



HAL
open science

Dépôt légal de données brutes de biodiversité – Caractérisation et évaluation des données sur les espèces en vue de leur diffusion dans le SINP

M Laville, L Poncet, J Touroult, P Dupont, J Panijel, F Vest, S Robert

► To cite this version:

M Laville, L Poncet, J Touroult, P Dupont, J Panijel, et al.. Dépôt légal de données brutes de biodiversité – Caractérisation et évaluation des données sur les espèces en vue de leur diffusion dans le SINP. PatriNat (AFB-CNRS-MNHN). 2019, pp.26. mnhn-04256194

HAL Id: mnhn-04256194

<https://mnhn.hal.science/mnhn-04256194>

Submitted on 24 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Dépôt légal de données brutes de biodiversité

Caractérisation et évaluation des
données sur les espèces en vue de
leur diffusion dans le SINP.

Décembre 2019



SINP
Système d'Information
sur la Nature et le Paysage



Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE LA TRANSITION
ÉCOLOGIQUE
ET SOLIDAIRE

Sommaire

Table des matières

I.	Contexte.....	2
II.	Analyse des dépôts.....	4
III.	Analyses complémentaires avant diffusion des données dans le SINP.....	19
IV.	Pistes d'améliorations.....	23
V.	Conclusion.....	25

Contexte

Depuis la promulgation de la loi Biodiversité du 8 août 2016, toute personne morale ou physique porteuse d'un projet d'aménagement ou d'un document de planification conduisant au recueil de données de biodiversité doit verser ces données afin qu'elles puissent être valorisées, mises à disposition du public et contribuer à l'inventaire du patrimoine naturel (IPN) (articles L411-1 et L122-1 du code de l'environnement). Ces articles concernent les études d'évaluations préalables ou les suivis d'impacts réalisés dans le cadre de documents de planification ou de projets d'aménagements soumis à l'approbation de l'autorité administrative.

Afin de permettre le dépôt des données brutes de biodiversité (DEPOBIO) et la description des métadonnées associées, un système spécifique a été mis en service et ouvert le 1^{er} juin 2018.

Nous présentons ici les résultats d'une analyse des données et métadonnées versées au cours des six premiers mois de fonctionnement du dispositif (entre le 1^{er} et le 25 janvier 2019), notamment en vue d'apporter un retour d'expérience sur celui-ci et pour organiser la diffusion des données dans le système d'information sur la nature et les paysages (SINP). Ces données portent exclusivement sur des données d'observations et de suivis d'espèces (le module permettant le versement de données d'habitats n'ayant pas encore été implémenté).

Cette étude vise à répondre aux questions suivantes :

- Que contiennent les données déposées dans DEPOBIO ?
Cela consiste donc à dresser un état des lieux sur la quantité, la précision et la typologie des données et métadonnées déposées. Il s'agit également d'identifier quelles sont les espèces, les acteurs et les types d'études représentés.
- Les données transmises sont-elles de qualité suffisantes pour intégrer le SINP ?
Un bilan est établi sur la qualité des données et métadonnées déposées, notamment au regard des exigences du standard de données d'occurrences de taxons pour le SINP et des contrôles de cohérence et de validation scientifique automatique mis en place au niveau de l'INPN¹.

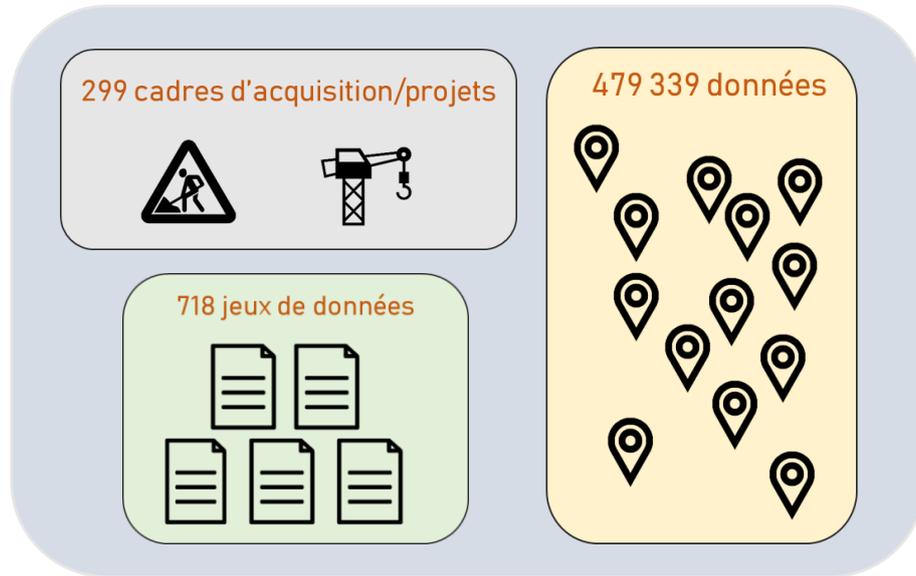
¹ <https://inpn.mnhn.fr/programme/donnees-observations-especes/references/validation>

- Quels sont les points d'améliorations possibles pour le dispositif ? Quelles mesures pourraient permettre une montée en qualité des données déposées ?

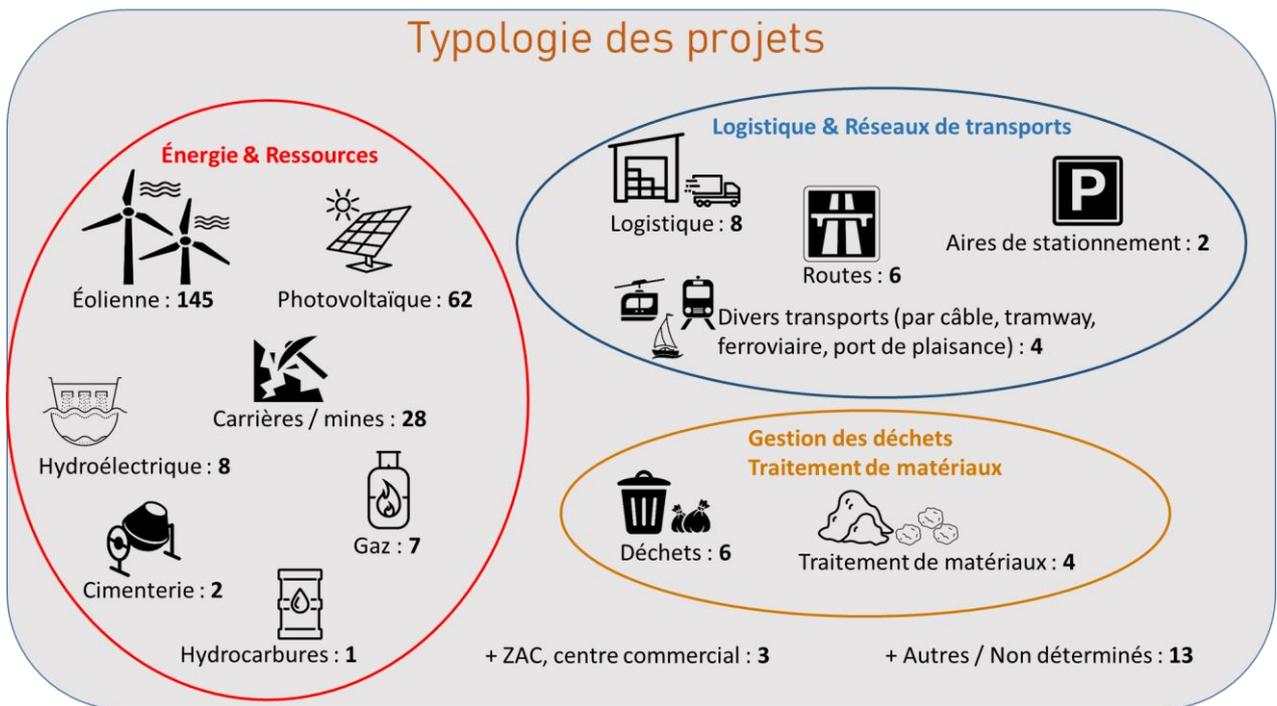


La présente analyse se situe schématiquement à l'interface entre le module dépôt des données GINCO et le SINP (flèche rouge du schéma).

Analyse des dépôts



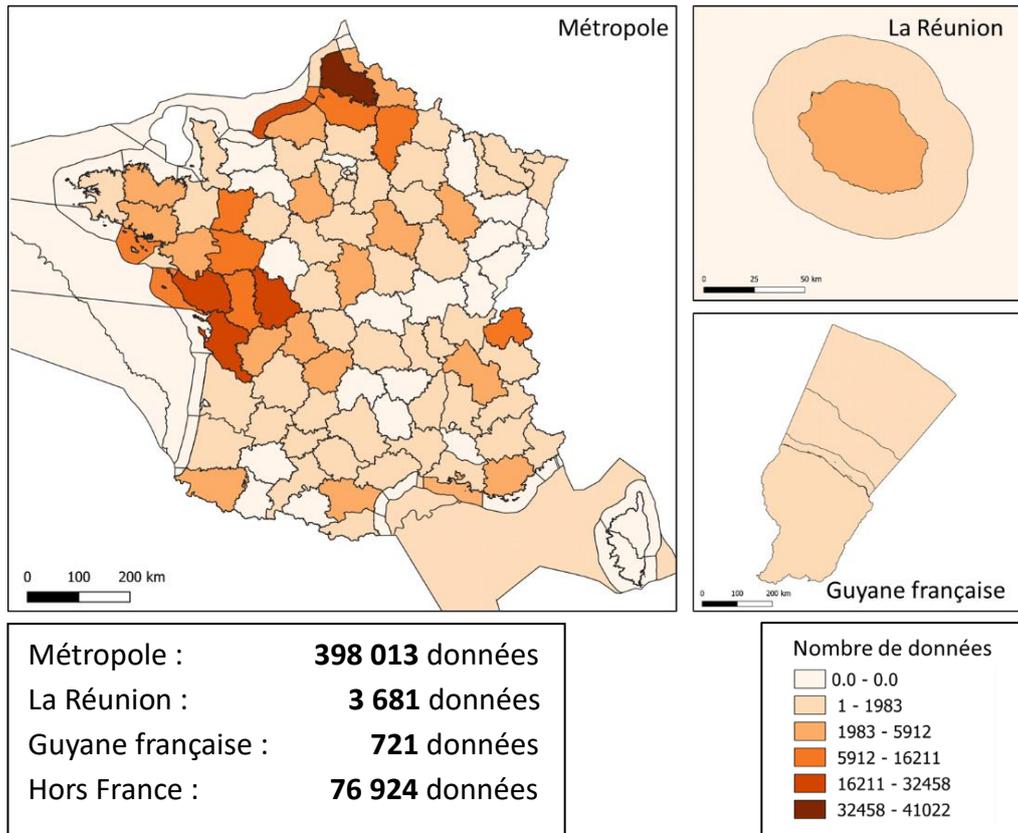
Typologie des projets



Typologie des projets : Le nombre de projets par catégories a été établi à l'aide des informations contenues dans les métadonnées des cadres d'acquisitions.

48 % des projets pour lesquels des données ont été déposées concernent le secteur éolien et 21 % le photovoltaïque.

Répartition spatiale des données



83 % des données déposées concernent la Métropole.

16 % des données déposées sont situées hors du territoire national.

19 % des données sont situées en milieu marin.

81 % des données sont situées en milieu terrestre.

Précision spatiale des données

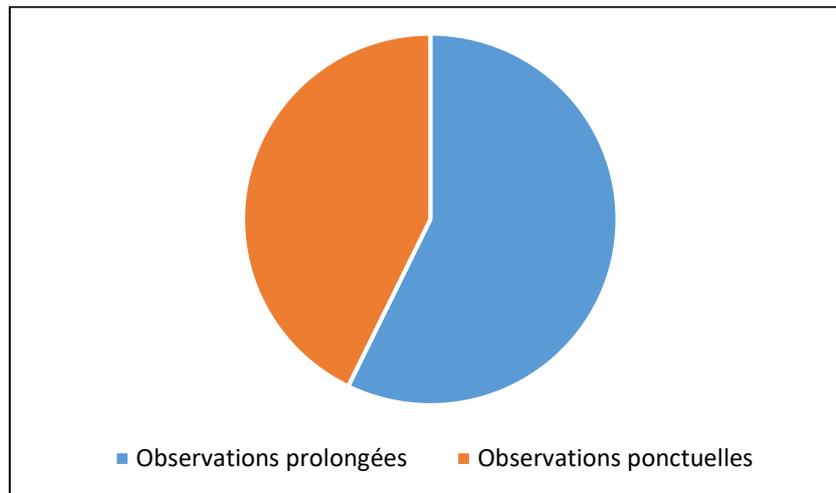
Données rattachées à une géométrie	Données rattachées à un référentiel
Point : 316 677 données (66 %)	Commune : 69 789 données (15 %)
Polygone : 65 673 données (14 %)	Maille 10 x 10 km : 1 789 données (< 1 %)
Ligne : 7 769 données (2 %)	Département : 1 376 données (< 1 %)
TOTAL = 390 119 données	TOTAL = 72 954 données

3% des données déposées ne disposent pas d'information géographique.

82 % des données déposées ont une information géographique correspondant à une géométrie (point, ligne ou polygone).

15 % des données déposées ont une information géographique se rattachant à un référentiel (commune, maille ou département).

Précision temporelle des données

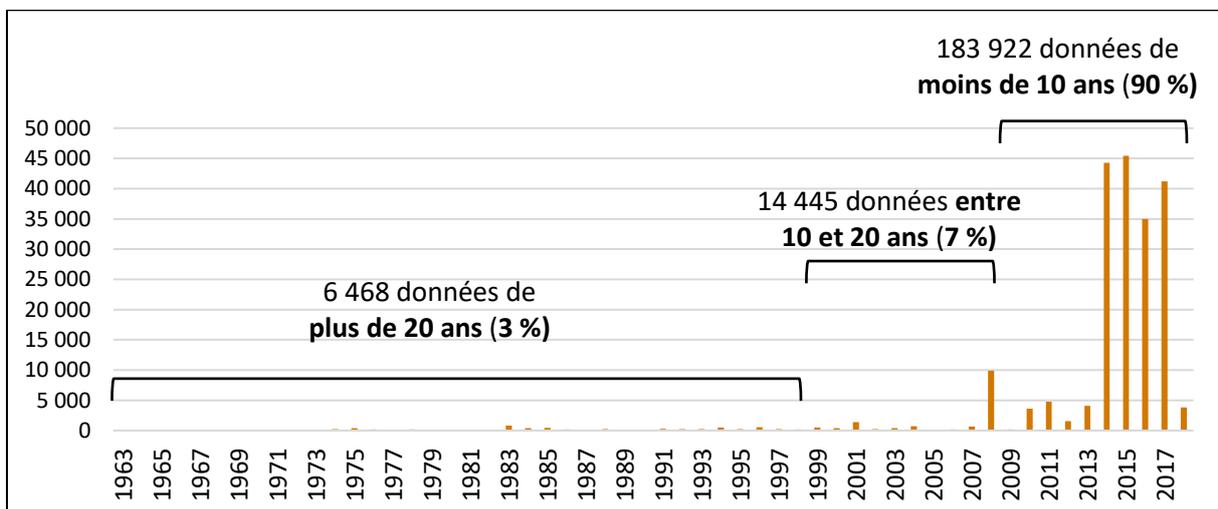


57 % des données correspondent à des observations sur une période (prolongées dans le temps).

43 % des données correspondent à des observations sur une journée (ponctuelles).

Répartition temporelle des observations

(pour les données ayant une précision temporelle ponctuelle).



90% des données ponctuelles déposées datent de moins de 10 ans.

83% des données ponctuelles déposées datent de moins de 5 ans.

Origine des données déposées

Statuts des données

Privée :	420 674 données (88 %)
Publique acquise :	29 411 données (6 %)
Publique :	17 997 données (4 %)
Indéterminé :	11 257 données (2 %)

Source des données

Terrain :	477 631 données (99 %)
Littérature :	25 données (< 1 %)
Collection :	2 données (< 1 %)
Indéterminé :	1 681 données (< 1 %)

88% des données déposées sont déclarées d'origine privée.

99,6% des données sont déclarées comme ayant été acquises directement sur le terrain.

Organismes cités

Liste des organismes les plus cités en tant que maîtres d'ouvrages

Organisme	Nombre de jeux de données
DIRECTION REGIONALE DE L'ENVIRONNEMENT DE L'AMENAGEMENT ET DU LOGEMENT AUVERGNE-RHONE-ALPES	37
RTE RESEAU DE TRANSPORT D ELECTRICITE	20
GLOBAL WIND POWER FRANCE	17
ENERGIETEAM SARL	16
FERME EOLIENNE DE FEUILLADE ET SOUFFRIGNAC	15
PARC EOLIEN DU CHATEAU	14
EOLIENNES EN MER DIEPPE LE TREPORT	14
SOC CONCASSAGE PREFABRICATION REUNION	14
ENERTRAG PAYS DE LOIRE I	14
EOLIENNES EN MER ILES D'YEU ET DE NOIRMOUTIER	13
RES	12
PARC EOLIEN D'ARGILLIERES	12
TERALTA GRANULAT BETON REUNION	11
ENERGIE DU CONFOLENTAIS	10
EDPR FRANCE HOLDING	10
W.E.B PARC EOLIEN DES VALLEES	10
FERME EOLIENNE DE BLANZAY	9
FERME EOLIENNE FLOTTANTE DE GROIX & BELLE-ILE	9
Réseau de Transport d'Electricité	9
PRD PERCIER REALIS ET DEVELOPPEMENT	8
...	...

243 noms d'organismes différents sont renseignés en tant que maîtres d'ouvrages.

Liste des organismes les plus cités en tant que collecteurs des données

Organisme	Nombre de données
Biotope	159 426
SARL EXEN	55 526
LPO VENDEE	26 754
CERA Environnement	16 010
TBM environnement	15 716
adev environnement	12 415
CBNA	11 115
Envol Environnement	10 329
PERISCOPE	7 794
BRETAGNE VIVANTE	7 414
cera environnement	6 826
CNRS Caen M2C	6 187
CSLN	5 878
ENCIS Environnement	5 117
ECOSPHERE	5 075
CPIE Loire Anjou	4 968
Picardie Nature	4 492
Calidris	3 979
FAUNA FLORA	3 688
CBNBL	3 371
...	...

196 noms d'organismes sont renseignés en tant que collecteurs.

Pour 6 % des données des mentions « Anonyme », « NSP » (Ne Sait Pas) ou « Inconnu » ont été renseignées pour désigner l'organisme collecteur.

678 noms d'observateurs différent sont mentionnés dans les données.

13,5 % des données portent toutefois les mentions « Anonyme » ou « Inconnu ».

Composition taxonomique des données déposées et validées

Les tableaux de synthèse présentés ci-dessous se basent sur les données déposées puis validées par des contrôles automatiques préalables à la diffusion de données dans le SINP (307 000 données). Ne sont ainsi pas pris en compte dans les synthèses les données non conformes, non cohérentes ou non valides.

Liste des rangs taxonomiques

Rang taxonomique	Nombre de données	Nombre de taxons
Espèce	280 318	5 189
Genre	14 844	325
Famille	8 081	24
Sous-Espèce	2 148	505
Ordre	908	6
Classe	370	2
Variété	341	97
Infra-Classe	28	1
Sous-Genre	27	2
Forme	17	10
Sous-Ordre	6	2
Race	3	3
Sous-Règne	3	1
Tribu	3	1
Agrégat	2	2
Super-Famille	1	1

Les données déposées et validées concernent à 92 % des rangs taxonomiques « espèce » ou infra espèces.

Liste des groupes taxonomiques

Groupe taxonomique	Précision groupe taxonomique	Nombre de données	Nombre de taxons	Pourcentages (total données)
Chordés	Mammifères	115 220	139	37,5 %
Chordés	Oiseaux	97 243	461	31,7 %
Plantes vasculaires	Angiospermes	55 763	2 912	18,2 %
Arthropodes	Insectes	10 351	1 054	3,4 %
Vers	Annélides	6 945	239	2,3 %
Arthropodes	Crustacés	4 700	222	1,5 %
Chordés	Poissons	3 923	105	1,3 %
Mollusques	Bivalves	3 019	90	1,0 %
Echinodermes	Autres	1 516	24	0,5 %
Mollusques	Gastéropodes	1 379	150	0,4 %
Chordés	Amphibiens	1 341	31	0,4 %
Bryophytes	Bryidae, Hépatiques et Anthocérotes	1 238	208	0,4 %
Chordés	Reptiles	1 229	42	0,4 %
Plantes vasculaires	Ptéridophytes	1 160	78	0,4 %
Arthropodes	Arachnides	810	168	0,3 %
Plantes vasculaires	Gymnospermes	396	38	0,1 %
Mollusques	Céphalopodes	258	8	0,1 %
Mollusques	Autres	214	10	0,1 %
Cnidaires	Hydrozoaires, Octocoralliaires et Autres	136	21	< 0,1 %
Siponcles	Autres	105	4	< 0,1 %
Bryozoaires	Autres	48	8	< 0,1 %
Arthropodes	Pycnogonides	24	4	< 0,1 %
Vers	Plathelminthes	17	2	< 0,1 %
Chordés	Ascidies	16	2	< 0,1 %
Algues	Ochrophyta	14	2	< 0,1 %
Arthropodes	Myriapodes	12	8	< 0,1 %
Autres	Chlorophytes, Charophytes et Autres	10	6	< 0,1 %
Porifères	Autres	8	2	< 0,1 %
Basidiomycètes	Autres	3	3	< 0,1 %
Algues	Rhodophytes	2	1	< 0,1 %

Les données concernent des espèces de faune vertébrés (71 %), de flore (19 %) et de faune invertébrés (10 %).

Liste des taxons les plus cités

Groupe taxonomique	Nom scientifique	Nombre de données	Pourcentages des données
Chiroptères	<i>Pipistrellus pipistrellus</i> (Schreber, 1774)	68 775	22,4 %
Chiroptères	<i>Pipistrellus kuhlii</i> (Kuhl, 1817)	10 776	3,5 %
Oiseaux	<i>Morus bassanus</i> (Linnaeus, 1758)	10 774	3,5 %
Chiroptères	<i>Myotis</i> Kaup, 1829	7 895	2,5 %
Chiroptères	<i>Pipistrellus nathusii</i> (Keyserling & Blasius, 1839)	6 499	2,1 %
Oiseaux	Alcidae Leach, 1820	5 391	1,8 %
Chiroptères	<i>Eptesicus serotinus</i> (Schreber, 1774)	4 619	1,5 %
Oiseaux	<i>Uria aalge</i> (Pontoppidan, 1763)	3 100	1,0 %
Oiseaux	<i>Fringilla coelebs</i> Linnaeus, 1758	2 956	1,0 %
Oiseaux	<i>Alauda arvensis</i> Linnaeus, 1758	2 925	1,0 %
Oiseaux	<i>Larus argentatus</i> Pontoppidan, 1763	2 775	0,9 %
Oiseaux	<i>Columba palumbus</i> Linnaeus, 1758	2 353	0,8 %
Chiroptères	<i>Barbastella barbastellus</i> (Schreber, 1774)	2 318	0,8 %
Oiseaux	<i>Corvus corone</i> Linnaeus, 1758	2 233	0,7 %
Chiroptères	<i>Plecotus</i> É. Geoffroy Saint-Hilaire, 1818	2 040	0,7 %
Oiseaux	<i>Turdus merula</i> Linnaeus, 1758	1 977	0,6 %
Oiseaux	<i>Buteo buteo</i> (Linnaeus, 1758)	1 779	0,6 %
Oiseaux	<i>Rissa tridactyla</i> (Linnaeus, 1758)	1 632	0,5 %
Oiseaux	<i>Larus</i> Linnaeus, 1758	1 437	0,5 %
Oiseaux	<i>Erithacus rubecula</i> (Linnaeus, 1758)	1 430	0,5 %
...

Parmi les taxons les plus cités présentés ci-dessus, 71 % des données se rapportent à des chiroptères et correspondent à cinq espèces et deux genres.

Répartition des données et des taxons par statut biogéographique²

Statut biogéographique	Nombre de données	Nombre de taxons
Présent	299 375	5 147
Statut inconnu	3 329	510
Introduit	2 667	222
Introduit envahissant	1 051	40
Introduit non établi	529	99
Occasionnel	118	12
Subendémique	23	7
Endémique	8	5

1,4 % des données concernent des espèces introduites

Répartition des données et des taxons par grand type d'habitat

Habitat	Nombre de données	Nombre de taxons
Terrestre	174 845	2 603
Continental (terrestre et/ou eau douce)	59 450	2 325
Marin	30 281	793
Marin et Terrestre	28 533	78
Non défini	13 303	148
Marin et Eau douce	360	13
Eau douce	323	81
Eau saumâtre	5	1

Les données concernent à 57 % des espèces purement terrestres, à 19 % des espèces continentales et à 10 % des espèces marines.

Nombre de données considérées comme sensibles selon le référentiel national du SINP³

Sensibilité	Nombre de données	Nombre de taxons
Données sensibles	4 426	77

4 426 données (1,5 % des données déposées et validées), portant sur 77 taxons distincts, sont considérées comme sensibles selon les critères utilisés dans le standard du SINP.

² Définitions des statuts biogéographiques : Gargominy, O., Terceirie, S., Régnier, C., Ramage, T., Dupont, P., Daszkiewicz, P. & Poncet, L. 2018. *TAXREF v12, référentiel taxonomique pour la France : méthodologie, mise en oeuvre et diffusion*. Muséum national d'Histoire naturelle, Paris. Rapport Patrinat 2018-117. 156 pp.

³ <https://inpn.mnhn.fr/programme/donnees-observations-especes/references/sensibilite>

Nombre de données déposées dont le taxons est concerné par un plan national d'action (PNA)

Statut de protection ou plan national d'action	Nombre de données	Nombre de taxons
Plan national d'action	82 950	37

27 % des données déposées et validées portent sur des espèces concernées par un plan national d'action.

Nombre de données dont le taxons a un statut de protection national⁴

Statut de protection	Nombre de données	Nombre de taxons
Protection nationale	130 304	489

42 % des données déposées et validées concernent des espèces faisant l'objet d'une protection nationale.

⁴ Liste des statuts de protection pris en compte :

- Liste des oiseaux représentés dans le département de la Guyane protégés sur l'ensemble du territoire et les modalités de leur protection : Article 4
- Liste des tortues marines protégées sur le territoire national et modalités de leur protection : Article 1
- Liste des oiseaux représentés dans le département de la Guyane protégés sur l'ensemble du territoire et les modalités de leur protection : Article 2
- Liste des mammifères marins protégés sur le territoire national et les modalités de leur protection : Article 2
- Liste des insectes protégés sur l'ensemble du territoire et les modalités de leur protection : Article 3
- Liste des mollusques protégés sur l'ensemble du territoire français métropolitain : Article 2
- Liste des mammifères terrestres protégés sur l'ensemble du territoire français et les modalités de leur protection : Article 2
- Listes des amphibiens et des reptiles protégés sur l'ensemble du territoire français et les modalités de leur protection : Article 2
- Listes des amphibiens et des reptiles protégés sur l'ensemble du territoire français et les modalités de leur protection : Article 5
- Mesures de protection sur tout le territoire national des oiseaux représentés dans les Terres australes et antarctiques françaises : Article 1
- Liste des mollusques protégés sur l'ensemble du territoire français métropolitain : Article 4
- Listes des amphibiens et des reptiles protégés sur l'ensemble du territoire français et les modalités de leur protection : Article 3
- Liste des espèces de poissons protégées sur l'ensemble du territoire français national : Article 1
- Liste des oiseaux protégés sur l'ensemble du territoire et les modalités de leur protection : Article 4
- Liste des mammifères marins protégés sur le territoire national et les modalités de leur protection : Article 3
- Listes des amphibiens et des reptiles protégés sur l'ensemble du territoire français et les modalités de leur protection : Article 4
- Liste des espèces végétales protégées sur l'ensemble du territoire français métropolitain : Article 1
- Liste des insectes protégés sur l'ensemble du territoire et les modalités de leur protection : Article 2
- Liste des espèces végétales protégées sur l'ensemble du territoire français métropolitain : Article 2
- Liste des oiseaux protégés sur l'ensemble du territoire et les modalités de leur protection : Article 3
- Liste des oiseaux représentés dans le département de la Guyane protégés sur l'ensemble du territoire et les modalités de leur protection : Article 3
- Liste des mammifères marins protégés sur le territoire national et les modalités de leur protection : Article 5
- Liste des mollusques protégés sur l'ensemble du territoire français métropolitain : Article 3
- Listes des amphibiens et des reptiles protégés sur l'ensemble du territoire français et les modalités de leur protection : Article 6
- Protection des écrevisses autochtones sur le territoire français métropolitain : Article 1

Répartition des données et des taxons par catégories de menaces selon la liste rouge mondiale de l'UICN

Statut de menace	Nombre de données	Nombre de taxons
Préoccupation mineure (LC)	156 071	1 169
Quasi menacée (NT)	5 287	53
Vulnérable (VU)	2 520	34
En danger (EN)	388	9
En danger critique (CR)	278	5
Données insuffisantes (DD)	720	30

Parmi les données déposées et validées relatives aux taxons évalués sur la liste rouge mondiale de l'UICN, 0,4 % concernent des espèces considérées « en danger » ou « en danger critique » et 94 % des espèces considérées en « préoccupation mineure » .

Répartition des données et des taxons par catégories de menaces selon la liste rouge européenne de l'UICN

Statut de menace	Nombre de données	Nombre de taxons
Préoccupation mineure (LC)	158 275	1 405
Quasi menacée (NT)	10 186	58
Vulnérable (VU)	5 792	30
En danger (EN)	680	6
En danger critique (CR)	195	2
Données insuffisantes (DD)	751	41
Non applicable (NA)	12	7

Parmi les données déposées et validées relatives aux taxons évalués sur la liste rouge européenne de l'UICN, 0,5 % concernent des espèces considérées « en danger » ou « en danger critique » et 90 % des espèces considérées en « préoccupation mineure » .

Répartition des données et des taxons par catégories de menaces selon la liste rouge des espèces de France métropolitaine

Statut de menace	Nombre de données	Nombre de taxons
Préoccupation mineure (LC)	103 084	2 305
Quasi menacée (NT)	75 755	104
Vulnérable (VU)	6 903	56
En danger (EN)	3 590	25
En danger critique (CR)	555	9
Données insuffisantes (DD)	745	32
Non applicable (NA)	2 620	204

Parmi les données déposées et validées relatives aux taxons évalués sur la liste rouge pour la France métropolitaine, 2,1 % concernent des espèces considérées « en danger » ou « en danger critique », 40 % des espèces considérées comme « quasi-menacées » 53 % des espèces considérées en « préoccupation mineure ».

Répartition des données et des taxons par catégories de menaces selon la liste rouge des espèces de Guyane

Statut de menace	Nombre de données	Nombre de taxons
Préoccupation mineure (LC)	211	90
Quasi menacée (NT)	7	6
Vulnérable (VU)	8	6
En danger (EN)	67	2
En danger critique (CR)	2	1
Données insuffisantes (DD)	95	13
Non applicable (NA)	3	2

Parmi les données déposées et validées relatives aux taxons évalués sur la liste rouge pour la Guyane, 17,5 % concernent des espèces considérées « en danger » ou « en danger critique » et 54 % des espèces considérées en « préoccupation mineure ».

Répartition des données et des taxons par catégories de menaces selon la liste rouge des espèces de la Réunion

Statut de menace	Nombre de données	Nombre de taxons
Préoccupation mineure (LC)	1 630	106
Quasi menacée (NT)	310	8
Vulnérable (VU)	44	10
En danger (EN)	326	8
En danger critique (CR)	11	5
Données insuffisantes (DD)	52	9
Non applicable (NA)	302	20

Parmi les données déposées et validées relatives aux taxons évalués sur la liste rouge pour la Réunion, 12,6 % concernent des espèces considérées « en danger » ou « en danger critique » et 61 % des espèces considérées en « préoccupation mineure ».

Apports de connaissances

Liste des espèces citées dans les données déposées pour lesquelles aucune données n'avaient été mentionnées dans l'INPN.

Nom scientifique	Nombre de données
<i>Apseudes holthuisi</i> Bacescu, 1961	35
<i>Dipolydora giardi</i> (Mesnil, 1896)	31
<i>Palposyllis prosostoma</i> Hartmann-Schröder, 1977	29
<i>Cochlodesma praetenuae</i> (Pulteney, 1799)	16
<i>Litocorsa stremma</i> Pearson, 1970	11
<i>Iphimedia eblanae</i> Bate, 1857	9
<i>Iphimedia perplexa</i> Myers & Costello, 1987	6
<i>Argissa hamatipes</i> (Norman, 1869)	4
<i>Stenothoe tergestina</i> (Nebeski, 1881)	4
<i>Astacilla damnoniensis</i> (Stebbing, 1874)	3
<i>Acanthomysis longicornis</i> (Milne Edwards, 1837)	2
<i>Aporrectodea zicsii</i> (Bouché, 1972)	2
<i>Eulalia ornata</i> Saint-Joseph, 1888	2
<i>Peltocoxa damnoniensis</i> (Stebbing, 1885)	2
<i>Processa nouveli</i> Al-Adhub & Williamson, 1975	2
<i>Corynis lateralis</i> (Brullé, 1832)	1
<i>Dicologlossa cuneata</i> (Moreau, 1881)	1
<i>Eucera cineraria</i> Eversmann, 1852	1
<i>Lumbricalus campoyi</i> Aguirrezabalaga & Carrera-Parra, 2006	1
<i>Paramysis helleri</i> (G.O. Sars, 1877)	1
<i>Pomatoschistus norvegicus</i> (Collett, 1902)	1
<i>Procerastea nematodes</i> Langerhans, 1884	1
<i>Sclerocheilus minutus</i> Grube, 1863	1
<i>Taraxacum eminens</i> Soest, 1977	1
<i>Todaropsis eblanae</i> (Ball, 1841)	1
<i>Tychobythinus glabratus</i> (Rye, 1870)	1
<i>Veronica x lackzschewickii</i> J.Keller, 1942	1

Parmi les données considérées comme valides, 27 espèces n'avaient pas encore été mentionnées dans les données disponibles au niveau de l'INPN, pour un total de 170 données.

Analyses complémentaires avant diffusion des données dans le SINP

Contrôles SINP en vue de l'intégration des données

Les données déposées ont été soumises aux contrôles en vigueur dans le SINP au niveau de la plateforme nationale⁵ (doublon, conformité, cohérence, validation scientifique automatique). Sur les 479 339 données analysées :

- > **92 838 données** ont été considérées comme non conformes ou incohérentes (la principale cause de rejet étant la localisation en dehors du territoire français).
 - > **80,6 % des données transmises sont conformes et cohérentes.**

- > **79 165 données** ont été identifiées comme invalides ou pour lesquelles la validité scientifique n'a pas pu être établie, principalement en raison de l'absence de CD_NOM (les contrôles de validité scientifique s'appliquant seulement sur les données comportant un code taxon issu du référentiel TaxRef⁶ = CD_NOM).

Au final, 64 % des données déposées sont jugées conformes, cohérentes et considérées comme de qualité suffisante pour être diffusées.

(307 336 données – 6 042 taxons)

⁵ <https://inpn.mnhn.fr/programme/donnees-observations-especes/references/validation>

⁶ <https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref>

Métadonnées

L'analyse des métadonnées des cadres d'acquisition et des jeux de données a permis de mettre en évidence que :

- Les métadonnées des cadres d'acquisition sont globalement renseignées mais que leurs contenus sont biaisés car centrées sur les projets et non sur les études en tant que telles.
- La description des métadonnées des jeux de données est absente dans plus de 40 % des cas.
- Certains éléments importants notamment la déclaration du protocole, des producteurs de données, etc. n'ont pas été correctement renseignés.

Il a également pu être mis en évidence, dans de nombreuses études, le manque de structuration des jeux de données. Pour beaucoup de cadres d'acquisition (69 % du total), un seul jeu de données est associé alors que les métadonnées renseignées suggéreraient qu'une structuration de données plus poussée aurait été pertinente.

Réutilisation de données du SINP

En se référant aux identifiants uniques SINP d'occurrence de taxons déclarés, il apparaîtrait qu'aucune donnée versée dans le cadre de DEPOBIO n'ait été préalablement versée au SINP.

Il est cependant difficile à ce stade de conclure sur le fait que les données versées soient réellement nouvelles pour le SINP puisque d'une part l'analyse se base sur un comparatif avec les données actuellement disponibles au niveau de l'INPN et que des données régionales peuvent ne pas être encore disponibles à ce niveau et, d'autre part, que l'utilisation des identifiants lors des échanges a pu ne pas être mis en place par les différents opérateurs de la chaîne de circulation des données.

Duplicatas de données au sein des versements

Différents types de duplicatas (doublons) ont pu être identifiés au sein des données déposées sur DEPOBIO :

- L'ensemble des attributs sont identiques entre deux données ou plus.
- Seuls les identifiants uniques SINP d'occurrence de taxons sont dupliqués entre plusieurs données mais le reste des attributs est différents.
- Les données sont identiques à l'exception des identifiants uniques SINP.

Ces duplicatas ont différentes origines, justifiées ou non :

- La réutilisation de jeux de données pour des projets géographiquement proches
- Versements multiples de jeux de données (tests ou erreurs probables)

Ainsi, pour les trois types de doublons cités, 25 821 données ont été identifiées en tant que doublon soit 5,4 % du total des données versées.

Utilisation du référentiel Taxref⁷

Lors du dépôt, plus de la moitié des données versées disposaient d'un CD_NOM (57,7 % - 276 754 données).

La réconciliation taxonomique automatique a permis d'attribuer un CD_NOM pour 58,4 % des données restantes (118 247 données).

Finalement 84 338 données⁸ ne disposent pas de CD_NOM après l'étape de réconciliation taxonomique automatique (7 476 nom cités différents). Le nom cité non réconcilié le plus représenté est « Pipistrelle commune » et concerne 12 795 données. A l'inverse 3 598 noms cités différents ne sont mentionnés qu'une seule fois (pour une seule observation).

Par ailleurs la présence de nom cités non exploitables a été mise en évidence tel que : « 0 », « divers passereaux », « pelote rejection », « Route », « - », « Chalutier ».

Près de la moitié de ces données, se référant à un nombre limité de noms cités (moins de 20), se verront attribuer un CD_NOM par expertise manuelle.

⁷ <https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref>

⁸ Dont 5 614 données n'ayant pas satisfait les contrôles de conformité et cohérence

Les 20 premiers noms cités en terme de nombre d'occurrence (nombre de données correspondantes) non réconciliés automatiquement

Nom scientifique	Nombre de données
Pipistrelle commune	12795
Eptesicus/Nyctalus/Vespertilio sp.	4779
Pipistrelle de Kuhl	4498
Pipistrellus kuhlii/nathusii	2858
Parasite	2841
Pipistrellus/Hypsugo sp.	2562
Goeland leucophee	1637
Pipistrelle Commune	702
Pinson des arbres	700
Puffin yelkouan	663
Serotine commune ⁹	617
Sterne caugek	575
Cortaderia selloana (Schult. & Schult.f.) Asch. & Graebn., 1900	500
Alouette des champs	486
Pipistrellus pipistrellus/nathusii	474
0	463
Corneille noire	384
Pigeon ramier	381
Mouette melanocephale	371
Nasse réticulée - Nassarius reticulatus	362
	38 648

⁹ On peut ici constater la présence de problème d'encodage dans certains noms cités

Pistes d'amélioration

L'analyse a mis en évidence qu'un travail important restait à mener autour des métadonnées afin que les jeux de données soient mieux organisés et mieux décrits. Ces améliorations pourraient passer par des évolutions de la procédure en place ainsi que par une meilleure démarche d'accompagnement des acteurs autour de ce sujet. Des documents d'accompagnement pourraient être proposés et être appuyés par une démarche d'animation relayée au niveau local et régional.

Aucune donnée déposée via DEPOBIO n'a pu être rattachée à des données existantes dans l'INPN, plateforme nationale du SINP au travers de l'identifiant unique. Ce constat pourrait signifier :

- que l'ensemble des données livrées à DEPOBIO sont nouvellement partagées au sein du SINP
- que des données livrées à DEPOBIO proviennent de plateformes régionales n'ayant pas encore transmis les données à l'INPN
- que l'identifiant unique du SINP n'a pas bien été véhiculé au cours des échanges

Quelques doublons ont été mis en lumière au cours de l'analyse. Les doublons identifiés ont été traités afin de ne pas être intégrés dans le SINP. La gestion des doublons pourrait être améliorée en mettant en place, à l'entrée du SINP, des contrôles plus poussés d'identification des données doublonnées. L'usage de l'identifiant unique du SINP, et son implémentation dans les bases de données des producteurs seraient également à même de limiter leur apparition. Là encore, une animation collective autour de l'usage de l'identifiant pourrait être à même de résorber ce problème.

Par ailleurs, une évolution fonctionnelle du télé-service pour une meilleure prise en charge des jeux de données « réutilisés » pourrait permettre de mieux gérer le dépôt de données déjà disponibles dans le SINP ou versées parallèlement dans le cadre de différentes études.

D'un point de vue taxonomique, il est apparu que le référentiel TaxRef était plutôt bien utilisé (transmission du code unique CD_NOM) bien que l'information ne soit pas obligatoire lors du versement des données. La majorité des données déposées sans le code TaxRef ont pu être réconciliées avec le référentiel en utilisant le processus classique mis en place pour l'intégration des jeux de données dans l'INPN (réconciliation automatique associée à une expertise manuelle complémentaire). L'étape de réconciliation taxonomique pourrait éventuellement être réalisée plus tôt dans la chaîne de traitement (directement lors du versement des données dans DEPOBIO) afin que les services instructeurs et les utilisateurs de la plateforme DEPOBIO puissent bénéficier du rattachement à TaxRef.

Enfin, 16 % des données déposées (environ 76 000) sont reliées à des entités (points, lignes ou polygones) situées hors du territoire national et donc inexploitable dans le cadre du SINP. Ce constat semble être dû à des erreurs dans la saisie des localisations. Pour y remédier, des contrôles portant sur la localisation géographique des entités géométriques pourraient être mis en place au début de la procédure de dépôt.

Conclusion

Cette étude qui vise à caractériser et évaluer les données sur les espèces déposées dans le cadre de DEPOBIO a été commanditée par le Ministère de la Transition écologique et solidaire (MTES). Un des objectifs finaux est d'objectiver la capacité de réutilisation de ces données et d'identifier les pistes d'amélioration concrètes. Cette étude a été faite sur les données transmises dans les six premiers mois de fonctionnement de la procédure de dépôt, sur environ 480 000 données réparties en 718 jeux de données. Les projets concernent majoritairement les secteurs d'activités liées à la production d'énergie, notamment éolienne et photovoltaïque, ou l'extraction de matériaux et sont localisés en métropole ainsi qu'à la Réunion et en Guyane.

Les données sont majoritairement récentes avec plus de 80 % d'entre elles datant de moins de 5 ans et principalement issues d'observations de terrain effectuées sur le milieu terrestre (4 données sur 5). Elles sont majoritairement localisées précisément, les deux tiers étant rattachées à un point et plus globalement à 80 % à une géométrie. Les espèces citées concernent majoritairement les vertébrés (71 %) et les plantes (19 %). Les invertébrés sont minoritaires (10 %) et la fonge quasi-absente. Les espèces représentées sont principalement liées à des habitats terrestres ou continentaux (terrestre et/ou eaux douces). Les données concernent majoritairement des espèces évaluées en préoccupation mineure sur les listes rouges de l'UICN que ce soit à l'échelle mondiale, européenne ou nationale. Concernant les listes nationales, le pourcentage de données d'espèces considérées comme en danger ou en danger critique varie de 2 % pour la métropole jusqu'à 17,5 % pour la Guyane. Les données validées portent sur des espèces réglementées à hauteur de 42 %. 27 % des données concernent des espèces pour lesquelles un plan national d'action (PNA) est en cours ou a été mis en place par le passé.

Globalement, il a également pu être mis en évidence, dans de nombreuses études, un manque de structuration des jeux de données. Pour beaucoup de cadres d'acquisition (69 % du total), un seul jeu de données est associé alors que les métadonnées renseignées suggéreraient qu'une structuration de données plus poussée aurait été pertinente. Cette problématique se retrouve globalement à l'échelle du SINP et les travaux doivent donc se poursuivre dans ce cadre afin de mieux accompagner les producteurs pour la caractérisation des jeux de données. En raison des retours d'expérience des plateformes régionales du SINP et de certains indices indirects suggérant le contraire, il n'a pas été possible de conclure sur le fait que les données versées soient réellement nouvelles pour le SINP. Cela souligne l'intérêt de la mise en place des identifiants uniques le plus tôt possible au cours de la chaîne de production et l'importance de l'accompagnement sur les enjeux de réutilisation des données.

Le contrôle de ces données par les processus automatique préalable à l'intégration dans la plateforme nationale du SINP, montre que 64 % des données ont pu être jugées conformes avec le standard en vigueur dans le SINP, cohérentes et valides scientifiquement au regard des connaissances actuelles sur les statuts biogéographiques, les habitats et la répartition des espèces. Les données invalidées (36 %) sont majoritairement des données sans le code taxon ou situées hors du territoire français. Pour autant, au regard des résultats de contrôle, les données déposées et validées sont de qualités

suffisantes pour être mises à disposition au sein du SINP. Seules 4 426 données (1,5 %) valides et considérées comme sensibles se verront floutées lors de la diffusion sous licence ouverte du SINP.

Parmi les données valides, 27 espèces n'avaient pas encore été citées dans l'INPN, pour un total de 170 données. Cela confirme l'intérêt de diffuser et rendre publique les données issues des études environnementales.

En conclusion, les données déposées dans le cadre de DEPOBIO sont globalement conformes avec les préconisations du SINP et vont apporter des connaissances nouvelles jusqu'ici non partagées publiquement. Après ces quelques mois de fonctionnement, il est mis en évidence qu'un accompagnement des acteurs pour améliorer la traçabilité, passant notamment par l'utilisation des identifiants uniques, et la structuration des jeux de données est important à organiser pour la suite du projet. Les producteurs des données étant aussi des partenaires du SINP, la dynamique d'accompagnement pourrait être mutualisée.

EN SAVOIR PLUS

Rapports complets d'intégration des données:

<https://inpn.mnhn.fr/docs-web/docs/download/302488>

Présentation du dépôt légal :

<http://www.naturefrance.fr/reglementation/depot-legal-de-donnees-brutes-de-biodiversite>

Plaquette de présentation DEPOBIO:

http://www.naturefrance.fr/sites/default/files/fichiers/ressources/pdf/18158_depobio_brochure_web.pdf

Partage des données dans le SINP :

<https://inpn.mnhn.fr/programme/donnees-observations-especes/presentation>

Liste des jeux de données déposés :

<https://depot-legal-biodiversite.naturefrance.fr/versement/published-jdds/>

Dépôt légal de données brutes de biodiversité

Caractérisation et évaluation des données sur les espèces en vue de leur diffusion dans le SINP.

Référence du rapport conseillé :

LAVILLE, M., PONCET, L., TOUROULT, J., DUPONT, P., PANIJEL, J., VEST, F., ROBERT, S. 2019. *Dépôt légal de données brutes de biodiversité – Caractérisation et évaluation des données sur les espèces en vue de leur diffusion dans le SINP*. UMS PatriNat, AFB/CNRS/MNHN. 26 p.

Résumé :

Depuis la promulgation de la loi Biodiversité du 8 août 2016, toute personne morale ou physique porteuse d'un projet d'aménagement ou d'un document de planification conduisant au recueil de données de biodiversité doit verser ces données afin qu'elles puissent être valorisées et mises à disposition du public (articles L411-1 et L122-1 du code de l'environnement). Afin de permettre le dépôt des données brutes de biodiversité (DEPOBIO) et la description des métadonnées associées, un système spécifique a été mis service et ouvert en le 1^{er} juin 2018.

Nous présentons ici les résultats de l'analyse des données et métadonnées visant à objectiver la capacité de réutilisation de ces données, d'organiser leur diffusion dans le SINP et d'identifier les pistes d'amélioration concrètes. Cette étude a été faite sur les données transmises durant les six premiers mois de fonctionnement de la procédure de dépôt, sur environ 480 000 données réparties en 718 jeux de données.

En conclusion, les données déposées dans le cadre de DEPOBIO sont conformes avec les préconisations du SINP pour 64 % d'entre elles et vont apporter des connaissances nouvelles jusqu'ici non partagées publiquement. Après ces quelques mois de fonctionnement, il est mis en évidence qu'un accompagnement des acteurs pour améliorer la traçabilité, passant notamment par l'utilisation des identifiants uniques, et la structuration des jeux de données est important à organiser pour la suite du projet.

**AGENCE FRANÇAISE
POUR LA BIODIVERSITÉ**
ÉTABLISSEMENT PUBLIC DE L'ÉTAT

www.afbiodiversite.fr



www.cnrs.fr



MUSÉUM
NATIONAL D'HISTOIRE NATURELLE

www.mnhn.fr