DATA MANAGEMENT PLAN

INFORMATIONS GENERALES

Renseignements administratifs

Acronyme : SALMEVOL-1041

Titre : Evolutionary Ecology of Kerguelen Islands Colonization by introduced salmonids

Nom du coordinateur : Labonne

Prénom du coordinateur : Jacques

Affiliation : INRAE

Contact concernant le PGD : Jacques Labonne

Version du PGD : 1.

Date : 16.06.2023

1. DESCRIPTION DES DONNEES ET COLLECTE OU REUTILISATION DE DONNEES EXISTANTES

Les données concernées par le présent PGD sont générées par l'UMR 1224 ECOBIOP UPPA/INRAE.

La donnée de base est constituée par des captures de poissons de différentes espèces, localisées dans le temps et l'espace. Les informations minimums sont donc :

- La date
- L'espèce
- Les coordonnées géographiques en UTM Mercator, ainsi que la rivière (lieu-dit).
- La méthode de capture.

A ces données de base peuvent être optionnellement associées des données biométriques et des échantillons de tissus.

Les échantillons sont potentiellement :

- Des écailles
- Des morceaux de nageoires
- Des rayons de nageoires
- Des otolithes.
- Des contenus stomacaux
- Des carcasses entières

Les données biométriques de bases sont :

- La taille corporelle en mm
- Le poids en g.

Des données additionnelles sont produites après analyse, principalement par scalimétrie :

- L'âge total
- L'âge en eau douce
- L'âge en mer
- L'âge de première maturation
- La croissance en eau douce
- La croissance en mer

Le travail de scalimétrie produit aussi des fichiers graphiques (photographies d'écailles).

Des données de type démographiques peuvent être produites par analyse des données de capture en lien avec certaines méthodologies d'échantillonnages :

- Présence ou absence d'espèce par bassin versant et par date
- Effectifs pêchés lors d'échantillonnage à passage unique ou multiples, avec ou sans mesure de l'effort de pêche

Enfin, des documents descriptifs des échantillonnages sur le terrain ont été numérisés, et contiennent des informations de diverses natures.

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Les données sont recueillies sur le terrain en fonction des expéditions programmées, en accord avec l'Institut Polaire Français et la Réserve Intégrale des Terres Australes et Antarctiques Françaises. Les analyses additionnelles faites au laboratoire qui produisent des données dérivées sont aussi considérées comme relevant du monitoring Ces données ainsi que les données déjà acquises sont partie intégrante du Programme SALMEVOL 1041 mené par l'UMR ECOBIOP 1224 UPPA/INRAE, et constituent donc un investissement stratégique pour cette UMR et ses tutelles.

La récolte des données sur le terrain se fait soit en utilisant des systèmes électroniques robustes (tablettes, PC portable) soit en utilisant des carnets étanches adaptées aux conditions météorologiques extrêmes. Les données sont ensuite retranscrites sur des tables et envoyées en métropole. Elles sont vérifiées et compilées par un technicien responsable du monitoring. Il est à noter que des données anciennes sont aussi régulièrement retrouvées dans des archives papier (les informations remontant jusqu'à 1954). Ceci produit une seule table de données dont l'unité de base (la ligne) est la capture d'un individu. En date du 1er janvier 2023, cette table contient plus de 30 colonnes et plus de 157 500 lignes.

Ensuite, différentes structures de données sont produites à partir de cette table, pour différents usages : illustration, analyse, mise à disposition du public ou de services de type Infrastructure de Recherche (IR Rare avec le CRB COLISA, IR RZA avec la Zone Atelier Antarctique et Terres Australes).

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Les données de captures d'individus, de traits d'histoire de vie, et de disponibilité des échantillons sont au format numérique (tableurs).

En particulier, la table de données principale est au format EXCEL : ce format est préféré par notre technicien référent, de façon à pouvoir modifier facilement les données, sachant que cette table fait l'objet de modification régulières liées au data mining dans nos archives. Les autres tables extraites de cette table principale ont des formats divers (.xlsx, .csv, .R) selon leur destination. Ces tables de données occupent un espace total inférieur à 1Go (01.06.2023).

Les images d'écailles sont numérisées au format JPG ou PNG. et requièrent environ 14Go d'espace de stockage pour 15883 fichiers (01.06.2023).

Les documents de pêche sont numérisés au format PDF, et requièrent environ 1.5Go d'espace de stockage pour 219 fichiers (01.06.2023).

2. DOCUMENTATION ET QUALITE DES DONNEES

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

La table principale des données comporte un feuillet qui décrit chaque variable/colonne. Ce feuillet donne aussi une description des modes d'échantillonnage. Les unités de mesure sont intégrées au nom des variables.

Une partie de ces données sont en cours d'intégration dans un observatoire de la biodiversité en lien avec la Zone Atelier Antarctique et Terres Australes, qui décrit les variables essentielles de biodiversité.

Une documentation est aussi fournie sur les méthodes d'ageage et de mesure de croissance en scalimétrie.

La qualité des données est assurée par le technicien référent au moment de l'intégration des données au retour des expéditions, par des vérifications croisées avec les preneurs de données, et sous la responsabilité du responsable scientifique du programme SALMEVOL.

3. STOCKAGE ET SAUVEGARDE PENDANT LE PROCESSUS DE RECHERCHE

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

Lors de la récolte des données sur le terrain, les données sont le plus souvent stockées sur des carnets de terrain étanche (en lien avec les conditions environnementales extrêmes) qui sont temporairement (durée de la campagne d'échantillonnage soit plusieurs mois) sauvegardées sous format numériques (photographies des carnets).

Une fois sur base (Port-aux-Français), ces données sont transposées sur des tableurs, via des ordinateurs portables, et une sauvegarde sur disque dur externe est effectuée. Si les communications le permettent, une copie est envoyée en métropole.

Une fois en métropole, les données sont vérifiées (recroisement entre carnets de terrain et tableurs), et intégrées à une table de donnée générale préexistante. Cette table est sauvegardée sur un disque dur externe, dans un espace de stockage NextCloud personnel du technicien référent, et sur un espace NextCloud commun (nommé kerguelen-data) de l'unité (UMR ECOBIOP 1224) (voir Figure 1). Ces espaces NextCloud sont localisés sur un DataCenter INRAE. Ce service est fourni à notre UMR par notre tutelle, et n'est pas dépendant de l'obtention de projets. Cet espace NextCloud ne peut être modifié que par 3 personnes : le technicien référent, le scientifique responsable du programme SALMEVOL, et le responsable informatique de l'UMR ECOBIOP.

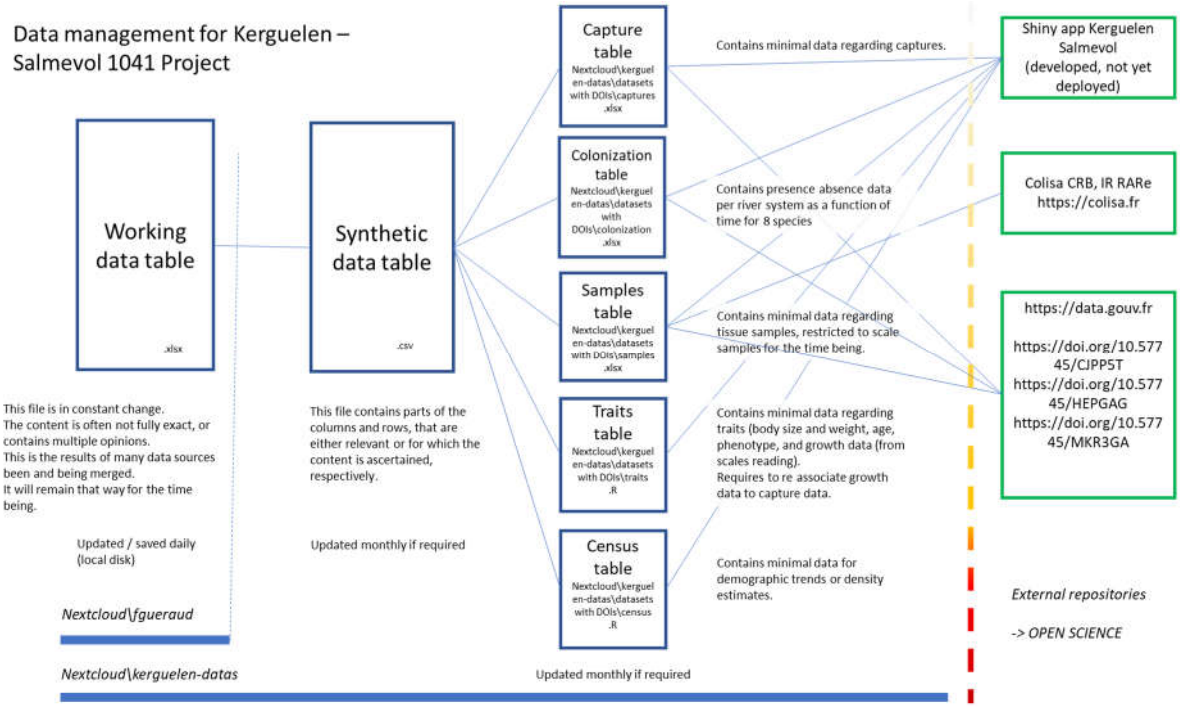

Figure 1 : structure des données et localisation en fonction des objectifs.

Après ce travail de curation effectué par le technicien référent, la totalité des données est conservée.

La table de données principale est utilisée pour produire des tables spécifiques ayant chacune un objectif spécifique. Certaines de ces tables spécifiques sont mises à disposition pour différentes utilisations ouvertes de type OPEN SCIENCE (voir § 4). Il contient plusieurs sous-dossiers, comme décrit dans la Figure 2.

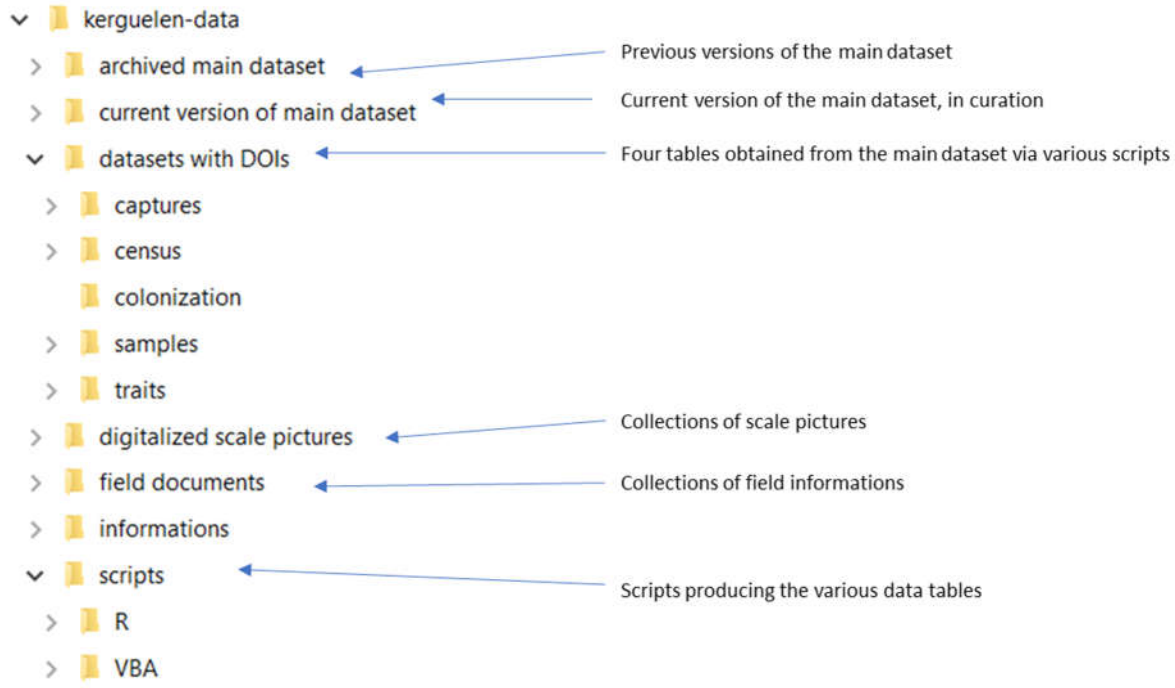

Figure 2 : arborescence du bucket kerguelen-data localisé sur le NextCloud de l'UMR ECOBIOP.

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du
processus de recherche ?

Concernant la phase d'acquisition des données, l'existence en format simultanément analogique et numérique des données garantit la possibilité de récupération en cas d'accident.

Pour les phases d'exploitation des données comme pour celles de mise à disposition, la sécurité des données est la responsabilité de INRAE qui est gestionnaire de ses propres Datacenters.

Nous ne produisons pas de données sensibles ou à caractère personnel.

4. EXIGENCES LEGALES ET ETHIQUES, CODES DE CONDUITE

4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les
données à caractère personnel et sur la sécurité des données sera-t-il assuré ?

Non concerné.

4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

Les données sont la propriété des tutelles de l'UMR ECOBIOP (INRAE, UPPA) ainsi que, pour partie, des autres entités qui peuvent participer plus ponctuellement à leur production (exemple : collaboration avec un institut ou une université tierce pour certains projets au sein de SALMEVOL).

Les personnes ayant accès aux données lors du processus de recherche sont l'ensemble des collaborateurs du programme SALMEVOL. Tout collaborateur du programme doit mentionner l'origine de ses données ainsi qu'intégrer un membre de l'UMR ECOBIOP dans les auteurs de publications. Il faut noter que les données ne perdent jamais leur caractère stratégique pour la recherche dans le programme SALMEVOL et pour l'UMR ECOBIOP. Ce point est central dans la démarche de recherche écologique à long terme (label LTER lié à Zone Atelier Antarctique et Terres Australes). Il faut enfin noter que l'acquisition des données est directement dépendante du support de l'Institut Polaire Français pour la mise en œuvre des programmes scientifiques. L'institut Polaire Français exige la mise à dispositions des meta-données et encourage la mise à disposition des données, ce que l'UMR ECOBIOP applique.

En effet, nous ouvrons une partie de ces données via diverses démarches (voir § 5), et l'ensemble de ces données peut être accessible si l'on devient collaborateur de SALMEVOL.

4c. Comment les éventuelles questions éthiques seront-elles prises en compte, et les codes déontologiques respectés ?

Tout protocole expérimental ainsi que les processus de monitoring à long terme sont évalués par un comité à l'éthique animal (le CEA 73 dans notre cas), ainsi que par le comité à l'environnement polaire (CEP) de la réserve naturelle intégrale de Terres Australes et Antarctiques (TAAF). Ces instances peuvent affecter la façon dont les données sont obtenues, mais elles n'ont pas d'impact sur l'utilisation et la mise à disposition des données.

5. PARTAGE DES DONNEES ET CONSERVATION A LONG TERME

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

L'ouverture et le partage des données au sein de SALMEVOL suit les principes institutionnels : maximiser l'impact scientifique et sociétal des investissements dans la recherche et la production de connaissance scientifique. Plus généralement, notre démarche est de s'inscrire dans les principes FAIR. La philosophie générale est de conserver un accès privilégié aux données pour les collaborateurs de SALMEVOL sur une période de 3 ans, le temps d'exploiter et publier les données récoltées. Au-delà de cette durée, les données peuvent être en partie ouvertes, par mise à disposition directe, par intégration dans des services, ou pour des objectifs d'éducation à la science.

Le terme « en partie » limite cette ouverture à des données ne relevant pas de l'interprétation : par exemple, la scalimétrie permet d'estimer des âges en eau douce, en mer, de la croissance, de l'âge de première reproduction. Pour autant, il s'agit là d'estimations, et ces estimations sont réalisées dans des contextes particuliers qui ne peuvent être résumés à des méta-données. Pour accéder à ces données interprétées, le demandeur doit faire une demande collaboration, afin que nous puissions contrôler l'utilisation de ces données dans un but de qualité de la science qui sera produite avec.

Les données sont donc partagées de différentes façons, après exploitation ou après un délai de 3 ans.

- Les données de capture sont mises à disposition sur un serveur ouvert (https://data.gouv.fr)
- Les données de colonisation sont mises à disposition sur un serveur ouvert (https://data.gouv.fr)
- Les données concernant la disponibilité des échantillons sont partagées via le CRB COLISA.
- Les données concernant les traits ne sont pas mises à disposition directement : elles relèvent de l'analyse, et requièrent une expertise pour leur utilisation qui dépasse la simple fourniture de métadonnées. Elles peuvent être communiquées dans le cadre d'une collaboration.
- Les documents de pêche ont vocation à être partagés.
- Les images d'écailles numérisées peuvent être partagées dans le cadre d'une collaboration.

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par
ex. un entrepôt de données ou une archive) ?

Le plan de partage des données est décrit dans la figure 2. Il souligne l'existence de 3 espaces séparés. Le premier concerne le jeu de données principal (Working data table dans la Fig. 2), sur lequel le technicien référent est susceptible de travailler régulièrement. Cette table est sauvegardée sur deux espaces NextCloud différents (décrits en § 3), et une copie locale sur un disque dur est effectuée quotidiennement. Un script permet d'extraire une version synthétique (Synthetic data table), qui contient une grande partie des informations nécessaires aux objectifs de recherche et de partage. Cette table synthétique est déposée sur le NextCloud partagé de l'UMR ECOBIOP dans le bucket kerguelen-data (voir figure 1). De cette table synthétique sont extraites 4 tables de données, chacune avec un objectif spécifique.

La table des captures (« captures ») fournit des données synthétiques sur tous les animaux capturés, et elle est mise en partage sur data.gouv.fr (https://doi.org/10.57745/CJPP5T).

La table des colonisations (« colonization ») fournit des données synthétiques sur la dynamique de colonisation (présence/absence) des différents bassins versants par huit espèces de salmonidés. Elle est mise en partage sur data.gouv.fr (https://doi.org/10.57745/HEPGAG).

La tables des échantillons (« samples ») permet d'alimenter le CRB COLISA de l'IR RARe (https://colisa.fr).

La table des traits d'histoire de vie (« traits ») regroupe des données de bases sur les âges des poissons, et les associe à des données de croissance (ces dernières ne venant pas de la table synthétique).

La table de census (« census ») permet de regrouper des informations qui pourront servir à l'établissement de modèles d'estimation de la densité.

Ces deux dernières tables sont générées pour des objectifs de question scientifique, et ne sont pas encore ouvertes (au 01.06.2023).

Les documents descriptifs de pêche sont destinés à être partagés via data.gouv.fr. Ils sont aussi stockés dans le NextCloud kerguelen-data (voir Figure 1)

Les images numérisées d'écailles ne sont pas partagées pour le moment (au 01.06.2023), et sont également stockées dans le NextCloud kerguelen-data.

Les métadonnées sont fournies pour toutes les données publiées sur data.gouv.fr.

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

Les formats utilisés sont accessibles par la majorité des logiciels en licence libre (en particulier R et les alternatives ouvertes à Microsoft Office).

Concernant les listes d'échantillons, leur consultation se fait via le CRB COLISA (https://colisa.fr) et les demandes se font par le même système.

L'accès aux données partagées sur data.gouv.fr se fait directement selon la charte ETALAB 2.0. Celle-ci implique le respect de la propriété intellectuelle et la citation du jeu de données utilisé par son DOI.

Pour les données qui ne sont pas directement partagées, se reporter à § 5a.

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de
données ?

En accord avec le plan décrit en Figure 2, les jeux de données publiés sur data.gouv.fr bénéficient d'un DOI.

6. RESPONSABILITES ET RESSOURCES EN MATIERE DE GESTION DES DONNEES

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-
à-dire le gestionnaire des données) ?

Un technicien référent ECOBIOP gère la compilation, la curation et la mise à disposition des donnée, assisté par un chercheur ECOBIOP (0.2 et 0.05 ETP respectivement). Cet investissement est prévu dans la gestion des compétences de l'UMR ECOBIOP, le programme SALMEVOL à long terme revêtant un caractère stratégique pour l'unité. Les infrastructures informatiques sont fournies par INRAE. LE DMP est mis en œuvre et révisé par ces personnels, sous la supervision d'un Assistant-Ingénieur impliqué dans la déclinaison des DMP au département ECODIV INRAE.

Le présent DMP est issu du format fourni par l'ANR, et sera révisé en fonction de la déclinaison institutionnelle des DMP de INRAE dans le département ECODIV.

6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

Voir § 6a.