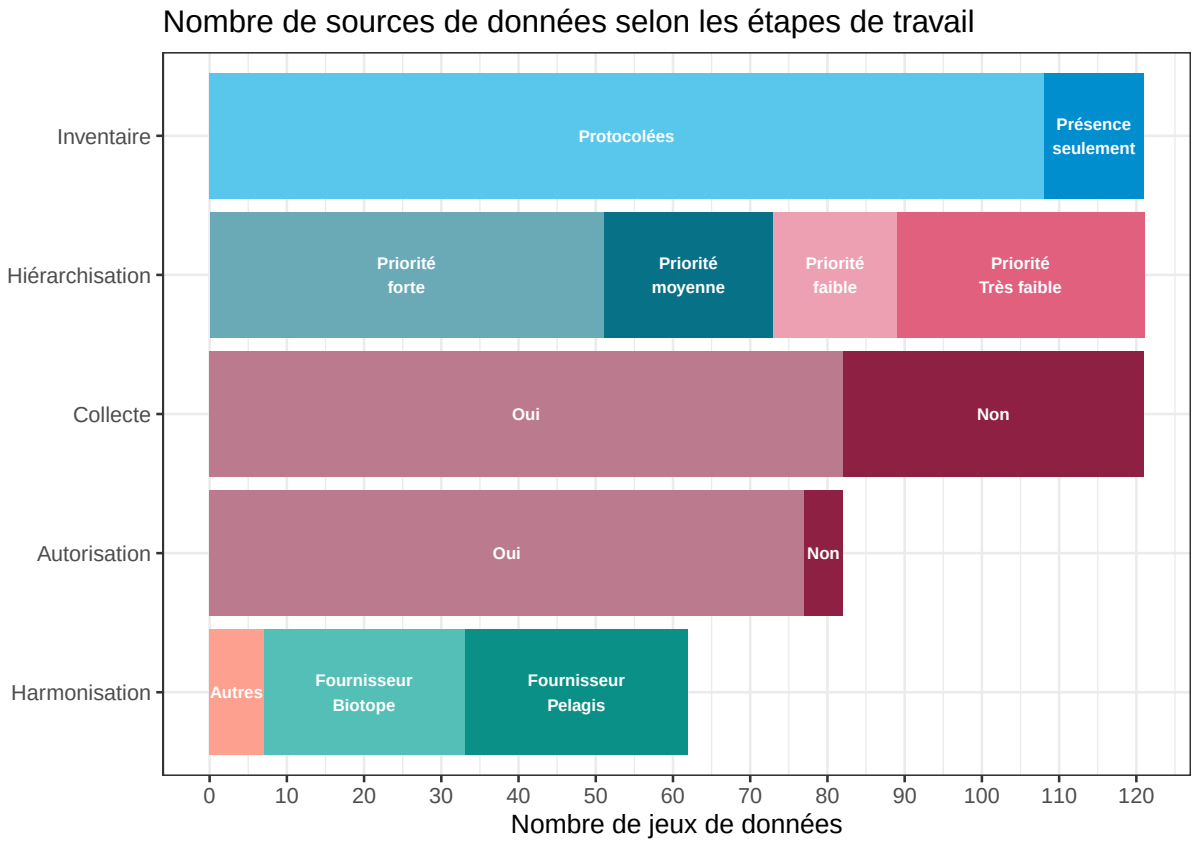


FIGURE 12 – Figure récaptitulative du nombre et du statut des sources de données aux différentes étapes de construction de la base de données MODELCET ATL.

Enfin, nous souhaitons conclure ce rapport en exprimant notre gratitude à l'égard de tous les fournisseurs et propriétaires de données pour avoir contribué à ces efforts en répondant à nos sollicitations et en accordant leur permission d'utiliser leurs données. Nous soulignons que cette base de données n'est pas seulement un outil au service du projet MODELCET ATL mais, avec le concours et la bienveillance des propriétaires des données et autres parties prenantes, constitue une ressource réutilisable, susceptible d'alimenter de nombreux projets à venir.

Références

Barve, Vijay, et Edmund Hart. 2022. *rinat : Access 'iNaturalist' Data Through APIs*. /url%7Bhttps://docs.ropensci.org/rinat/%7D.

Chamberlain, Scott, Vijay Barve, Dan Mcglinn, Diego Oldoni, Peter Desmet, Lisa Geffert, et Karthik Ram. 2025. « rgbif : Interface to the Global Biodiversity Information Facility API ». https://CRAN.R-project.org/package=rgbif.

Dorémus, Ghislain, et Olivier Van Canneyt. 2025. « Protocole d'observation de la mégafaune marine depuis les campagnes océanographiques – Programme MEGASCOPE ». https://www.observatoire-pelagis.cnrs.fr/wp-content/uploads/2025/03/PROTOCOLE_OBSERVATION_Megascope_2025.pdf.

GECC. 2025. « OBSenMER : plateforme collaborative pour la collecte et l'analyse des observations en mer ». https://www.obsenmer.org/.

Observatoire Pelagis. 2021. « PelaObs : Visual sightings of scientific surveys ». https://pelabox.univ-lr.fr/pelagis/PelaObs/.

———. 2025. « SAMMOA : logiciel d'observation aérienne de la mégafaune marine ». https://www.observatoire-pelagis.cnrs.fr/les-outils/sammoa/.

Owens, Hannah, Vijay Barve, et Scott Chamberlain. 2024. *spocc : Interface to Species Occurrence Data Sources*. /url%7Bhttps://github.com/ropensci/spocc%7D.

Plard, Floriane, Hélder Araújo, Amaia Astarloa, Maite Louzao, Camilo Saavedra, José Antonio Vazquez Bonales, Graham John Pierce, et Matthieu Authier. 2024. « Using fusion effects to decrease uncertainty in distance sampling models when collating data from different surveys ». *Marine Mammal Science* 40 (3) : e13104. https://doi.org/10.1111/mms.13104.

Savouré-Soubelet, Audrey, Stéphane Aulagnier, Patrick Haffner, François Moutou, Olivier Van Canneyt, J-B Charrassin, et Vincent Ridoux. 2016. *Atlas des mammifères sauvages de France : Volume 1. Mammifères marins*.

10 Annexes

10.1 Nomenclature des jeux de données de suivis pour les parcs éoliens offshore

10.1.1 Abréviation des noms de parc éolien

Abréviation du nom du parc	Nom du parc
AO5	Parc éolien flottant 'Bretagne Sud'
AO7	Parc éolien 'Sud-Atlantique' (Oléron)
CEM	Parc éolien 'Centre Manche' (AO4)
CEM2	Parc éolien 'Centre Manche 2' (AO8)
COUR	Parc éolien en mer de 'Courseulles' (AO1)
DUN	Parc éolien en mer de 'Dunkerque' (AO3)
FEC	Parc éolien en mer de 'Fécamp' (AO1)
NOY	Parc éolien en mer de 'Noirmoutiers' (AO2)
PIL	Parc éolien pilote flottant de 'La Rochelle et Belle-île'
SNA	Parc éolien en mer de 'Saint-Nazaire' (AO1)
STB	Parc éolien en mer de 'Saint-Brieuc' (AO1) ou 'Bretagne Nord'
TRE	Parc éolien en mer de 'Dieppe - Le Tréport' (AO2)

10.1.2 Abréviation de la phases de projet

Abréviation de la phase de projet	Phase de projet
EI	État initial
EDR	État de référence
Con	Phase de construction

10.1.3 Abréviation de la méthode d'observation

Abréviation de la méthode d'observation	Méthode d'observation
plane	Suivi aérien visuel

Abréviation de la méthode d'observation	Méthode d'observation
digital	Suivi aérien digital (haute altitude)
boat	Suivi nautique

10.2 Extraits de codes

Listing 1 `data/cetaceans/metadata/datasets/DUN_EI_plane.yaml` Exemple de fichier YAML de métadonnées, ici pour le jeu de données 'DUN_EI_plane'. Le début di fichier à trait aux informations générales utiles au plan de gestion des données et au référencement du jeux de données dans un catalogue.

```yaml
source:
  owner:
    name: EDF Renouvelables
    contact: Caroline PIQUET
    email: Caroline.Piguet@edf-re.fr
  provider:
    name: Biotope
    contact: Arnaud GOVAERE
  performing:
    name: Biotope
    contact: Pauline DE ROCK

license: MODELCET-restricted

citation: |
  Biotope. (2021). Parc éolien en mer de Dunkerque -
  Suivi de l'avifaune et de la mégafaune marine - Etat initial.

title: Etat initial du parc éolien en mer de Dunkerque

name: DUN_EI_plane

description: ""

url:
  - <https://www.eoliennesenmer.fr/facades-maritimes-en-france/facade-manche-mer-du-nord/dunkerque>
  - <https://parc-eolien-en-mer-de-dunkerque.fr/wp-content/uploads/2022/04/Ateliers-environnement_1-Etat-initial_Presentation_MM-chiropteres_VDEF-CORPUS-min.pdf#page=13>

survey_type: Aerial strip transect

categories:
  - "Façade: Manche"
  - "Theme: OWF"
  - "Type: Protocolled"
  - "Protocol: Plane"
  - "Source: Biotope"

image: https://data.ofb.fr/catalogue/srv/api/regions/geom.png?geomsrs=EPSG:4326&geom=POLYGON ((1.250836 50.71884, 3.329025 50.71884, 3.329025 51.86271, 1.250836 51.86271, 1.250836 50.71884))
```

Listing 2 `data/cetaceans/metadata/datasets/DUN_EI_plane.yaml` Exemple de fichier YAML de métadonnées, ici pour le jeu de données 'DUN_EI_plane'. La seconde partie du fichier a trait aux données techniques concernant l'échantillonnage : détails du protocole, des dates et de la zone géographique échantillonnées.

```yaml
protocols:
  - name: Aerial strip transect
    params:
      n_observer:
        val: 2
        unit: ~
      strip_width:
        val: 200
        unit: m
      plane_height:
        val: 180
        unit: m
      plane_speed:
        val: 90
        unit: knot
      plane: Britten-Norman BN-2 Islander (with Bubble-windows)

sampling_design:
  - protocol: Aerial strip transect
    space:
      epsg: 4326
      bbox:
        xmin: 1.250836
        ymin: 50.71884
        xmax: 3.329025
        ymax: 51.86271
    time:
    - 2020-03-04
    - 2020-03-07
    - 2020-05-25
    - 2020-05-26
    - 2020-05-27
    - 2020-07-13
    - 2020-07-14
    - 2020-07-15
    - 2020-09-21
    - 2020-09-22
    - 2020-10-17
    - 2020-10-19
    - 2020-11-05
    - 2020-11-06
    - 2020-11-07
    - 2021-02-04
    - 2021-02-05
    - 2021-02-06
```

Listing 3 `data/cetaceans/metadata/data_specs/biotope_aerial_2020.yaml` Exemple de fichier YAML de spécification du format des données, ici celui utilisé pour le jeu de données 'DUN_EI_plane'

```yaml
effort:
  read_fun: function(file) sf::st_read(file, layer = "Plane effort")
  vars:
    SURVEY_ID: ~
    DATE: date
    TIME: ~
    TRANSECT_ID: transect
    LEG_ID: [date, transect]
    LINE_BUILDER: [date, transect]
    ELEVATION: ~
    GLARE_SEVERITY: ~
    SEASTATE_BEAUFORT: SEASTATE_BEAUFORT
    SEASTATE_BEAUFORT_LOWER: SEASTATE_BEAUFORT_LOWER
    SEASTATE_BEAUFORT_UPPER: SEASTATE_BEAUFORT_UPPER
    SEASTATE_PETERSEN: ~
    SEASTATE_DOUGLAS: ~
    TURBIDITY: ~
    CLARITY: ~
    SUBJECTIVE: VISIBILITY
    EFFORT_LENGTH: effort_length
    EFFORT_AREA: ~
    LON: lon
    LAT: lat
    WKT: wkt_transect
    WARNING: WARNING
    WARNING_DETAIL: WARNING_DETAIL
    COMMENT: COMMENT
  format:
    TIME: NA_POSIXct_

obs:
  read_fun: function(file) sf::st_read(file, layer = "Plane sightings")
  vars:
    SURVEY_ID: survey
    DATE: date
    TIME: ~
    TRANSECT_ID: transect
    LEG_ID: [date, transect]
    TAX_ORIGIN: species
    GROUP_SIZE: group_size
    PERP_DIST: distance
    LON: lon
    LAT: lat
  format:
    TIME: NA_POSIXct_
```