SINP Occitanie

Protocole de validation scientifique des données d'occurrences de taxons du pôle Flore / Fonge / Habitats

Conservatoire botanique national méditerranéen de Porquerolles
Conservatoire botanique national des Pyrénées et de Midi-Pyrénées

Historique des versions du document :

Version	Date	Commentaire
1	29/12/2021	Document présenté au GT connaissance du CSRPN le 13/01/2022

1. Rappel du contexte : la validation dans le cadre du SINP

La démarche générale de la validation dans le SINP comprend quatre phases (les définitions ci-dessous sont issues du guide méthodologique national de Robert S. et al, 2016[1]) :

- l'**identification des doublons**

Contrôle préliminaire visant à éviter l'intégration de doublons potentiels dans le SINP et qui est mis en place à l'entrée de la plateforme régionale, par les pôles taxinomiques, de manière à vérifier que la donnée ou la métadonnée intégrée n'est pas déjà présente.

- le **contrôle de conformité**

Cette étape consiste à vérifier la conformité des fichiers échangés au regard des standards d'échange du SINP (données et métadonnées). Ces vérifications concernent la présence des champs obligatoires, le type, les règles et le format des fichiers échangés, mais également la vérification des valeurs faisant appel à un vocabulaire contrôlé sous forme de listes de valeurs/nomenclatures ou de référentiels.

- le **contrôle de cohérence**

Contrôle de la cohérence des informations transmises dans les données et les métadonnées (par exemple : date de fin > date de début).

- la **validation scientifique**

Elle consiste en des processus d'expertises visant à renseigner sur la fiabilité (désigne le degré de confiance que l'on peut accorder à la donnée). Ces processus font intervenir des bases de connaissance et/ou de l'expertise directe.

➔ **Le présent document porte sur la validation scientifique des données d'occurrences de taxon.**

La validation des données dans le SINP est organisée en 3 niveaux : niveau producteur, niveau régional et niveau national. Les différents niveaux de validité coexistent et circulent avec la DEE (Donnée Elémentaire d'Echange), sont indépendants les uns des autres et ne se substituent pas entre eux.

La validation dite « producteur » permet au producteur d'auto-évaluer les données qu'il a lui- même produit ou de transmettre le résultat d'une validation tierce réalisée sur ses données. Le producteur de données qui contribue au SINP peut être n'importe quelle structure ou individu produisant de la donnée naturaliste (association, bureau d'étude, collectivité, CBN, CEN, espace naturel ou protégé,..).

La validation régionale, coordonnée par la plateforme régionale, est mise en œuvre par les pôles taxinomiques. Parmi ces pôles, le pôle FFH qui administre les données sur la flore (plantes

[1] Robert S. *et al*. 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.

vasculaires, bryophytes, algues), la fonge et les habitats naturels, est animé par les 2 Conservatoires botaniques nationaux : CBN Méditerranéen (CBNMed) et CBN des Pyrénées et de Midi-Pyrénées (CBNPMP).

La validation nationale, coordonnée par la plateforme nationale, est réalisée de manière globale en s'appuyant sur les réseaux d'experts nationaux et sur les retours des utilisateurs (via le site de l'INPN ou encore via les travaux de réutilisation des données dans le cadre de programmes nationaux). Pour la mise en œuvre le MNHN, responsable de la plateforme nationale, peut mandater ses réseaux partenaires le cas échéant.

2. Organisation de la validation régionale par le pôle Flore / Fonge / Habitats

Chaque CBN assure la gestion et la validation des données d'occurrence situées sur son territoire d'agrément, à savoir les départements 09, 12, 31, 32, 46, 65, 81 et 82 pour le CBNPMP et 11, 30, 34, 48 et 66 pour le CBNMed.

Bien que chaque CBN utilise pour ce faire son propre système d'information (Lobelia pour le CBNPMP et Simethis pour le CBNMed) les processus de qualification et de gestion des données sont similaires. Les détails techniques propres à chaque outil sont disponibles en ligne (SIMETHIS-processus_integration_validation et LOBELIA-validation)

Les CBNMed et CBNPMP partagent respectivement leurs outils avec, d'une part, les CBN Alpin et Corse, et d'autre part, les CBN Sud-atlantiques, Massif central et Bassin parisien. Ceci assure une cohérence inter-régionale dans la procédure de validation des données d'occurrences de taxons, ce qui répond à l'une des recommandations du GT national « validation ».

Le producteur de données peut, soit utiliser directement un des outils des CBN pour la saisie de ses données, auquel cas la donnée est directement conforme aux exigences du SINP, soit le producteur met à disposition ses données sous forme d'un export (fichier ou flux de données), auquel cas le CBN doit s'assurer de la vérification des doublons, de la conformité, ou de la mise en conformité lorsque cela est possible, et de la cohérence des données.
La vérification des doublons est basée sur une comparaison de la date d'observation, de(s) observateur(s), du taxon et de la localisation entre les observations à intégrées et celles déjà dans la base de données du pôle. Les doublons repérés par cette méthode ne seront pas intégrés ou provoqueront une mise à jour de la donnée préexistante en accord avec le producteur.
La conformité consiste à vérifier que les champs minima requis soient bien fournis par le producteur : date de l'observation, nom de l'observateur (et structure le cas échéant), taxon observé, localisation.
Des tests de cohérence sont aussi appliqués : cohérence sur la date, cohérence spatiale,…
Si nécessaire les données sont également rattachées aux référentiels nationaux en vigueur (taxref pour la taxinomie, code insee, …) et au catalogue régional de métadonnées (cadres d'acquisition et jeux de données mis en commun pour les différents pôles taxinomiques, eux-mêmes mis en cohérence avec le catalogue national des métadonnées).

Dans un second temps, les données vont être validées scientifiquement et recevoir un niveau de validité selon le standard national (extension validation pour occurrence de taxons) :

	Validation automatique	**Validation manuelle**
Certain - très probable	La donnée présente un haut niveau de vraisemblance (très majoritairement cohérente) selon le protocole automatique appliquée. Le résultat de la procédure correspond à la définition optimale de satisfaction de l'ensemble des critères du protocole automatique, par exemple, lorsque la localité correspond à la distribution déjà connue et que les autres paramètres écologiques (date de visibilité, altitude, etc.) sont dans la gamme habituelle de valeur.	La donnée est exacte. Il n'y a pas de doute notable et significatif quant à l'exactitude de l'observation ou de la détermination du taxon. La validation a été réalisée notamment à partir d'une preuve de l'observation qui confirme la détermination du producteur ou après vérification auprès de l'observateur et/ou du déterminateur.
Probable	La donnée est cohérente et plausible selon le protocole automatique appliqué mais ne satisfait pas complètement (intégralement) l'ensemble des critères automatiques appliqués. La donnée présente une forte probabilité d'être juste. Elle ne présente aucune discordance majeure sur les critères jugés les plus importants mais elle satisfait seulement à un niveau intermédiaire, ou un ou plusieurs des critères automatiques appliqués.	La donnée présente un bon niveau de fiabilité. Elle est vraisemblable et crédible. Il n'y a, a priori, aucune raison de douter de l'exactitude de la donnée mais il n'y a pas d'éléments complémentaires suffisants disponibles ou évalués (notamment la présence d'une preuve ou la possibilité de revenir à la donnée source) permettant d'attribuer un plus haut niveau de certitude.
Douteux	La donnée concorde peu selon le protocole automatique appliqué. La donnée est peu cohérente ou incongrue. Elle ne satisfait pas ou peu un ou plusieurs des critères automatiques appliqués. Elle ne présente cependant pas de discordance majeure sur les critères jugés les plus importants qui permettraient d'attribuer le plus faible niveau de validité (invalide).	La donnée est peu vraisemblable ou surprenante mais on ne dispose pas d'éléments suffisants pour attester d'une erreur manifeste. La donnée est considérée comme douteuse
Invalide	La donnée ne concorde pas selon la procédure automatique appliquée. Elle présente au moins une discordance majeure sur un des critères jugés les plus importants ou la majorité des critères déterminants sont discordants. Elle est considérée comme trop improbable (aberrante notamment au regard de l'aire de répartition connue, des paramètres biotiques et abiotiques de la niche écologique du taxon). Elle est considérée comme invalide.	La donnée a été infirmée (erreur manifeste/avérée) ou présente un trop bas niveau de fiabilité. Elle est considérée comme trop improbable (aberrante notamment au regard de l'aire de répartition connue, des paramètres biotiques et abiotiques de la niche écologique du taxon, la preuve révèle une erreur de détermination). Elle est considérée comme invalide.

	Validation automatique	Validation manuelle
Non réalisable	La donnée a été soumise à l'ensemble du processus de validation mais l'opérateur (humain ou machine) n'a pas pu statuer sur le niveau de fiabilité, notamment à cause des points suivants : état des connaissances du taxon insuffisantes, ou informations insuffisantes sur l'observation.	

Cette validation scientifique s'accompagne de métadonnées permettant de préciser le contexte de la validation : date de validation, méthode de validation, nom et organisme du validateur,….

Etant donné le nombre important de données à valider annuellement, cette validation scientifique se déroule en 2 étapes complémentaires, une validation automatique et/ou une validation manuelle par expert.

3. Procédure de validation scientifique automatique

La procédure de validation automatique consiste à comparer de manière automatique la donnée en cours de validation aux observations du même taxon déjà validées et d'en déduire un niveau de validation. Ainsi les CBN ayant validé des centaines de milliers de données manuellement à dire d'expert pour la flore vasculaire, ils peuvent s'appuyer sur cette base de connaissance pour alimenter leur modèle de validation automatique. Ce n'est pas le cas pour les bryophytes, les champignons et les lichens, pour lesquels les connaissances sont plus parcellaires et la taxinomie encore évolutive. Pour ces derniers groupes, la validation automatique est rarement appliquée par le CBNMed, et le CBNPMP a fait le choix de valider pour l'instant « manuellement » ces données.

La validation automatique est appliquée grâce à des développements mis en œuvre au sein des systèmes d'information propres à chacun des CBN. Ces outils étant différents, le fonctionnement de la validation automatiques est lui aussi légèrement différent. Néanmoins les 2 CBN utilisent les mêmes types de critères qui sont : les catalogues départementaux (présence / absence d'un taxon), la cohérence territoriale et altitudinale et la difficulté de détermination.

Ces fonctionnements sont présentés dans les 2 paragraphes suivants :

➢ *Pour le CBNMed*

La validation scientifique automatique fait appel à plusieurs critères :

- **Catalogue départemental :** comparaison avec le catalogue de référence listant la présence des taxons par département.
- **Cohérence territoriale :** vérification si le taxon a déjà été observé récemment (>= 2000) dans un rayon de 5 kilomètres autour du lieu de l'observation.
- **Cohérence altitudinale :** vérification que l'altitude de l'observation est dans la gamme altitudinale connue pour le taxon (statistique sur les observations valides et précises géographiquement présentes dans SIMETHIS).

- **Détermination :** absence de la liste des taxons complexes à déterminer définie par les experts (listes départementales).
- **Enjeu réglementaire ou de conservation :** taxon non protégé et non fortement menacée (EN, VU dans les listes rouges régionales).

Les observations répondant positivement à l'ensemble des critères sont automatiquement tagguées au niveau de validité "probable" et non spécialement considérée par les experts. Les observations répondant négativement à au moins l'un des critères sont orientées vers la validation manuelle.

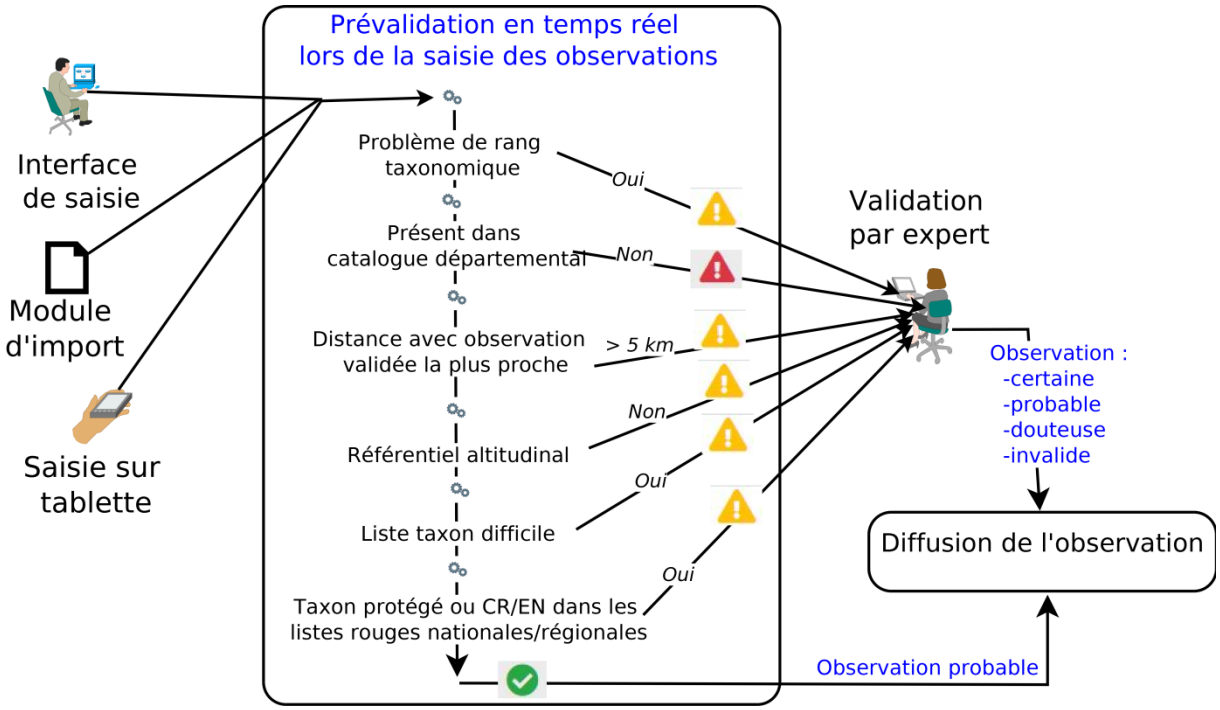


Schéma général de description des étapes du processus de validation scientifique au sein de Simethis

➢ *Pour le CBNPMP*

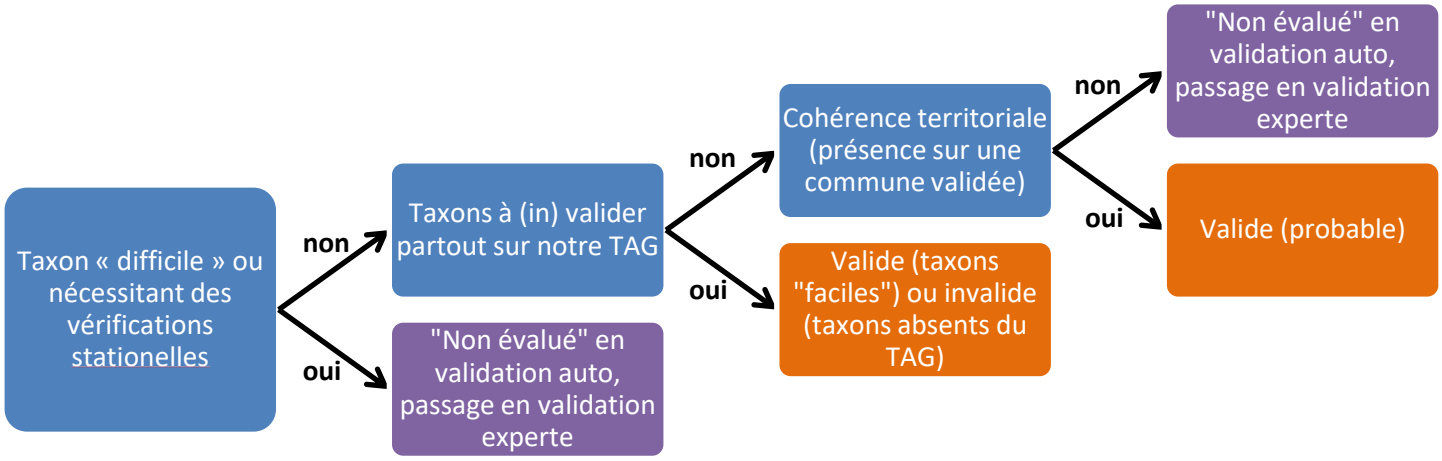
La procédure exposée ci-dessous est celle appliquée jusqu'au 31/12/2021. A partir de janvier 2022, le CBNPMP fait évoluer ses outils de gestion et de validation de données (mutualisation à 4 CBN – CBNPMP, Sud-atlantique, Massif central et Bassin Parisien - autour d'un outil métier commun) et la procédure de validation automatique sera revue en collaboration avec ces 3 CBN. Des développements informatiques concernant validation automatique dans Lobélia sont ainsi prévus en 2022. La procédure sera néanmoins cohérente avec l'existant dans chaque région et sera partagée avec les CBNMed et Alpin avec qui les pôles Flore des SINP Occitanie et Auvergne-Rhône-Alpes sont co-animés. Le protocole de validation sera alors mis à jour.

Plusieurs critères sont appliqués au cours de la validation automatique :

- **Taxons « difficiles »**. Une liste de taxon établie à dire d'expert sur le territoire d'agrément est exclue de la validation automatique. Il s'agit des taxons difficiles à déterminer ou pour lesquels il faut vérifier l'altitude de présence ou la localisation précise. Si le taxon est dans la

liste des taxons difficiles, son statut reste « non évalué », il doit passer par la validation « manuelle ».

- **Taxons « faciles »**. Une liste de taxons communs et faciles à identifier avec une faible probabilité d'erreurs a été établie à dire d'experts. Ils sont validés automatiquement partout sur le territoire d'agrément.
- **Cohérence territoriale** et distribution de référence (la donnée est validée « probable » si elle se trouve sur une commune de présence validée manuellement pour ce taxon)
- **Catalogue départemental** : la donnée est automatiquement invalidée si le taxon est assurément absent d'un territoire, par exemple taxon littoral se trouvant dans les Pyrénées centrales.



Etapes de la validation automatique

4. Procédure de validation scientifique « manuelle » ou experte

Si la validation automatique ne permet pas de déduire un niveau de validité, la donnée sera alors analysée par un expert. Ce dernier, de part ses connaissances et les éléments de contextes en sa possession (expérience de l'observateur, période d'observation, condition écologiques, etc.) ou suite à des échanges avec le producteur, pourra alors attribuer manuellement un niveau de validation.

Ces experts peuvent intervenir sur les données, quel que soit le niveau de validité attribué automatiquement à la donnée, y compris si la donnée n'a pas suivi la phase automatique. C'est la validation dite « manuelle » qui prime sur la validation automatique. Ainsi une donnée validée automatiquement pourra être invalidée ou indiquée douteuse par un expert (ceci permettant de mieux calibrer la modèle de validation automatique).

Outils pour la validation experte :

- Aire de répartition des taxons (SI Flore, catalogues départementaux de la région Occitanie permettant de mettre en relief les taxons nouveaux pour un département lors du processus de validation)
- Flores, ouvrages de référence

Le validateur peut se permettre d'intervenir sur le rattachement taxonomique, pour préciser un rang taxonomique plus fin par exemple (pour une donnée à l'espèce il s'agit d'indexer la donnée sur la seule sous-espèce présente dans la région par exemple). Le nom d'origine du taxon fourni par le producteur est dans tous les cas conservé dans le champ « nom cité ».

Dans les cas où la modification du référentiel taxinomique est de nature à modifier le niveau de validité (exemple d'un splittage), les données concernées repassent dans le circuit de validation et la date de validation est bien sûr mise à jour.

L'expert décidera d'attribuer manuellement à l'observation un des niveaux de validité suivant :

- Certain : si une preuve de l'observation (part d'herbier, photographies) a été vue par l'expert ou s'il a eu vérification auprès de l'observateur / déterminateur.
- Probable : si les éléments en sa possession lui rendent l'observation crédible.
- Douteux : si les éléments en sa possession lui rendent l'observation peu crédible sans toutefois pouvoir vérifier qu'il s'agit d'une erreur.
- Invalide : si les éléments de vérification prouvent une erreur
- Non réalisable : en cas d'absence totale d'éléments pouvant aider à une vérification (nom inconnu ou non identifié dans les référentiels taxonomiques en vigueur).

Les validateurs sont des personnes référentes dans leur domaine de compétence et sur un territoire, dans chacun des CBN :

- CBNPMP : 4 botanistes, 1 bryologue et 2 mycologues
- CBNMED : 3 botanistes et 1 bryologue

5. Prise en compte des validations des producteurs et échanges sur la validation régionale

Le producteur de données peut avoir attribué un niveau de validation à ses observations. La validation du producteur est mentionnée dans le format de donnée élémentaire d'échange du SINP et est associée à la donnée au même titre que la validation régionale et la validation nationale. Cette information est utilisée pour orienter la validation régionale experte notamment lorsque le producteur estime lui-même que sa donnée est douteuse ou invalide.

Le résultat de la validation régionale, ainsi que les métadonnées associées et les éventuelles modifications faites sur le rattachement taxinomique sont envoyées en retour au producteur de données, a minima annuellement, pour échange et, si possible, archivage.

Des échanges, en bilatéral entre le producteur et l'un des CBN, ou en réunions de pôles par groupe taxinomique (flore vasculaire, champignons) sont organisés a minima annuellement et permettent d'affiner la validation scientifique des données et de discuter notamment des groupes difficiles à déterminer et des erreurs fréquentes.