RAPPORT

Conservatoire d'espaces naturels Provence-Alpes-Côte d'Azur

Organisation de l'intégration des données faunistiques dans Silene - 2022

Novembre 2022

Conservatoire d'espaces naturels Provence-Alpes-Côte d'Azur

Rapport

Organisation de l'intégration des données faunistiques dans Silene - MAJ 2022

Document réalisé par :

Conservatoire d'espaces naturels de Provence-Alpes-Côte d'Azur

Coordination :

Julie DELAUGE – Adjointe à la Direction - Responsable Connaissance et Programmes

Rédaction :

Géraldine KAPFER – Responsable du Pôle Biodiversité Régional

Paul Honoré – Chargé de mission – base de données

Fanny Guillaud – Chargée de mission assistance technique et scientifique

Hélène CHAUVIN – Animatrice Silene

Relecture :

Julie DELAUGE – Adjointe à la Direction - Responsable Connaissance et Programmes

Géraldine KAPFER – Responsable du Pôle Biodiversité Régional

Participation :

Stéphane Bence

Sonia Richaud

Cédric Roy

Fanny Guillaud

Date de réalisation : novembre 2022

Citation recommandée :

Delauge J., Kapfer, G. Honoré P., Guillaud F., Chauvin H., 2022. Organisation de l'intégration des données faunistiques dans Silene. Conservatoire d'espaces naturels de Provence-Alpes-Côte d'Azur – MAJ 2022. Sisteron, 30 p.

PRÉFET DE LA RÉGION PROVENCE- ALPES-CÔTE D'AZUR
Liberté Égalité Fraternité

RÉGION SUD PROVENCE ALPES CÔTE D'AZUR

Table des matières

Préambule

Silene est un outil public et collectif au service de la prise en compte de la biodiversité. Il s'inscrit dans la dynamique générale de mise à disposition de l'information environnementale (convention d'Aarhus, directive Inspire) et plus particulièrement le Système d'Information de l'iNventaire du Patrimoine naturel (SINP). Soutenu par la DREAL et le Conseil Régional, Silene est développé et administré par :

- les Conservatoires Botaniques Nationaux Méditerranéen et Alpin (CBNMED et CBNA) pour les volets flore et habitats ;
- le Conservatoire d'Espaces Naturels de PACA pour le volet faune.

Ensemble, ils partagent un cadre commun de référence : la charte Silene et ses principes de gouvernance.

De nombreuses structures participent à Silene Faune en y intégrant leurs données. On distingue les fournisseurs :

- « partenaires » qui transmettent volontairement leurs données. Ils sont liés par une convention et choisissent la périodicité de l'envoi et de l'intégration de leurs données.
- par obligation de restitution (programmes soutenus par des financements publics, retour de donnée suite à un droit d'accès ponctuels, etc.). Ils sont liés par la convention d'accès aux données et envoient leurs données selon une périodicité variable.

En raison de la multitude des formats de bases de données reçues, la mise en conformité des lots de données avec le format de données de Silene est nécessaire pour le traitement et l'intégration dans la base de diffusion. Cette mise en compatibilité du fichier reçu avec le format Silene nécessite obligatoirement une procédure contenant plusieurs phases de traitement et visant la meilleure intégration possible des données dans Silene.

Ce rapport présente cette procédure afin de fournir un cadre détaillé explicitant les choix et évolutions apportés annuellement au gré des problèmes rencontrés et des améliorations techniques possibles.

La présente note intègre donc les évolutions apportées :

- en 2014 et 2015 concernant :
  - le traitement de la donnée anonyme
  - le traitement des données hors PACA
  - la mise à jour des données descriptive de Silene ;
- en 2016, sur le traitement des doublons ;
- en 2017, sur le déroulement d'une mise à jour de Silene ;
- en 2018 et 2019, sur le traitement des doublons ;
- en 2020, sur le traitement des données situées sur des communes ayant fusionnées, et sur l'anonymisation des observateurs en raison du RGPD qui a été mise en place ;
- en 2021 : la migration de Silene sous GeoNature avec deux outils (Silene Nature et Silene Expert) et la mise en place d'une plateforme de dépôt de fichiers de données Silene.

Introduction

Afin d'être intégrée dans Silene, une donnée faune doit comporter quatre champs obligatoires conformes et cohérents :

- le nom de l'espèce,
- la date,
- la commune,
- l'observateur. Exception pour les observations avec observateur non mentionné et l'anonymisation à la demande de l'observateur (RGPD).

L'intégration des données d'un fichier dans Silene implique plusieurs phases de travail (Figure 1 - ci-dessous) s'échelonnant depuis la réception du fichier à la diffusion de ces données sur Silene :

- Phase 1 : procédure de réception
- Phase 2 : mise en correspondance des fichiers pour traitement automatique des données / Création de la métadonnée
- Phase 3 : analyse technique des données (cohérence / conformité)
- Phase 4 : gestion des doublons et traitement différentiel
- Phase 5 : qualification et validation scientifique de la donnée
- Phase 6 : retour sur l'intégration

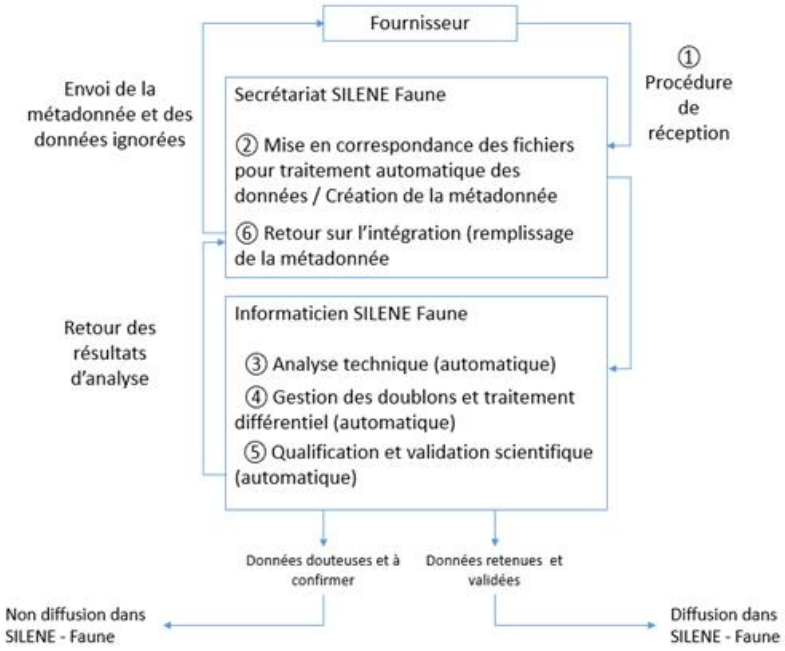

FIGURE 1 : SCHEMA SIMPLIFIE DU PARCOURS D'UN FICHIER

Déroulement du traitement d'un fichier

Phase 1 : Réception et pré-contrôles

a. Etape préliminaire (réception, sauvegarde, contrôle)

La transmission des données faune et flore en vue de leur intégration à Silene se fait principalement depuis la plateforme de dépôt mise ne place en septembre 2021 : https://silene.eu/contribuer/interface-de-depot-de-fichier/

Le nom du fichier reçu, la date de réception et le nom du fournisseur sont saisis dans un fichier de suivi qui permettra de suivre les étapes de traitement du fichier (non conforme et non intégrable, mise en conformité jusqu'à l'envoi pour intégration).

Un contrôle préliminaire visant à éviter l'intégration de doublons potentiels dans le système est effectué. Ce contrôle consiste en un pré-tri à réception du fichier. Il s'agit de vérifier que le jeu de données n'est pas déjà présent dans le système (jeux de données en doublons).

Des échanges peuvent avoir lieu à ce moment précis avec le fournisseur de données afin de compléter certaines informations importantes manquantes (sans nécessité de traitement informatique).

b. Pré-contrôle manuel de conformité et de cohérence du fichier de données

A la réception d'un fichier de données, la forme globale du fichier est analysée afin de repérer les éventuels problèmes non détectables par le script de contrôle automatique et de s'assurer de la conformité et la cohérence avec le format de SILENE. Les contrôles sont les suivants :

- Conformité : contrôle de la présence et du remplissage systématique des quatre champs obligatoires : commune, espèce, date, observateur.
- Conformité et cohérence : contrôle de la précision de localisation. La localisation d'une donnée peut être précise, au lieu-dit ou communale mais ce niveau de précision n'est souvent pas indiqué dans le fichier. Ce champ peut être interprété manuellement ou demandé au fournisseur.
- Conformité : Contrôle des données négatives. Silene n'intègre pas les données d'absences. Les données comportant « 0 » dans le champ effectif ou nombre sont donc recherchées et analysées (lecture du champ « commentaires » ou échange directe avec le fournisseur) pour savoir s'il s'agit d'une donnée d'absence ou d'une absence d'information. S'il s'agit d'une donnée d'absence, le « 0 » est remplacé par « absence » afin d'être rejeté automatiquement. S'il s'agit d'une absence d'information le « 0 » est remplacé par « indéterminé ».

Cette phase se fait l'objet d'échanges avec le fournisseur de données afin que les prochains échanges se basent sur des standards plus partagés.

c. Cas particulier des fichiers MNHN

Les données MNHN sont réparties dans une dizaine de fichiers qu'il convient de réagréger pour le traitement classique des fichiers Silene.

Il convient dans un premier temps de séparer la faune et la flore. Les données sans localisation précise et sans commune ne sont pas prises en compte.

Ensuite, il faut déterminer, parmi les données faune, lesquelles sont déjà présentes dans Silene. La base du MNHN est alimentée par de multiples sources et c'est pourquoi le transfert du niveau national au niveau régional amène un lot conséquent de doublons qui nécessite un pré-traitement. Le volume des données reçues implique ce traitement particulier en amont du traitement classique.

Un premier comparatif est réalisé sur les données déjà reçues en 2015, 2018 et 2020.

On procède ensuite à plusieurs comparatifs draconiens avec les données déjà présentes dans Silene, ce qui permet d'aboutir à un jeu réduit de données équivalent à un peu plus de 10% du jeu de données faune fourni par le MNHN.

On procède enfin à l'agrégation des données en un seul fichier qui sera tout de même soumis au traitement classique d'intégration.

Phase 2 : Mise en correspondance - Création de la métadonnée

Le fichier de données est ensuite mis en correspondance avec le format standard SILENE en précisant les libellés des champs afin qu'ils soient automatiquement reconnus. Cette analyse permet d'intégrer tous les champs pouvant être standardisés au format Silene. Les champs présents dans le fichier et non présents dans le format Silene sont concaténés dans le champ « Remarque » afin de ne perdre aucune information (utilisation du libellé « recup_Titre libre »).

Il existe plus de 76 libellés différents permettant d'intégrer les informations relatives à l'espèce observée, à la date d'observation, aux observateurs et déterminateur, à la localisation et à diverses précisions comme l'effectif, le stade de développement ou encore la plante support et la météo.

Pour exemple, ci-dessous un tableau listant les libellés relatifs à la localisation :

Info	Libellé	Description	Format
Localisation	dep	numéro du département	sur 2 digits (ex : 13)
	commune	Nom de la commune	en majuscule, sans article ni accent ni espace (ex : ISLE-SUR-LA-SORGUE)
	codeinsee	code insee	sur 5 digits
	lieudit	lieu-dit de la base bdnyme	visualisable dans BDnyme
	secteur	lieu-dit de la base bdnyme	visualisable dans BDnyme
	alt	altitude	nombre
	lat	latitude en degré décimal WGS84	
	lon	longitude en degré décimal WGS84	
	lond	partie degré de la longitude	
	lonm	partie minute de la longitude	
	lons	partie seconde de la longitude	
	latd	partie degré de la latitude	
	latm	partie minute de la latitude	
	lats	partie seconde de la latitude	
	lambert93x	coordonnée X en Lambert 93	
	lambert93y	coordonnée Y en Lambert 93	
	lambertx	coordonnée X en Lambert II étendu	
	lamberty	coordonnée Y en Lambert II étendu	
	utm31e	coordonnée Est en UTM 31	
	utm31n	coordonnée Nord en UTM 31	
	utm32e	coordonnée Est en UTM 32	
	utm32n	coordonnée Nord en UTM 32	
	wbs	système de coordonnées	champ libre
	prec	type de précision de localisation : Commune (C), Lieu-dit (Topologie), Point précis (P)	C / T / P
	remarque	commentaire relevé ou localisation	commentaire relevé ou localisation : champ libre max ?

Une fois les champs mis en correspondance, le fichier est enregistré au format texte séparateur tabulation et envoyé à l'informaticien administrateur de données pour analyse technique et scientifique automatique. L'envoi de ce fichier s'accompagne de la précision du fournisseur du fichier que l'informaticien renseignera automatiquement comme « source » pour toutes les données du fichier.

a. Traitement particulier des fichiers

Un travail plus important de mise en conformité et correspondance peut être fait lors de la première intégration de données pour un fournisseur adhérent au SINP. Cette opération, réalisée au cas par cas en fonction du format du fichier, consiste à récupérer le maximum d'informations en mettant en forme le jeu de données. Cette étape permet de faire évoluer au besoin le format d'échange du fournisseur avec un accompagnement et/ou de créer un script spécifique de mise en correspondance.

Par exemple, le fichier de la SFO au format CILIF contient de très nombreux champs spécifiques au protocole CILIF. Il contient le détail de l'effectif observé d'une espèce en fonction des différents stades de développement et comportements observés. Ce champ a été transposé au format Silene de façon à renseigner les champs stade de développement et comportement dans Silene. Sans cette mise en correspondance et reformatage, l'information n'aurait pas pu être retenue par le script de contrôle automatique qui est généraliste.

b. Création de la métadonnée descriptive

Le fichier source mis en forme est archivé et renommé avec le format suivant : ANNEEDERECEPTION_FOURNISSEUR_numéro-incrémenté (ex : 2017_PROSERPINE_1). La création des métadonnées peut alors avoir lieu.

La métadonnée du fichier reçu, nommée « Métadonnées_NOMDUFICHIER » (ex : Métadonnées_2017_PNPC_3) sert en premier lieu à décrire le fichier d'origine de façon générale et ultérieurement à rapporter le résultat des analyses techniques et scientifiques (cf. §.6.b. Finalisation de la métadonnée). La partie renseignée à cette étape est présentée ci-dessous (Figure 2) :

Element de la métadonnée		Libellé	Remarque
Intitulé de la ressource		2022_SEGED_5	
Nom des fichiers reçus		Tableau SILENE_Les Mées_vf	
Description de la ressource		Données Faune	
Type de ressource		Série de données géographiques ponctuelles	
Etendue spatiale des données		commune de LES MEES	
Etendue temporelle des données		du 12/03/2017 au 28/06/2018	
Nombre de données		190	
Identifiant unique des données	présent dans la ressource	NON	
	conservé par le partenaire	-	
Groupe(s) taxonomique(s) concerné(s)		Faune	
Date de mise à disposition		05/08/2022	
Champs obligatoires renseignés	Espèce	OUI	
	Observateur	OUI	
	Date	OUI	
	Localisation	OUI	
	Origine de la donnée	OUI	
	Diffusion de la donnée	OUI	
Echelle de localisation		POINT PRECIS	
Citation de la source des données		OUI	
Respect du cadre de la convention		OUI	
Validité des données de la ressource		Validation nécessaire	
Localisation de la ressource sur le réseau		HélèneChauvin\OneDrive - CEN PACA\1_SILENE\Administration_donnees\Silene_donnees_a_integrer\2022\2022.08\Seged_Environnement	
Fournisseur		SEGED-Environnement	
Personne référente au sein de la	Nom	Bruno CATALDO	
structure fournisseur	Contact mail	bcataldo@seged-environnement.com	
Programme(s) d'acquisition		Déviation d'une canalisation de gaz sur la commune des Mées	
Date des métadonnées		29/08/2022	

FIGURE 2 : EXEMPLE D'UNE METADONNEE REMPLIE A LA RECEPTION D'UN FICHIER

Phase 3 : Analyse technique des données – Contrôles de conformité et de cohérence des données

L'analyse technique des données comprend plusieurs opérations permettant de :

- mettre en conformité chaque donnée en recherchant et traitant les champs obligatoires, leurs champs associés et d'autres champs complémentaires ;
- rechercher les éventuels doublons en comparant les données du fichier entre-elles et avec celles intégrées à Silene ;

Lorsque les champs obligatoires sont non conformes et ne peuvent pas être mis en conformité, la donnée est rejetée et le producteur en est informé. Il doit alors corriger les informations non conformes pour que les données puissent être intégrées dans Silene.

Lorsque les attributs non conformes concernent des champs facultatifs, ils sont déportés dans un champ dédié, les champs non conformes sont mis à blanc. La donnée devient conforme et peut alors circuler dans le système.

En fonction du résultat de ces opérations, la donnée est soit :

- conforme et peut passer à l'étape de l'analyse scientifique ;
- rejetée dès lors qu'elle ne passe pas un contrôle et ne peut pas être mise en conformité. La donnée est alors répertoriée dans un fichier des données rejetées et accompagnée de la raison du rejet.

Cette analyse technique est automatisée grâce à plusieurs scripts de traitement de la donnée (non détaillés ici mais faisant l'objet de scripts commentés). Les contrôles et mises en conformité sont présentés ci-dessous, dans l'ordre de leur mise en œuvre.

a. Analyse préliminaire

Une analyse préliminaire permet d'identifier les lignes vides et de vérifier via un script que les champs minimums requis sont renseignés.

Un identifiant unique Silene est affecté ensuite à chaque donnée. Un UUID SINP est créé ultérieurement lors du transfert des données vers l'outil GeoNature (hors données MNHN, PNE et PNM qui possèdent déjà un UUID).

Certains fichiers bénéficient également de traitement particulier au préalable : MNHN / Parc national des écrins / Parc national du Mercantour. En effet, pour ces fournisseurs, l'intégralité des données étant à chaque fois réceptionnée, il convient d'éliminer les données prises en compte lors des précédents traitements. Il faut également reporter les modifications ou les suppressions de données qui ont eu lieu entre temps.

b. Taxon (mise en conformité)

Taxon	
Champ obligatoire	OUI
Elément(s) retenu(s)	Genre, Espèce, Sous-espèce

Différents champs peuvent renseigner l'espèce observée : le nom latin, le nom vernaculaire, le cd_nom, etc. Le script recherche la correspondance de ces champs avec le référentiel taxonomique **TAXREF v15** modifié (ajout d'espèces non présentes dans TAXREF et présentes en PACA, modification de noms d'espèces erronées, doubles taxons, etc.) en ne retenant que les taxons de rang générique, spécifique et sous-spécifique.

Pour l'intégration des données aquatiques il existe une table de correspondance qui permet de transformer certains noms de genre, pour lesquels une seule espèce est présente en PACA, en espèce.

Le schéma ci-dessous résume le principe du script analysant le nom d'espèce.

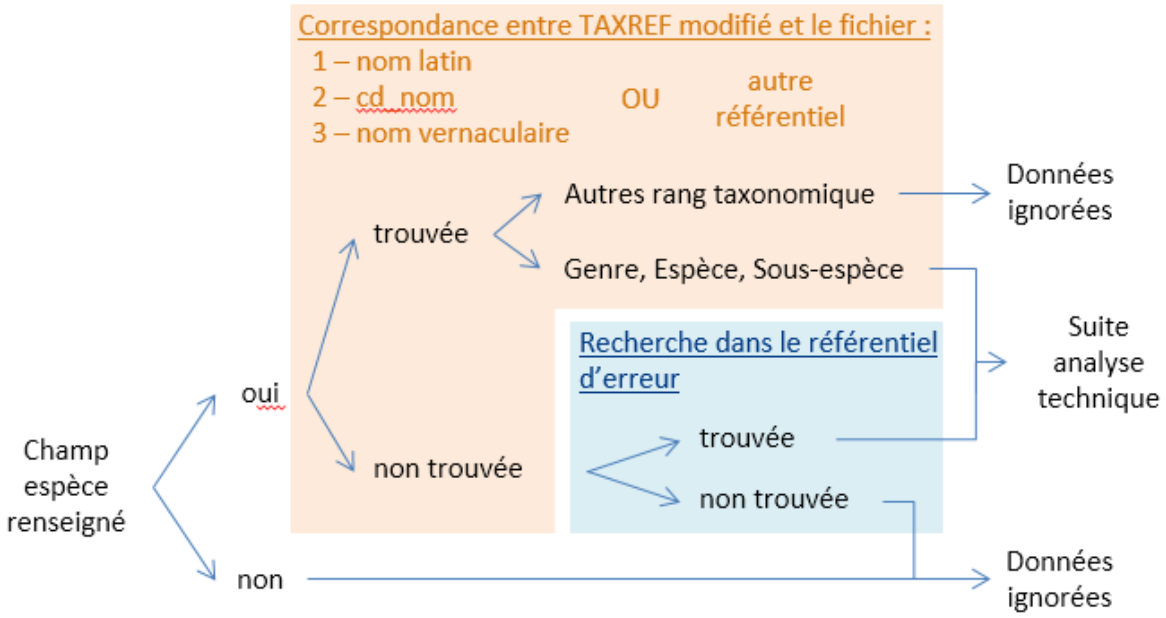

FIGURE 3 : SCHEMA GENERAL D'ANALYSE DU NOM D'ESPECE

Le script recherche prioritairement la correspondance avec le nom latin, puis avec le cd_nom, avec le nom vernaculaire et enfin le CODEESP (code historique).

Une donnée est rejetée lorsque le nom d'espèce :

- n'est pas renseigné ;
- correspond à un rang taxonomique non pris en compte dans Silene (ex : Famille, Ordre…) ;
- est absent du référentiel taxonomique.

Dans le cas d'un nom latin ou d'un nom vernaculaire mal orthographié, un référentiel des erreurs résolues permet la correspondance avec TAXREF modifié. Ce référentiel s'améliore en continu et permet d'optimiser l'intégration de la donnée.

Le nom affiché dans Silene est le nom valide TAXREF v15 et le nom présent dans le fichier est enregistré dans le champ « nom origine » par souci de traçabilité.

c. Observateur (mise en conformité)

Observateur	
Champ obligatoire	OUI
Elément(s) retenu(s)	Nom de l'observateur

L'auteur de l'observation est analysé à partir du référentiel « auteur Silene » (référentiel nominatif).

Si l'auteur n'est pas présent dans ce référentiel, il y est ajouté. Ce référentiel permet la mise en correspondance entre les noms valides, les initiales et les noms erronés trouvés antérieurement. Ce référentiel est en constante évolution.

En 2014, il a été convenu avec le comité d'administrateurs et les référents faune de retenir la donnée anonyme afin de la rendre disponible.

Définition : une donnée anonyme est une observation sans observateur mentionné ; le champ est vide (case non remplie) ou absent (colonne absente) de la base de données transmise. L'observateur non reconnu (code ou prénom ou initiales…) n'est pas une donnée anonyme et elle ne sera pas intégrée.

Le nom de la structure à la source de la donnée qui peut être différent du fournisseur du fichier est recherché dans le champ observateur, organisme ou autres (recup_producteur).

Si elle est identifiée : le champ observateur est complété par <Non mentionné – NOM DE LA STRUCTURE>

Si elle n'est pas identifiée : le champ observateur est complété par <Non mentionné – NOM DU FOURNISSEUR>

L'indication « Non mentionné » : peut être issu de l'anonymisation RGPD ou d'une absence d'information.
Le schéma suivant résume le principe du script analysant l'observateur.

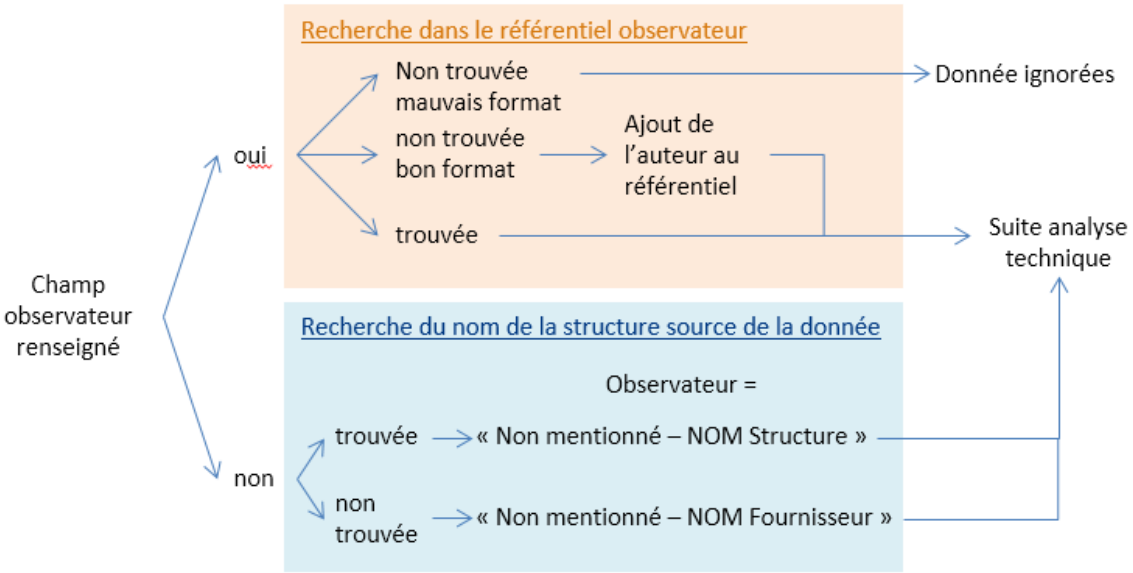

FIGURE 4 : SCHEMA GENERAL D'ANALYSE DE L'OBSERVATEUR

d. Date (mise en conformité et contrôle de cohérence)

Date	
Champ obligatoire	OUI
Elément(s) retenu(s)	Date au format JJ/MM/AAAA ou MM/AAAA ou AAAA

La date doit comporter au moins l'année d'observation.

Problème	Conséquence
Date postérieure à la date de réception	Donnée rejetée
Date absente	Donnée rejetée
Date erronée	Donnée rejetée

e. Précision de date (mise en conformité)

Précision de date	
Champ obligatoire	NON
Elément(s) retenu(s)	Texte : Précise, + ou – 1 jour, + ou – 3 jours, + ou – 1 semaine, + ou – 2 semaines, + ou – 3 mois, + ou - 6 mois

Problème	Conséquence
Si Precdate non rempli ou pas de champ	Analyse automatique de la date (si jour, mois et année sont présent : Date précise
Si date à l'année	01 juillet + ou – 6 mois
Si date au mois	15 du mois + ou – 2 semaines

f. Commune (mise en conformité)

Commune	
Champ obligatoire	OUI (NON si code INSEE)
Elément(s) retenu(s)	Nom de la commune

Une vérification est effectuée sur le nom de la commune. S'il y a une faute d'orthographe dans le nom de la commune, le script recherche le bon nom dans un référentiel de noms de communes erronés.

Problème	Conséquence
Commune inconnue	Recherche de la bonne commune, si pas de résultat : donnée rejetée
Commune absente	Donnée rejetée (si pas de code INSEE)
Commune hors PACA	Donnée rejetée

g. Code INSEE (mise en conformité et contrôle de cohérence)

Code INSEE	
Champ obligatoire	NON (OUI si pas de commune)
Elément(s) retenu(s)	Nombre

Le code INSEE est utilisé lorsque la commune est absente. Une vérification de cohérence est réalisée entre la commune et le code INSEE lorsque ces deux champs sont remplis. Le référentiel utilisé aujourd'hui est le ref INSEE 2020 et non le COG (Code Officiel Géographique) qui est le standard INPN.

La fusion de certaines communes devant entrainer une perte de précision pour la donnée communale, un travail spécifique a été apporté :

Pour les communes ayant fusionnées :

1. Sur le référentiel des lieux-dits :
    a. changer les nom de lieu-dit par « Nom ancienne commune – Nom Lieu-dit »
    b. créer autant de lieu-dit que d'ancienne commune nommé « Nom ancienne commune – centroïde »
2. Sur les données historiques (script en phase de test) :
    a. Communales sans lieu-dit dans le champ lieu-dit => champ commune : nouvelle commune / champ lieu-dit : « Nom ancienne commune – centroïde » / champs précision : lieu-dit
    b. Communales avec lieu-dit non reconnu dans le champ lieu-dit => champ commune : nouvelle commune / champ lieu-dit : « Nom ancienne commune – nom du lieu-dit cité » / champs précision : lieu-dit
    c. Lieu- dit reconnu => « Nom ancienne commune – Nom Lieu-dit ».

Grace à des tableaux d'équivalences, lorsque les communes renseignées ne sont pas celles du référentiel actuel, le nom de l'ancienne commune est identifié comme le lieu-dit et le nom de la nouvelle commune remplacé par le nouveau nom.

NB : dans GeoNature, les coordonnées finales de l'objet (polygone, ligne ou point) déterminent la commune.
NB : il conviendra de passer à une spatialisation de la donnée d'occurrence et non plus d'avoir un code INSEE en base de données. Ceci permettra de mettre une zone de présence pour une donnée communale qui sera la commune en date d'observation et de traiter également des données à la maille.

h. Coordonnées géographiques (mise en conformité et contrôle de cohérence)

Coordonnées géographiques	
Champ obligatoire	NON
Elément(s) retenu(s)	Nombre positif

Le système de coordonnées de Silene est le Lambert 93. Cependant, le script gère tous les systèmes de coordonnées (latitude/longitude en degré décimal ou en degré minute seconde, Lambert II étendu, UTM 31 et 32). Chaque système est reconnu et est transformé en Lambert 93.

Deux tests de cohérence géographique sont effectués successivement :

- les valeurs X et Y sont situées en région PACA ;
- la commune indiquée dans le fichier est la même que la commune obtenue par requête géographique (les communes limitrophes sont acceptées).

Les données présentant une incohérence sont analysées par l'opérateur afin d'identifier la source d'erreur (mauvaise conversion, inversion champs X/Y…) et sont corrigées le cas échéant.

i. Précision de localisation (mise en conformité et contrôle de cohérence)

Précision de localisation	
Champ obligatoire	OUI si présence de coordonnées géographiques
Elément(s) retenu(s)	Indéterminé=absent / Commune (C) / Lieu-dit (T) / Précis (P)

Lorsque que des coordonnées géographiques sont présentes, il est indispensable de connaître leurs niveaux de précision car elles peuvent correspondre au centroïde d'une commune, d'un lieu-dit ou à un pointage précis :

- si le champ est Non renseigné ou absent => la précision est donc supposée « indéterminée » ;
- s'il n'y a pas de coordonnée mais un lieu-dit renseigné et reconnu => la précision sera au lieu-dit ;
- s'il n'y a pas de coordonnée mais un lieu-dit non reconnu => la précision sera à la commune ;
- s'il n'y a pas de coordonnée mais une commune => la précision sera à la commune.

j. Lieu-dit (mise en conformité et contrôle de cohérence)

Lieu-dit	
Champ obligatoire	NON
Elément(s) retenu(s)	Texte libre

La précision de l'observation sera définie au lieu-dit pour :

- le champ lieu-dit est renseigné et reconnu dans le référentiel géoréférencé (BDnyme) mais la donnée n'a pas de coordonnées géographiques, les coordonnées du lieu-dit sont récupérées ;
- une donnée identifiée en précision communale mais possédant également un champ lieu-dit renseigné et reconnu dans le référentiel géoréférencé (BDnyme).

Si le lieu-dit n'est pas renseigné dans le référentiel, les coordonnées prises sont celles du centre de la commune (cf. § 3.f. : fusion des communes). La précision de l'observation est alors communale. La BDnyme s'améliorant en continu, de plus en plus de données communales peuvent être précisées au lieu-dit.

k. Département (mise en conformité et contrôle de cohérence)

Département	
Champ obligatoire	NON
Elément(s) retenu(s)	Nom du département

Le département est analysé afin de vérifier la cohérence entre la commune et le département d'observation. S'il n'y a pas de département renseigné et que la commune n'a pas de synonyme dans un autre département, alors le département est récupéré à partir du référentiel INSEE.

l. Nombre (mise en conformité et contrôle de cohérence)

Nombre	
Champ obligatoire	NON
Elément(s) retenu(s)	Nombre entier

Lorsque le champ nombre est rempli avec un 0, le script transforme le champ en champ vide.

Pour certains fichiers, notamment issus de protocole de suivi d'espèce, le 0 signifie qu'il s'agit d'une donnée négative, c'est-à-dire que l'espèce a été recherchée mais n'a pas été contactée.

Il existe deux manières de traiter la donnée négative, toutes deux mises en place :

- pour un fichier de suivi où les 0 correspondent à des données négatives, on remplace le 0 par « Absence » et ces données seront rejetées lors des scripts d'intégrations ;
- lorsque les données 0 correspondent à un nombre « indéterminé » (trace de présence par exemple) ou que l'information est non vérifiable, on vide la case et la donnée apparaîtra en base comme « NULL ».

m. Statut de la donnée (publique/privée)

Statut de la donnée = DSPubliqueValue	
Champ obligatoire	OUI
Elément(s) retenu(s)	Pr (Privé), Pu (Publique) et NSP (ne sait pas)

Depuis 2021, les fichiers réceptionnés doivent renseigner ce nouveau champ obligatoire (via la plateforme de dépôt ou dans le jeu de données).

Si pas renseigné → NSP

Si erreur de saisie → NSP

Les jeux de données historiques seront tagués par défaut dans Silene en NSP sauf si le fournisseur donne l'information Pr/Pu/NSP dans son jeu de données ou par jeu de données.

Le taguage se fait à la donnée et non au jeu de données (sauf mention explicite que tout le jeu de données a le même statut).

n. Autorisation de diffusion (mise en conformité)

Autorisation de diffusion = NiveauPrecisionValue	
Champ obligatoire	OUI
Elément(s) retenu(s)	Nombre entier : 1 à 5

En fonction du statut de la donnée et d'information sur le niveau de diffusion autorisé :

Si DSPubliqueValue =	Alors NiveauPrecisionValue =
NSP	1
Pu	5
Pr – sans information ou avec limite de diffusion	1
Pr – avec autorisation de diffusion à la précision maximale	5

Phase 4 : Gestion des doublons et traitement différentiel

a. Détection des doublons

Les doublons de données correspondent à une donnée source transmise deux fois en raison :

- d'une erreur humaine (double saisie, duplication de ligne), les doublons sont alors présents dans la même base et reçus dans le même fichier ;
- d'échanges et de transmissions de données à travers plusieurs bases, inévitablement accompagnés de mises en forme propres à chaque base et d'une éventuelle dégradation de la précision de l'information (localisation, auteur, date…). Les doublons sont issus alors de différents fichiers sources.

Les doublons créent un problème de représentation du nombre d'observations d'une espèce et faussent diverses analyses écologiques et biologiques (suivi d'espèce, détectabilité, répartition fine…).

La détection des doublons se fait sur la base de contrôle de redondance à partir des attributs relatifs à la date de l'observation, au lieu, au taxon ou encore à l'observateur.

Plusieurs types de détection des doublons sont mis en place :

- recherche des doublons pour un même fournisseur, en se basant sur les critères suivants :

  mêmes CDREF, Observateur - ou Non mentionné -, Date, Code INSEE, Lieu-dit et/ou coordonnées géographiques identiques à 5 mètres près ;
- recherche des doublons entre le fichier source et Silene, en se basant sur les critères suivants :

  mêmes CDREF, Observateur- ou Non mentionné -, Date, Code INSEE, et/ou coordonnées géographiques identiques à 300 mètres près.

b. Traitement des doublons

Le traitement a lieu chaque semaine et après l'étape de validation. Il est constitué de cinq étapes :

i. Données à la précision communale

Sont recherchées les données communales en doublons selon les critères suivants : même CD_REF, Code INSEE, Premier observateur et Date d'observation.

On conserve la donnée selon le fournisseur en privilégiant :

1/ Autres fournisseurs

2/ DREAL PACA - N2000

3/ DDT05

4/ MNHN.

Si les fournisseurs sont identiques ou des fournisseurs identifiés autres que la DREAL PACA, DDT05 et MNHN, on garde la donnée la plus anciennement transmise.

ii. Données au Lieu-dit

Sont recherchées les données au lieu-dit en doublons selon les critères suivants : même CD_REF, Code INSEE, Premier observateur, Date d'observation et lieu-dit.

On conserve la donnée selon le fournisseur en privilégiant :

1/Autres fournisseurs

2/ DREAL PACA -N2000

3/ DDT05

4/ MNHN.

Si les fournisseurs sont identiques ou des fournisseurs identifiés autres que la DREAL PACA, DDT05 et MNHN, on garde la donnée la plus anciennement transmise.

iii. Identification des doublons d'un fournisseur identique – critère distance

On recherche les données précises en doublons selon les critères suivants : même CD_REF, Code INSEE, Premier observateur, Date d'observation, pour lesquels les coordonnées sont situées dans un rayon de 5 m. On garde la donnée la plus anciennement transmise.

iv. Identification des doublons de fournisseurs différents – critère distance

On recherche les données précises ou au lieu-dit en doublons selon les critères suivants : même CD_REF, Code INSEE, Premier observateur, Date d'observation, pour lesquels les coordonnées sont situées dans un rayon de 300 m.

On conserve la donnée selon le fournisseur en privilégiant :

1/Autres fournisseurs

2/ DREAL PACA -N2000

3/ DDT05

4/ MNHN.

La donnée la plus anciennement transmise est conservée.

v. Données Communales vs données au Lieu-dit ou Précises

Sont éjectées également pour cause de doublon toutes les données communales pour lesquelles il existe des données ayant les mêmes CDREF/Codes INSEE/Observateur1/Date avec un niveau de précision de localisation plus important (lieu-dit ou précis).

vi. Cas particulier des données INPN vs autres fournisseurs :

Toutes données INPN communales pour lesquelles il existe des données ayant les mêmes CDREF/Codes INSEE/Date issues d'un autre fournisseur seront rejetées.

c. Intégration différentielle

La méthode intégrale d'intégration (annule et remplace) n'est plus utilisée, c'est la Méthode différentielle qui est mise en place : le jeu de données intégrés ne contient que les nouvelles occurrences ou leurs mises à jour, l'ensemble des informations de ces occurrences sont implémentées en base de données.

Deux cas de figure existent :

- le fournisseur n'envoie que les nouvelles données et données à modifier dont les suppressions ;
- le fournisseur envoie l'ensemble de sa base et c'est le traitement des données qui permet de faire le différentiel.

La gestion des doublons « Recherche des doublons pour un même fournisseur » est un comparatif entre ce qu'il existe déjà en base et ce qui est nouveau (nouvelles données) et qui donc fait partie de l'intégration différentielle.

Phase 5 : Qualification et validation des données

La validation des données fait l'objet d'un rapport spécifique : cf. Delauge J., Kapfer, G., Honoré, P., Guillaud, Fanny, 2021. Protocole de validation des données naturalistes faunistiques dans Silene. Conservatoire d'espaces naturels de Provence-Alpes-Côte d'Azur – MAJ 2021. Sisteron, 13 p.

Ci-dessous un schéma représente la qualification automatique de la donnée et sa validation manuelle. Le principe consiste à réaliser une série de test de cohérence scientifique sur les connaissances disponibles actuellement en fonction de groupes taxonomique comme la répartition et la période d'observation par exemple.

La validation manuelle peut parfois être réalisée par un ou plusieurs experts avant l'envoi à l'informaticien. Dans ce cas, le fichier ne passera par une qualification automatique.

A l'issue de cette qualification automatique les données sont « Retenues » ou « A valider ». Les données « A Valider » sont soumises à une validation manuelle par les experts. Elles sont ensuite qualifiées en « Validée », "Retenue", « A confirmer » ou « Douteuse ». Seules les données « Validées » et « Retenues » sont mises à disposition dans Silene.

Les données « A confirmer » et « Douteuse » font partie des données rejetées.

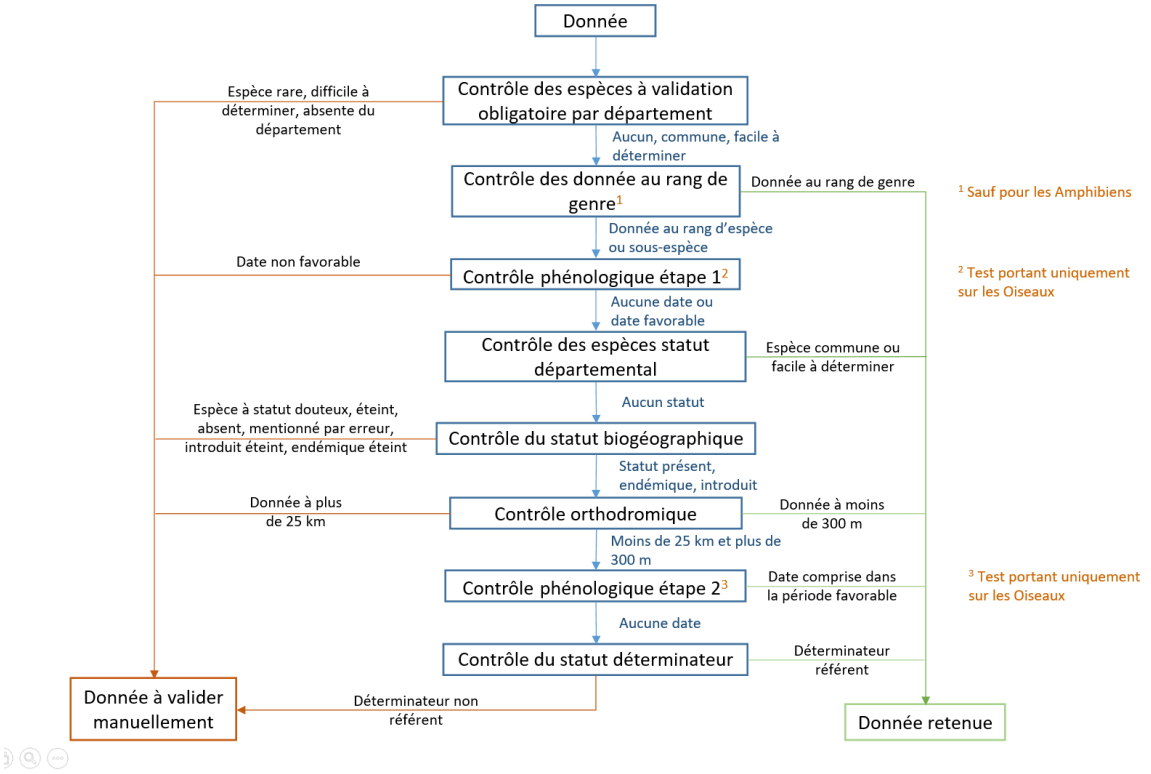

FIGURE 5 : SCHEMA SIMPLIFIE DE LA QUALIFICATION AUTOMATIQUE ET LA VALIDATION MANUELLE DE LA DONNEE

Phase 6 : Retour sur l'intégration dans SILENE

a. Création d'un fichier de données rejetées

A l'issue de l'analyse technique, un fichier des données rejetées, appelé « données rejetées » est produit. Il contient toutes les données rejetées par le script. Ces fichiers font l'objet d'une vérification manuelle afin de voir si des données peuvent être récupérées.

Ce fichier contient tous les champs du fichier au format Silene dont le champ ID_PERSO mentionnant l'identifiant dans la base du fournisseur. Un champ supplémentaire explicitant le type de rejet est créé. Les causes de rejet peuvent être :

- boublon avec Silene ;
- doublon dans le fichier ;
- donnée d'absence ;
- espèce inconnue ;
- donnée invalide / en attente (à confirmer ou douteuse) ;
- problème de dates ;
- problème de localisation ;
- donnée hors PACA.

b. Finalisation de la métadonnée

A la fin des analyses, il est produit un fichier synthétique regroupant les retours des analyses techniques et scientifiques et le retour d'intégration. Le fichier reprend les résultats des différents process de mise en conformité et de validation, à savoir :

- Pour l'analyse technique :
- le nombre total de données retenues et rejetées par cette analyse ;
- le nombre de données rejetées par cause de rejet (doublon, espèce, observateur, date, localisation, données négatives).

- Pour l'analyse scientifique :
- le protocole de validation (validation manuelle, automatique ou les deux) ;
- le nombre de données retenues et rejetées par cette analyse.

- Pour le retour d'intégration dans Silene :
- le statut des données (Pu/Pr/NSP) ;
- le nombre de données intégrées ;
- la date d'intégration ;
- la précision de localisation (précis, lieu-dit, commune) indiquée en nombre de données ;
- le nombre de données anonymes garanties par le fournisseur.

La Figure 6 présente la fiche de métadonnée explicative.

Retour de l'analyse technique		Libellé	Remarque
Nombre de données retenues			
Nombre de données rejetées			
Causes de rejets	Doublons (avec SILENE)		
	Problèmes sur le nom d'espèce		
	Problèmes sur le nom d'observateur		
	Problèmes sur la date		
	Problèmes sur la localisation		
	Données négatives		
	Données en attente		
	Observations hors PACA		
	Autres		

Retour de la validation scientifique	Libellé	Remarque
Protocole de validation		
Nombre de données retenues		
Nombre de données rejetées		

Retour d'intégration		Libellé	Remarque
Nombre de données intégrées dans SILENE			
Date d'intégration des données dans SILENE			
Précision de localisation	Commune		
	Lieu-dit		
	Point précis		
Nombre de données anonymes			

FIGURE 6 : METADONNEE EXPLICATIVE DES DONNEES INTEGREES D'UN FICHIER

Ces informations sont inclues dans les statistiques Silene et dans le catalogue des données Faune, réalisé à chaque mise à jour des données Silene.

c. Retour au fournisseur de données

Pour chaque fichier intégré, un rapport est renvoyé au fournisseur de données. Ce rapport comprend :

- les métadonnées du fichier ;
- le fichier de données rejetées (avec la cause de rejet) si le fournisseur en fait la demande.

Les producteurs de données ont ainsi un retour sur leurs données non intégrées et les causes de rejet. Les producteurs de données sont incités à travailler ces données afin de les renvoyer à l'administrateur de Silene pour intégration.

Conclusion

Bien qu'il soit observé un nombre croissant de producteurs de données et de fichiers transmis, l'intégration de la donnée se doit de poursuivre le processus précis mis en place afin de porter à connaissance une donnée techniquement et scientifiquement valide.

Le traitement des jeux de données s'accélère depuis 2019 avec l'augmentation de la ressource humaine affectée à l'administration de données grâce à une augmentation de la subvention de la DREAL PACA.

Conservatoire d'espaces naturels Provence-Alpes-Côte d'Azur

Siège :
4, avenue Marcel Pagnol
Immeuble Atrium Bât B.
13 100 Aix-en-Provence
Tél : 04 42 20 03 83
Fax : 04 42 20 05 98
Email : contact@cen-paca.org
www.cen-paca.org

Pôle Biodiversité Régionale
18 avenue du Gand
04200 SISTERON
Tél : 04 92 34 40 10

Le Conservatoire d'espaces naturels
de Provence-Alpes-Côte d'Azur
est membre de la Fédération
des Conservatoires d'espaces naturels

Conservatoires d'espaces naturels

Ce travail a été réalisé grâce au soutien financier des partenaires suivants :

agence de l'eau RHÔNE MÉDITERRANÉE CORSE établissement public de l'État

RÉGION SUD PROVENCE ALPES CÔTE D'AZUR

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Direction régionale de l'Environnement, de l'Aménagement et du Logement