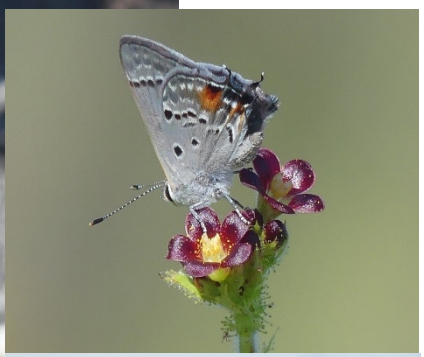
Bivouac Naturaliste

2022

PRÉFET DE LA RÉGION GUADELOUPE
Liberté
Égalité
Fraternité

Parc national de la Guadeloupe

Protocole de validation des données naturalistes dans Karunati



KARUNATI
PORTAIL SINP GUADELOUPE

GEO Nature

Protocole de validation des données naturalistes dans Karunati

Document réalisé par :

Association Bivouac Naturaliste

Coordination :

Marion GESSNER - Chargée de mission données environnementales, Pôle Biodiversité/Service Ressources naturelles, DEAL Guadeloupe

Rédaction :

Benjamin FERLAY – Bivouac Naturaliste

Marion GESSNER - Chargée de mission données environnementales, Pôle Biodiversité/Service Ressources naturelles, DEAL Guadeloupe

Relecture :

Nils SERVIENTIS – Bivouac Naturaliste

Alice ARMAND – Bivouac Naturaliste

Marion GESSNER - Chargée de mission données environnementales Pôle Biodiversité/Service Ressources naturelles, DEAL Guadeloupe

Alain FERCHAL – Chef de service, Service Système d'Information, PN de la Guadeloupe

Sarah LE CŒUR – Expert indépendant

Toni JOURDAN - Expert indépendant

Laurent MALGLAIVE - Expert indépendant

Date de réalisation : Juin 2022

Citation recommandée : Gessner M., Ferlay B., 2022. Protocole de validation des données naturalistes dans Karunati. Bivouac Naturaliste – Juin 2022.

Table des matières

Préambule

A travers plusieurs applications, les professionnels de l'environnement ainsi que les amateurs naturalistes ont la possibilité de transmettre de l'information sous forme de données d'observation ou données d'occurrence de taxon, qu'elles soient floristiques ou faunistiques. Que ce soit via le masque de saisie, avec l'application Géonature (saisie web) ou Occtax (saisie mobile), toutes ces données se retrouvent sur la plateforme de visualisation KARUNATI. Grâce à cette plateforme, le grand public peut visualiser la répartition des espèces à l'échelle de la Guadeloupe. Bien que le réseau des contributeurs soit en cours de développement pour ce SINP régional, il est indispensable d'accorder de l'importance à la qualité de la donnée transmise. De cette manière, une validation des données transmises est nécessaire pour renseigner la fiabilité de la donnée, en lui donnant un degré de confiance. La validation de la donnée s'effectue à trois niveaux :

a) un contrôle de conformité (respect des règles du standard) ;

b) un contrôle de cohérence logique (respect de la logique des données : date de début d'observation inférieure ou égale à la date de fin d'observation) ;

c) une validation scientifique (niveau de fiabilité de la donnée). Cette validation scientifique peut être automatique (informatique) et manuelle (intervention d'un expert).

Les processus de validation automatique font intervenir des bases de connaissance sur les données (référentiels taxonomiques, répartition géographique…). Les processus de validation manuelle font appel à de l'expertise directe à travers des traitements manuels (avis d'expert suite à l'analyse des informations transmises). Dans ce dernier cas, plusieurs experts de différents groupes taxonomiques sont désignés pour assurer ce contrôle qualité manuel des données dans le cadre du SINP.

Ce présent document a pour but de détailler le protocole de validation scientifique des données d'occurrence de taxon faune et flore appliqué par le SINP Guadeloupe, c'est à dire pour une donnée qui a validé le contrôle de conformité (a) et le contrôle de cohérence logique (b).

,

Objectif et principe du protocole de validation

La validation des données ne s'effectue que sur de la donnée conforme (validation de conformité et de cohérence logique validés). Cela correspond à une donnée comprenant un minimum d'informations indispensables à son intégration, informations reconnues et standardisées : un taxon, un observateur, une localisation et une date.

Pour toutes les données conformes aux standards, une question se pose sur la cohérence scientifique de cette donnée, et vient donc les différentes étapes de validation de ces données, en fonction de sa fiabilité. Est-il cohérent de retrouver cette plante en Guadeloupe, aux Saintes ou à la Désirade ? Est-ce cohérent d'observer cette espèce d'oiseau à cette date ? Cette observation faite par telle personne nécessite-t-elle une validation de la donnée manuelle ? Ainsi, toutes les données d'observations suivent ce traitement, en prenant uniquement en compte ces quatre critères (taxon, lieu, date, observateur).

Dans un premier temps, les données subissent un traitement automatique en suivant ces différents critères et sont ensuite qualifiées. Ces traitements automatiques sont d'autant plus efficaces que le jeu de donnée est important, un grand nombre de données permettant d'affiner cette première qualification automatique. En résumé, cette qualification a pour vocation d'évoluer au cours du temps et de s'améliorer. A l'issu de ce premier traitement, la donnée peut être qualifiée de probable et est validée informatiquement par défaut (ce qui n'en fait pas une donnée valide) ou être qualifiée de donnée à confirmer. Les données à confirmer vont donc subir un deuxième traitement, cette fois-ci manuel.

Dans KARUNATI, la validation scientifique manuelle consiste à attribuer à la donnée différents statuts :

- Validée[1]
- Probable[2]
- Douteux[3]
- Invalide[4]
- Non réalisable[5]

[1] Les informations apportées ne permettent aucun doute
[2] Donnée retenue mais non valide, les informations sont cohérentes d'un point de vue taxonomique, écologique, biogéographique, mais il manque une information pour en faire une donnée valide
[3] Donnée à confirmer taxonomiquement par un autre expert en cas de groupe taxonomique complexe, ou d'une répartition exceptionnelle
[4] Donnée sans preuves suffisantes ou avec des preuves réfutant l'observation
[5] Donnée non conforme

En fonction des statuts attribués, la donnée sera diffusée dans KARUNATI ou non. De cette manière, les données considérées comme douteuses et invalides ne seront pas transmises.

De plus, afin d'assurer un suivi et éviter les doubles évaluations, les données non évaluées ont également un statut « à valider » qui est attribué automatiquement.

Processus de validation scientifique automatique

Comme évoqué précédemment, la validation automatique fait appel à des bases de données existantes, des référentiels taxonomiques régionaux ainsi que des modèles probabilistes.

Contrôle de reconnaissance par TAXREF

Contrôle de présence et du statut biogéographique

La taxonomie utilisée dans KARUNATI correspond au référentiel taxonomique national le plus récent. Ainsi, les noms scientifiques valides accompagnés du CD_NOM sont interrogés, en prenant en compte le statut biogéographique du taxon.

De cette manière, si le CD_NOM du taxon observé existe dans la version de TAXREF en vigueur, et étant annoncé comme présent[1] sur le territoire de la Guadeloupe, alors le résultat du contrôle est positif.

A l'inverse, si le CD_NOM du taxon observé n'existe pas dans la version de TAXREF en vigueur, ou existe mais étant signalé comme absent[2] du territoire de la Guadeloupe, alors le résultat du contrôle est négatif et la donnée est considérée comme invalide au niveau de la validation intermédiaire. La donnée rentrera dans un processus de traitement manuelle[4] pour confirmer ou infirmer la statut invalide de cette donnée.

En revanche, si le CD_NOM du taxon observé existe dans la version de TAXREF en vigueur avec un statut biogéographique incertain[3] ou que l'espèce n'est plus observée[3] sur le territoire de la Guadeloupe, alors le résultat du contrôle est douteux et demandera directement un traitement manuel[4].

[1]Une espèce est considéré comme présente pour les statuts biogéographiques suivant : Présent (indigène ou indéterminé), Occasionnel, Endémique, Subendémique, Cryptogène, Introduit, Introduit envahissant et Introduit non établi (dont domestique et cultivé).

[2]Une espèce est considérée comme absente pour les statuts biogéographiques suivant : Absent et Mentionné par erreur.

[3]Cela concerne les espèces dont le statut biogéographique est le suivant : Douteux, Disparu, Éteint, Introduit éteint / disparu et Endémique éteint.

[4]Cela permet de ne pas négliger une observation exceptionnelle d'une espèce indigène ou exotique non observée depuis longtemps, ou dont le statut est douteux. Ces cas sont relativement fréquents sur des territoires comme la Guadeloupe où demeurent d'importantes lacunes sur la connaissance.

Contrôle de l'habitat

Il n'existe actuellement pas de référentiel d'habitat d'espèces pour la Guadeloupe, sans cartographie régionale précise de ces habitats naturels et avec d'importantes lacunes sur l'écologie des différentes espèces présentes. Un contrôle d'habitat à échelle plus large peut cependant être mis en place, en prenant en compte l'appartenance au domaine marin et continental d'un taxon. Le but est alors de vérifier que l'habitat associé au taxon[1] soit cohérent avec la situation géographique de la maille 1 km² dans laquelle le taxon a été observé. Chacune de ces mailles porte un attribut concernant le domaine écologique.

Ainsi, une maille peut être continentale (elle appartient à 100% au domaine continental) ou marine (elle appartient à 100% au domaine marin). Naturellement, et particulièrement sur un système insulaire, plusieurs mailles peuvent être à la fois continentale et marine.

De cette manière, si la maille est strictement continentale et que l'habitat du taxon est cohérent avec ce domaine (Eau Douce, Terrestre, Marin & Eau douce, Marin & Terrestre, Eau saumâtre et Continental), alors le résultat du contrôle est positif.

En revanche, si la maille est strictement continentale et que l'habitat du taxon est Marin, lors le résultat du contrôle est négatif.

Aussi, si la maille est strictement marine et que l'habitat du taxon est Marin, Marin & Eau douce, Marin & Terrestre et Eau saumâtre, alors le résultat du contrôle est positif.

En revanche, si la maille est strictement marine et que l'habitat du taxon est Eau Douce, Terrestre ou Continental, alors le résultat du contrôle est négatif.

Dans le cas des mailles mixtes, aucun contrôle n'est effectué et la mention non applicable est attribué à l'observation.

[1]Il existe 8 grands types d'habitats pouvant être attribués aux différents taxons du TAXREF : Marin, Eau Douce, Terrestre, Marin & Eau douce, Marin & Terrestre, Eau saumâtre, Continental (terrestre et/ou eau douce) et Continental (terrestre et eau douce).

Contrôle de la période d'observation

Ce contrôle concerne uniquement la faune pour certains groupes taxonomiques, principalement l'avifaune et autres groupes où la périodicité a un impact direct sur les observations. Ainsi, pour certains taxons, il existe des périodes d'observations favorables et défavorables (cas des migrations). L'objectif est donc de vérifier la cohérence d'une observation d'un taxon avec les dates d'observations favorables de ce même taxon.

Pour réaliser ce contrôle, il est nécessaire de créer un référentiel des périodes d'observations de la faune.

De cette manière, pour un taxon observable à tout moment de l'année, le résultat du contrôle est positif.

Pour les taxons étant observables pendant une ou plusieurs périodes de l'année en Guadeloupe, plusieurs niveaux de fiabilité peuvent être attribués à une observation :

- **Fiable** : Le taxon est observé pendant la période d'observation favorable.
- **Possible** : Le taxon est observé pendant la période d'observation favorable mais peu fréquente ou très rare.
- **Non fiable** : Le taxon est observé hors de la ou des périodes d'observations favorables.
- **Non applicable** : La date d'observation du taxon est imprécise.

Contrôle de l'identification du taxon

Alors que certaines espèces sont relativement simples à identifier, sans laisser de doute quant à leur identification, certains taxons ou groupes taxonomiques sont complexes à identifier, souvent pour des raisons de forte ressemblance, de critères d'identifications discriminant difficiles à appréhender ou encore du manque de documents de références permettant la détermination. L'identification peut donc être une source d'erreur importante dans la donnée. Ce contrôle nécessite cependant de disposer d'un référentiel des niveaux de difficulté de détermination des espèces présentes en Guadeloupe.

Plusieurs niveaux de difficulté devront être retenu à la création de ce référentiel :

- **Facile** : Taxon facilement identifiable, que ce soit à vue, sur photo, au chant ou par un indice de présence fiable. Il existe peu ou pas de risques de confusion. Plusieurs documents de références permettant une identification certaine sont disponibles (portefolio, articles, guides…).

- **Intermédiaire** : Taxon identifiable à vue, sur photo, au chant ou par un indice de présence fiable mais demandant davantage de compétences ou demandant d'avantage d'attention. Il existe des taxons proches rendant un risque de confusion possible mais faible. Quelques documents de références sont disponibles mais plus difficilement accessibles (littérature en anglais, ouvrages plus édités…).

- **Difficile** : Taxon nécessitant une observation attentive, avec une capture ou un prélèvement. Plusieurs taxons proches ou très proches rendant le risque de confusion élevé. Il existe très peu de documents de références disponibles.

- **Très difficile** : Taxon complexe, nécessitant une capture ou une récolte ainsi que du matériel adapté (binoculaire, microscope, appareil biométrique…) ou une analyse en laboratoire pour pouvoir assurer la détermination. Le risque de confusion avec d'autres taxons est très élevé. Les documents de références sont très rares, incomplets ou inexistants.

Attribution de la note finale au contrôle automatique

La combinaison des résultats des différents contrôles développés précédemment donne lieu à une traduction en termes de niveau de validité. Ainsi, chaque observation se voit attribuer automatiquement un niveau de validité qui détermine le degré de confiance que l'on peut lui accorder.

De cette manière 6 différents niveaux sont utilisés :

- **Très probable** : La donnée rassemble tous les résultats de contrôle positifs et de critères favorables.

- **Probable** : La donnée rassemble une majorité de résultats de contrôles positifs et de critères favorables, la rendant probable mais ne satisfaisant pas intégralement l'ensemble des critères automatiques. Il n'y a pas de discordance majeure et elle est satisfaisante au niveau intermédiaire.

- **Douteux** : La donnée concorde peu selon les différents contrôles automatiques appliqués. Elle est peu cohérente, ne satisfaisant pas ou peu de critères automatiques appliqués, sans pour autant présenter de discordances majeures sur les critères jugés les plus importants.

- **Invalide** : La donnée ne concorde pas selon différents contrôles automatiques appliqués. Elle n'est pas cohérente, ne satisfait pas les critères automatiques appliqués et présente des discordances majeures sur les critères jugés les plus importants.

- **Non réalisable** : La donnée peut être valide ou non mais les informations sont insuffisantes pour pouvoir statuer sur le niveau de fiabilité.

- **Non évalué** : La donnée n'a pas encore été soumise à l'opération ou l'opération n'est pas terminée.

Certains contrôles ont plus d'importances que d'autres dans l'attribution du niveau de fiabilité. Ils sont donc hiérarchisés et ordonnés de manière à faire passer les contrôles les plus importants en premier plan, pour passer aux contrôles secondaires dans le cas où le premier contrôle est validé.

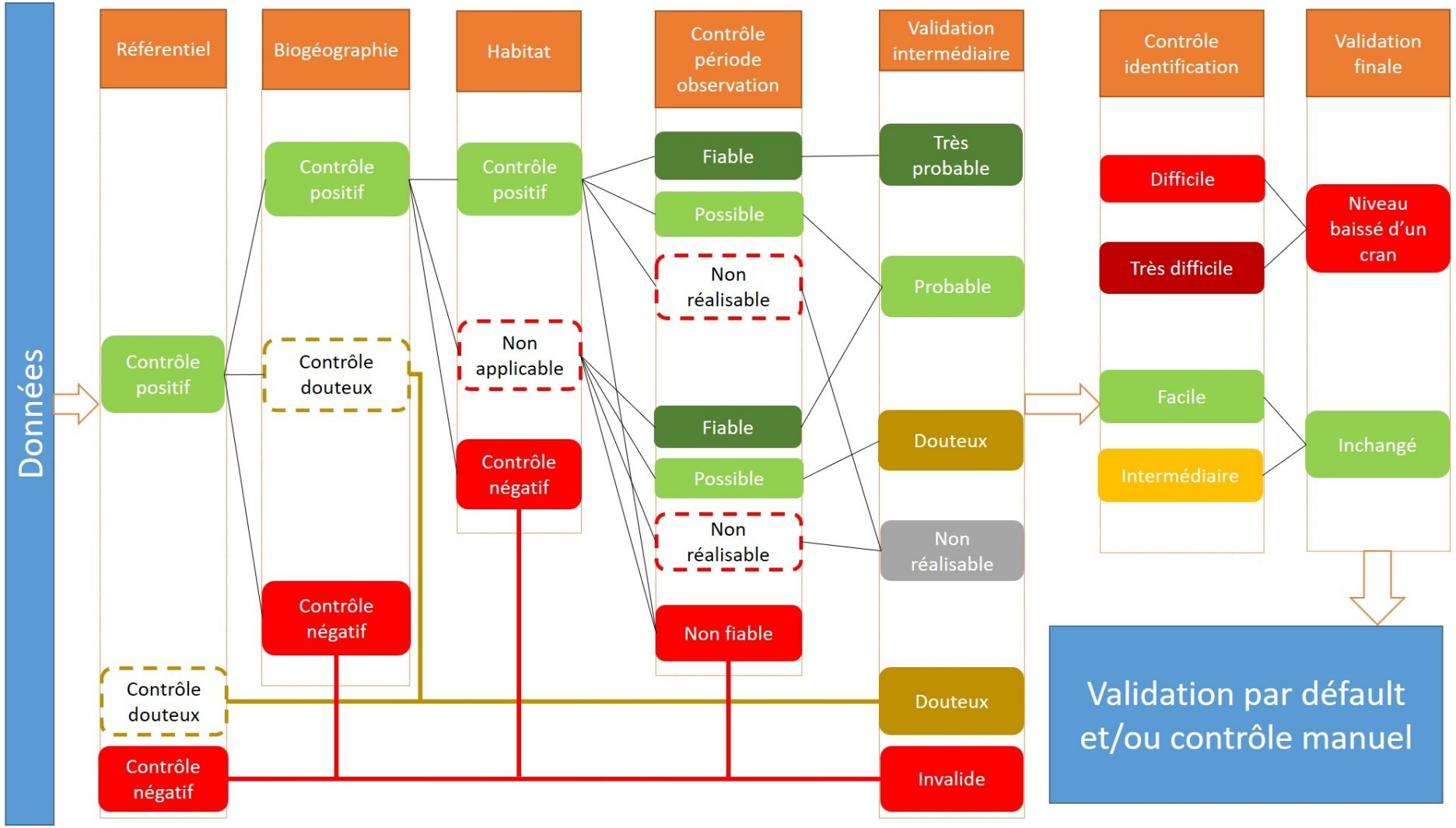


Figure 1 Processus de contrôle automatique pour aboutir à la note finale

Processus de validation manuel

Après le processus de contrôle automatique, une validation manuelle peut être réalisée sur certaines données. Cette validation va permettre de confirmer ou modifier le niveau de validation attribué par le processus automatique, au cas par cas. Ce contrôle au cas par cas permet d'étudier et de donner une note à des données qui n'ont pas pu être traitées dans le processus automatique (« Non réalisable »).

Cette étape de validation manuelle est réalisée par un regroupement d'experts, composés de membres spécialistes d'un ou plusieurs groupes taxonomiques. A travers des interfaces propres aux administrateurs, il leur est possible de contrôler et de corriger la note finale résultant de la validation automatique des données. Ces experts ont accès aux différents résultats de chacun des critères du processus de validation automatique et aux bases de connaissances disponibles. Ils ont également accès aux contacts des différents contributeurs afin d'obtenir des informations complémentaires sur l'observation comme des détails ou des preuves photographiques, afin de qualifier cette donnée de valide ou d'invalide. Ils ne peuvent en revanche pas corriger ni modifier les données brutes sans avoir l'autorisation de l'observateur.

Les experts se concentreront en priorité sur les données non traitées par le processus automatique (« Non réalisable »), ainsi que sur les données douteuses et les groupes taxonomiques complexes.

De cette manière, les validateurs se doivent de regarder et d'étudier la probabilité de présence d'un taxon au lieu et à la date mentionnés en fonction de sa propre connaissance du taxon (taxinomie, répartition, biologie…). Ils peuvent redéterminer un taxon si une photo est disponible ou s'ils obtiennent des compléments de l'observateur. Ils étudieront également les profils des observateurs afin de connaitre au mieux les observateurs fiables et les observateurs plus amateurs.

La validation manuelle fait foi par rapport à la validation automatique.

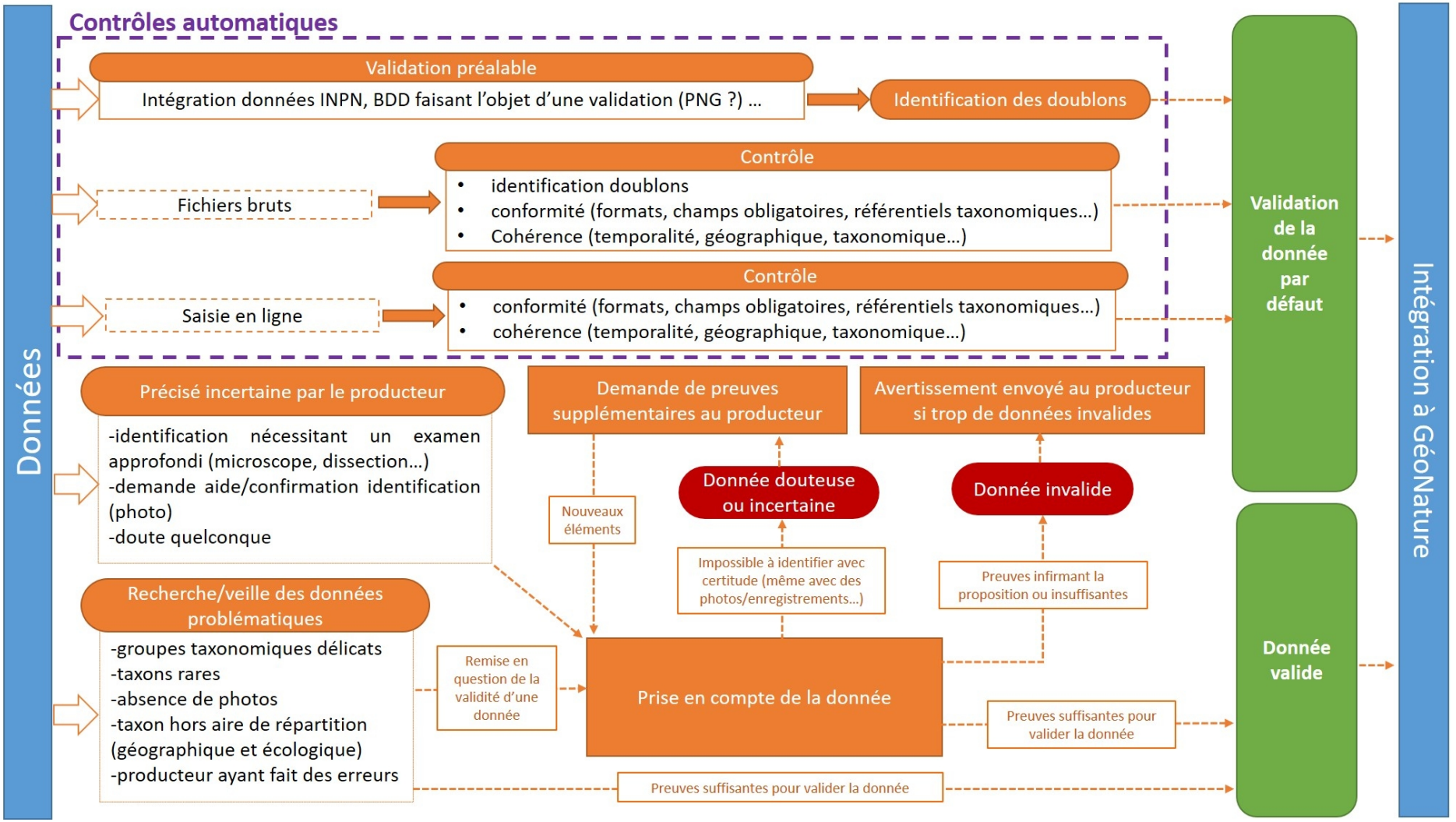


Figure 2 Synthèse schématique du protocole de validation des données naturalistes

Validation régionale, échange SINP et diffusion INPN

La finalité de tout le processus est la validation des données à l'échelle régionale. En fonction de niveau de validité de la donnée, et en suivant la procédure nationale (Robert *et al.* 2017), les données peuvent être échangées entre les plateformes SINP et transmises à l'INPN.

Pour les échanges entre les plateformes SINP, tous les niveaux de validité des données sont concernés. Ainsi, les données « Très probables » à « Invalides » peuvent être échangées. En revanche, pour les transferts à l'INPN, seules les données « Très probables », « Probables », « Non réalisables » et « Non évaluées » sont concernées, en écartant les données « Douteuses » et « Invalides ».

Afin de faciliter les mises à jour et d'effectuer la validation des données, le contrôle automatique sera effectué une fois par semaine, idéalement à partir du vendredi soir jusqu'au samedi en début d'après-midi. Sont concernées les données ayant un statut « A valider ».

Bibliographie

Delauge J., Kapfer, G., Honoré, P., Guillaud, F., 2021. Protocole de validation des données naturalistes faunistiques dans Silene. Conservatoire d'espaces naturels de Provence-Alpes-Côte d'Azur – Janvier 2021. Sisteron, 13

FAUNA, 2020. Procédure de validation régionale des données d'occurrence de taxon de
L'Observatoire FAUNA. Version 1.5. 16 p.

Robert S., Barneix M., Body G., Castanet J., Caze G., Cellier P., Desse A., de Mazières J., Fromage P., Gourvil J., Jomier R., Juste A., Landry P., Lebeau Y., Lecoq M.E., Lescure J., Marage D., Meyer D., Pamerlon S., Papacotsia A., Poncet L., Quaintenne G., Saltré A. & Touroult J. 2016. *Guide méthodologique pour la conformité, la cohérence et la validation scientifique des données et des métadonnées du SINP – Volet 1 : occurrences de taxons, Version 1.* Rapport pour le SINP, rapport MNHN-SPN 2016-77, 63 p.

Robert S., Dupont P., de Mazières J., Poncet L., Touroult J., 2017. *Procédure nationale de validation scientifique des données élémentaires d'échanges du SINP pour les occurrences de taxons. Version 1.* Service du patrimoine naturel, Muséum national d'histoire naturelle, Paris. Rapport SPN 2017 – 2. 16 p.

Robert S., Jomier R., Milon T., Panijel J., Vest F., Barneix M., Fromage P. 2017. *Procédure de conformité et de cohérence des données et des métadonnées circulant entre les plateformes du SINP - Thématique : Occurrences de taxon, Version 1.0.* Rapport pour le SINP. UMS 2006 Patrinat (AFB/CNRS/MNHN), Paris, 26 pp.