



GT Validation des données d'occurrences du SINP
Volet 1 : occurrences de taxons

**Guide méthodologique pour la
conformité, la cohérence et la validation
scientifique des données et des
métadonnées du SINP**

Version 1

Mars 2016

Historique des versions du document :

<i>Versions</i>	<i>Date</i>	<i>Commentaires</i>
0	18/08/2015	Document préliminaire
0.1	08/10/2015	Document modifié suite à la tenue d'une réunion du groupe de travail
0.2	20/01/2016	Document modifié suite aux relectures du groupe et à la tenue d'une nouvelle réunion du groupe de travail
1	03/03/2016	Document validé suite aux relectures du groupe de travail

Coordinatrice du groupe de travail et rédactrice du guide méthodologique:

Solène Robert – MNHN/SPN

Composition du groupe de travail / contributeurs:

DREAL Aquitaine	Andy Papacotsia
DREAL Bourgogne	Damien Marage
DREAL Provence Alpes Côte d'Azur	Dorothee Meyer
FCBN & CBN	Alexis Desse Anaïs Just Grégory Caze Johan Gourvil
LPO	Gwenaël Quaintenne
MEDDE	Yannick Lebeau
MNHN/GBIF	Marie-Elise Lecoq Sophie Pamerlon
MNHN/SPN	Solène Robert Jeanne de Mazières Julien Touroult Laurent Poncet Rémy Jomier
OAFS	Marie Barneix Paul Fromage
OBHN	Arnaud Saltré
ONCFS	Guillaume Body Philippe Landry
SHF	Jacques Castanet Jean Lescure
TELA BOTANICA	Pierre Cellier

Relecteurs du guide méthodologique :

V0	Guillaume Body, Jeanne de Mazières et Julien Touroult
V0.1	Guillaume Body, Johan Gourvil et Jeanne de Mazières
V0.2	Marie Barneix, Pascal Dupont, Johan Gourvil, Rémy Jomier, Damien Marage, Jeanne de Mazières, Laurent Poncet, Arnaud Saltré et Julien Touroult

Autres contributeurs :

Virgile Noble (CBN Méditerranéen de Porquerolles)

Julie Delauge (CEN PACA)

Pascal Dupont (MNHN/SPN)

Référence du guide :

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., Quintenne 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.

Documents connexes produits dans le cadre du GT Validation:

Mandat du groupe de travail « Validation des données d'occurrence », Version 4, 2015.

http://www.naturefrance.fr/sites/default/files/fichiers/ressources/pdf/mandat_gt_qualification_sinp_v4_0.pdf

Recensement de l'existant pour la validation des données d'occurrence du SINP, 2015.

Robert S., de Mazières J., Poncet L., Desse A., Saltré A., Gourvil J., Meyer D., Touroult J., Leveugle F., Delauge J., Gauthier E., Filoche S., Bouix T., Archambeau A.S. & Lebeau Y. 2015. *Recensement de l'existant pour la validation des données d'occurrence du SINP*. Rapport pour le SINP, rapport MNHNSPN 2015-44, 54 p.

Synthèse des réponses au questionnaire pour l'identification des besoins des différents acteurs du SINP en matière de validation, 2015

Robert S., Poncet L., de Mazières J., Gourvil J., Touroult J. 2015. *Synthèse des réponses au questionnaire pour l'identification des besoins des différents acteurs du SINP en matière de validation*. Rapport pour le SINP, rapport MNHN-SPN 2015-62, 19 p.

SOMMAIRE

I. Contexte et objectifs	7
II. Portée du guide	8
III. Définition des concepts	9
1. Conformité	9
2. Cohérence	9
3. Validation scientifique.....	9
4. Qualification	10
IV. Périmètre du guide et prérequis	11
V. La validation dans l'architecture du SINP.....	13
1. Doublons.....	14
2. Conformité et cohérence	15
3. Validation scientifique.....	17
VI. Procédure pour l'identification des doublons, pour la conformité et la cohérence	19
1. Principes	19
2. Protocole de doublons, de conformité et de cohérence	21
3. Référentiels	25
4. Rapport de doublons, de conformité et de cohérence	26
VII. Procédure pour la validation scientifique	27
1. Principes	27
2. Règles de validation.....	29
3. Référentiels et bases de connaissance.....	34
4. Protocole de validation	35
5. Niveaux de validité	37
6. Rapport de validation scientifique	40
VIII. Regroupements de données	41
IX. Mises à jour	42
1. Cas pouvant nécessiter une mise à jour inhérente à la validation.....	42
2. Démarche de mises à jour des contrôles de conformité et de cohérence	44
3. Démarche de mise à jour des niveaux de validité scientifiques et informations associées .	44
X. Exigences fonctionnelles	46
1. Contrôle de doublons.....	46
2. Contrôles de conformité et de cohérence	47

3. Validation scientifique.....	49
XI. Autres besoins pour assurer la validation	51
1. Echange d'informations entre les plateformes.....	51
2. Gestion des plateformes	52
3. Diffusion	52
4. Outils	52
XII. Bonnes pratiques pour améliorer la qualité des données et faciliter leur validation.....	53
XIV. Perspectives	55
XV. Glossaire.....	56

I. Contexte et objectifs

Une multitude de jeux de données sur la nature et les paysages est produite chaque année par de nombreux acteurs sur les territoires français (métropole et outre-mer). Ces données¹ sont collectées selon des méthodes et protocoles différents et pour des objectifs variés. Elles concernent des périmètres taxonomiques et géographiques divers et les compétences et moyens des observateurs sont hétérogènes.

Dans le cadre du SINP (Système d'Information sur la Nature et les Paysages), les différents utilisateurs¹ doivent être en mesure de connaître le niveau de confiance associé aux données (objectifs du SINP listé dans l'Article 3 du protocole² de 2013 : « définir et mettre en œuvre des critères de qualité des données »). Cela permet notamment une utilisation appropriée des données au regard des différents usages.

Les différents opérateurs en charge de la mise en œuvre de la procédure de validation dans le cadre du SINP doivent s'accorder sur l'application d'un ensemble de règles communes afin de garantir une organisation opérationnelle et une démarche générique de ces opérations.

Les travaux du groupe de travail (GT) SINP sur la validation ont été définis par un mandat², version du 02/02/2015, qui a planifié les travaux du GT selon les trois phases suivantes.

Phase 1 : Etablissement d'un recensement de l'existant et restitution sous la forme d'un rapport de synthèse.

Ce recensement a été réalisé via un appel à contribution à l'ensemble des contacts du cercle 1 identifiés dans le mandat du GT qui ont également fait appel à leur réseau. Un rapport de synthèse des éléments recueillis a été produit².

Robert S., de Mazières J., Poncet L., Desse A., Saltré A., Gourvil J., Meyer D., Touroult J., Leveugle F., Delauge J., Gauthier E., Filoche S., Bouix T., Archambeau A.S. & Lebeau Y. 2015. *Recensement de l'existant pour la validation des données d'occurrence du SINP*. Rapport pour le SINP, rapport MNHNSPN 2015-44, 54 p.

Phase 2 : Partage des définitions conceptuelles, identification des besoins, précision du périmètre et formalisation des objectifs.

Les définitions des concepts sont diffusées dans le glossaire du site de Nature France.

L'identification des besoins a été réalisée via la mise en place d'un questionnaire en ligne visant à identifier :

- les grands types de besoins par grandes catégories d'acteurs (producteur¹, gestionnaire¹, administrateur¹, utilisateur¹, etc.),
- leurs priorités et leurs pratiques en matière de validation.

Une synthèse des éléments recueillis a été produite et diffusée².

¹ Voir XV. Glossaire

² Diffusé sur Nature France : www.naturefrance.fr

Robert S., Poncet L., de Mazières J., Gourvil J., Touroult J. 2015. *Synthèse des réponses au questionnaire pour l'identification des besoins des différents acteurs du SINP en matière de validation*. Rapport pour le SINP, rapport MNHN-SPN 2015-62, 19 p.

Phase 3 : Analyse de l'existant et adaptation au SINP/rédaction du guide méthodologique sur la validation.

II. Portée du guide

Le présent guide, produit de la phase 3 du mandat du groupe, porte sur les données³ d'occurrence de taxons et les métadonnées³ associées (jeux de données³ et dispositifs de collecte³). Il vise à décrire la démarche de validation au sein du SINP et les éléments nécessaires à sa mise en œuvre.

³ Voir XV. Glossaire

III. Définition des concepts

Dans le cadre du SINP, les terminologies retenues pour caractériser la validation sont les suivantes :

1. Conformité

La conformité désigne le respect des règles fixées dans le cadre de la mise en œuvre des formats standards⁴ de données et de métadonnées autant sur les aspects physiques que conceptuels (renseignement des champs obligatoires, format⁴, utilisation des référentiels et des listes de valeurs/nomenclatures).

Exemple : la date transmise est de type date norme ISO8601 comme spécifié par le standard de données élémentaires d'échange (DEE⁴). L'attribut⁴ X est renseigné avec des valeurs conformes à la nomenclature définie dans le standard dédié. L'attribut Y est renseigné avec le type d'information spécifié dans le standard (un nombre par exemple).

2. Cohérence

La cohérence désigne le respect de la logique combinatoire des informations transmises au sein des données, au sein des métadonnées et entre les données et les métadonnées.

Exemple : la date de début de l'observation⁴ est inférieure ou égale à la date de fin de l'observation.

3. Validation scientifique

La validation scientifique consiste en des processus d'expertises⁴ 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.

Exemple :

*- la donnée d'*Atrichum angustatum* (Brid.) Bruch & Schimp., observée le 5 janvier 2015 et située sur la maille 10kmL93E077N634 présente un bon niveau de fiabilité au regard des éléments fournis et de la répartition connue. La donnée est probable.*

*- la donnée de *Tursiops truncatus* (Montagu, 1821), observée le 3 août 2014 sur le département de l'Yonne est considérée comme trop improbable au regard de l'aire de répartition connue et de la niche écologique du taxon. La donnée est considérée comme invalide.*

⁴ Voir XV. Glossaire

On distingue :

- La validation scientifique dite automatique qui consiste en une validation faisant appel à des résultats d'expertises préalables, (des référentiels, des bases de connaissances, etc.). Ce processus est réalisé de manière automatique (informatique) dans le cadre du SINP.
- La validation scientifique dite manuelle (nonobstant le fait qu'une interface informatique peut être utilisée) qui consiste en une validation des informations faisant appelle à une expertise directe (avis d'expert suite à l'analyse des informations transmises). Ce processus est réalisé par un ou des experts dans le cadre du SINP.

4. Qualification

La qualification est l'aptitude à répondre aux usages. Elle est le résultat d'une sélection selon des critères inhérents à l'objectif visé et est donc variable en fonction de cet objectif.

Exemple : une donnée d'oiseau nicheur sur la zone géographique de l'observation sera qualifiée « utilisable » pour un atlas des oiseaux nicheurs mais ne sera pas « utilisable » pour un atlas des oiseaux hivernants.

IV. Périmètre du guide et prérequis

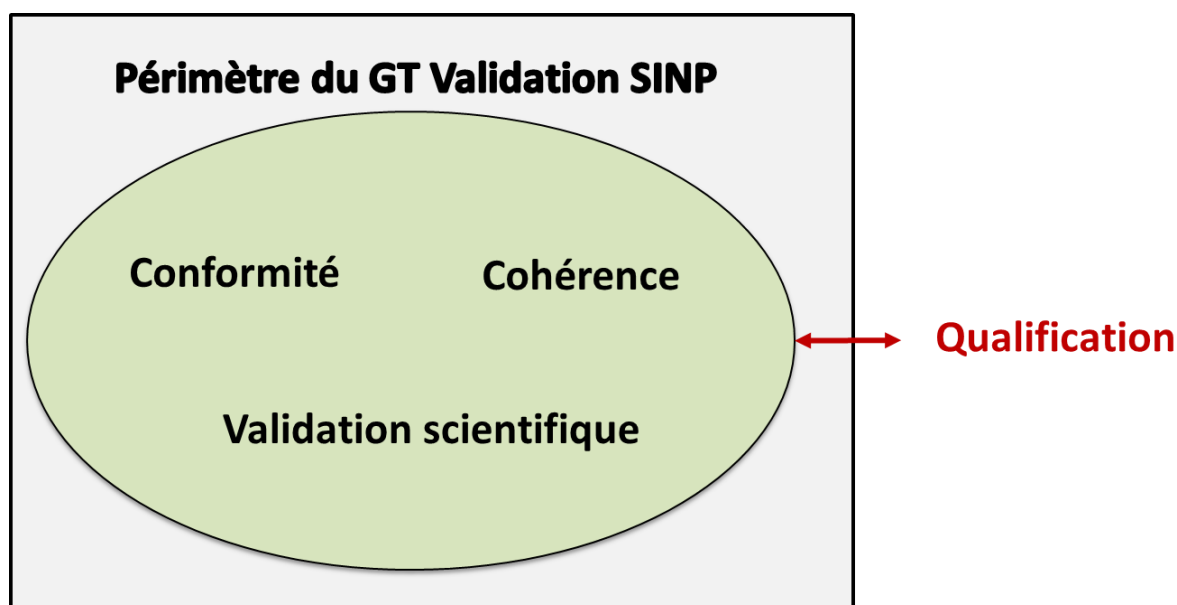
Les travaux du GT validation du SINP portent sur les données élémentaires d'échange (DEE) situées sur les territoires français (métropole et outre-mer) et les métadonnées associées (jeux de données et dispositifs de collecte).

Les procédures de validation scientifique pourront être cependant réalisées indifféremment sur la DEE ou sur la DS (Donnée Source) sous réserve que l'ensemble des attributs de la DS soit retranscrit au sein de la DEE.

Les travaux du GT ne portent pas sur les données dites « de synthèse » (atlas, cartes de répartition, etc.) qui sont destinées à des usages ou des exploitations spécifiques, mais concernent les données d'occurrences permettant, notamment, la réalisation de ces synthèses. L'élaboration et le contrôle des couches de référence et autres données de synthèse sont ainsi considérés comme hors périmètre du GT.

La qualification d'une donnée est directement liée aux usages que l'utilisateur souhaite en faire. Ces usages sont multiples et indéfinis ; ils sont propres à chaque thématique⁵ et à chaque usager.

Les travaux du groupe ne portent donc pas sur la qualification des données et de leurs métadonnées mais visent à fournir aux utilisateurs les informations sur la qualité des données mises à disposition afin qu'ils puissent sélectionner plus efficacement les données à prendre en compte selon leurs objectifs d'exploitation (adéquation à l'usage).



Périmètre du groupe de travail sur la validation des données et des métadonnées

⁵ Voir XV. Glossaire

La démarche de validation doit permettre d'assurer toute la traçabilité⁶ nécessaire à la valorisation des différents acteurs intervenant sur les opérations de validation et qui fournit aux différents utilisateurs des données les éléments nécessaires permettant de tracer le circuit de validation et les différents résultats des procédures mises en œuvre.

Lorsqu'un contact avec le producteur de la donnée ou de la métadonnée est nécessaire, celui-ci est réalisé, dans la mesure du possible, au plus près de la production de données (par le premier point d'entrée de la donnée dans le SINP). Ce principe permet de limiter les sollicitations multiples des producteurs par des interlocuteurs multiples. Cependant, pour la validation scientifique, des sollicitations directes sont tolérées sous réserve que celles-ci n'interfèrent pas avec les sollicitations réalisées par la plateforme d'entrée dans le système.

Les contrôles de conformité et de cohérence ainsi que la validation scientifique (opérations organisées par les plateformes) sont des opérations mises en œuvre **selon la précision la plus fine disponible pour la DEE** sans préjuger des opérations de normalisation inhérentes à la sensibilité des données ou aux modalités de diffusion⁷ qui pourront être mises en place par les différentes plateformes.

La validation scientifique du producteur est réalisée sur la DS (Donnée source).

La démarche de validation des données dans le cadre du SINP est basée sur un **principe d'indépendance** des niveaux de validité producteur, régional et national.

A l'instar des grands principes du SINP, la démarche de validation définie par le GT Validation respecte le principe de non intrusion dans les systèmes existants, et ce, tout particulièrement concernant la mise en œuvre de la validation dite « producteur ».

La démarche proposée dans ce guide méthodologique est issue d'un compromis entre une uniformisation des pratiques en matière de validation et la prise en compte des organisations, des systèmes et des protocoles de validation existants.

Le présent guide ne traite pas de la question de la diffusion⁷ des données par les plateformes. Les modalités de diffusion en fonction des niveaux de validité et des différents groupes d'utilisateurs ne sont ainsi pas abordées dans le présent document mais mériteront toutefois d'être discutées, dans un second temps, au sein des différentes instances du SINP concernées.

⁶ Voir XV. Glossaire

⁷ La diffusion comprend ici la recherche et la visualisation ainsi que le téléchargement / Voir XV. Glossaire

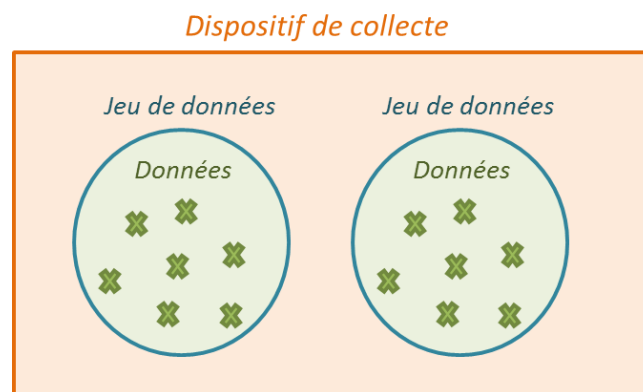
V. La validation dans l'architecture du SINP











La démarche générale de la validation dans le SINP comprend quatre phases :

- l'identification des doublons
- le contrôle de conformité
- le contrôle de cohérence
- la validation scientifique

Les métadonnées⁸ sont concernées par l'identification des doublons et les contrôles de conformité et de cohérence.

Les données sont concernées par l'identification des doublons, les contrôles de conformité et de cohérence et la validation scientifique.



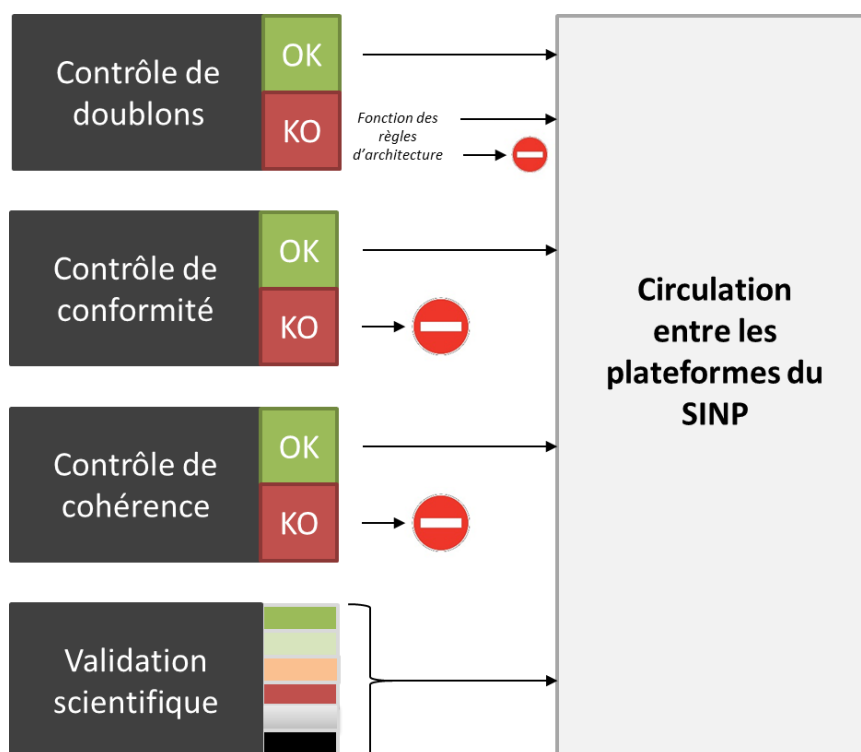
Doublons	Conformité	Cohérence	Validation scientifique
  	  	  	

L'identification de doublons doit donner lieu à des échanges entre la plateforme qui les identifie, le fournisseur et le producteur des données ou des métadonnées concernées. Le GT Validation ne saurait statuer sur les règles de mises en circulation de ces doublons dont la définition relève de l'architecture du SINP.

Le caractère non conforme ou incohérent d'une donnée, d'un jeu de données ou d'un dispositif de collecte donne lieu à un refus à l'entrée dans le système : ils ne peuvent alors pas circuler entre les plateformes.

⁸ La notion de métadonnées regroupe les dispositifs de collecte et les jeux de données

En revanche, la validation scientifique n'induit pas d'obstruction à la circulation. Toutes les données, quel que soit leur niveau de validité peuvent être échangées entre les plateformes.



Le processus de circulation des flux de données et de métadonnées (workflow) au sein des plateformes est défini par les travaux du GT Architecture du SINP.

Les responsables des plateformes régionales peuvent mandater, en fonction de l'organisation en place, leurs pôles régionaux⁹ pour la mise en œuvre des différentes opérations.

1. Doublons

Les données et les éléments de métadonnées sont concernés par les vérifications relatives à la présence de doublons⁹.

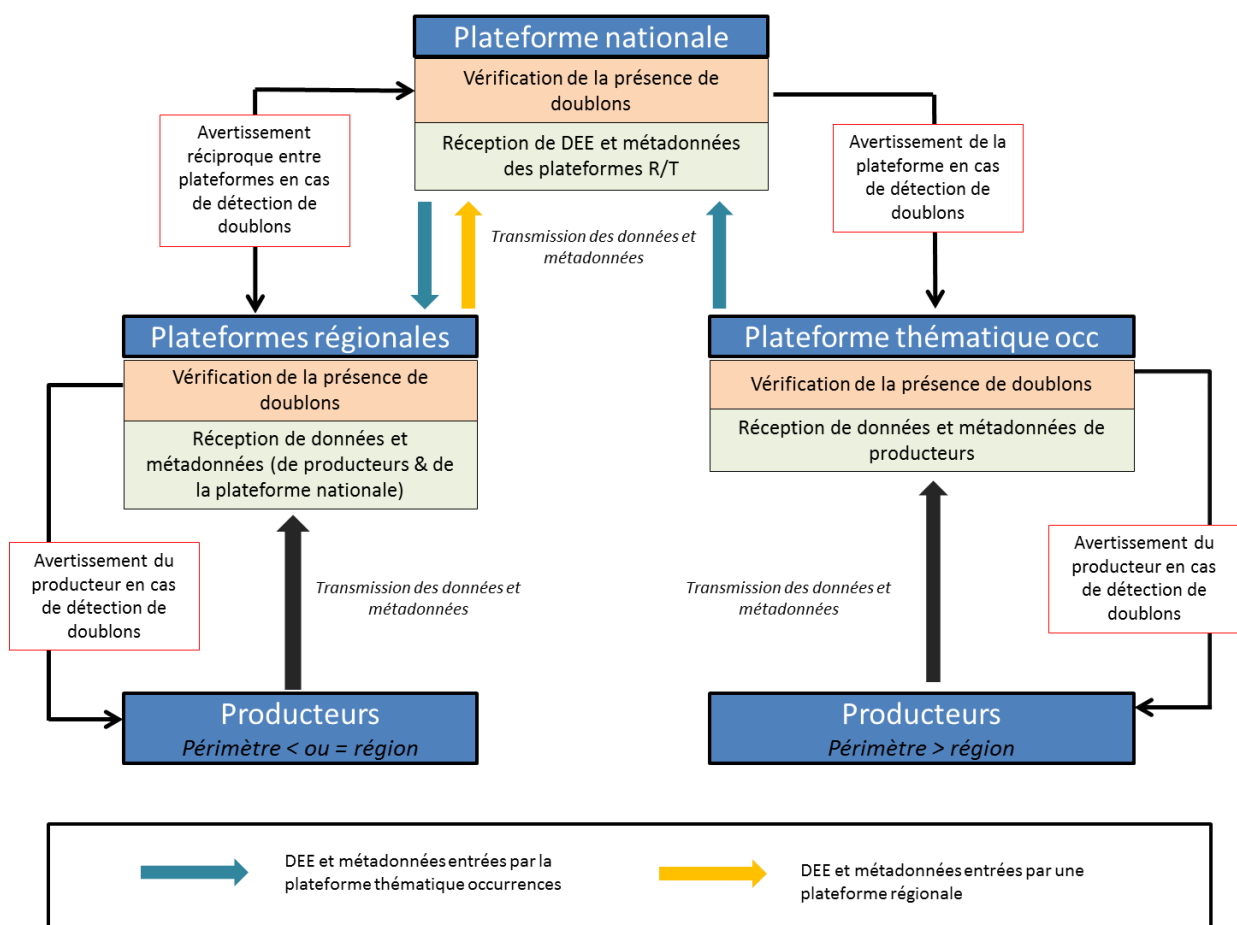
Pour rappel, l'architecture du SINP prévoit que les données ou métadonnées ne doivent entrer dans le système que par un point unique. C'est-à-dire qu'une même donnée ou métadonnée ne doit, en principe, pas être livrées par le producteur à plusieurs plateformes d'entrée dans le système (plateformes régionales et plateforme thématique occurrences de taxon). La vigilance est donc de mise lors de l'intégration des données. Une animation coordonnée des flux de données par les animateurs des plateformes minimiserait les risques de doubles intégrations.

⁹ Voir XV. Glossaire

En préalable à la réalisation des contrôles de conformité, de cohérence et à la validation scientifique, le présent GT identifie ainsi un contrôle préliminaire visant à éviter l'intégration de doublons potentiels dans le système. Ce contrôle doit être mis en place à l'entrée de chaque plateforme (régionale, thématique et nationale), de manière à vérifier que la donnée ou la métadonnée intégrée n'est pas déjà présente dans le système.

L'enjeu de ce traitement consiste en l'identification des doublons. Leur traitement devra être discuté dans le cadre des travaux du GT Architecture.

En cas d'identification de doublons potentiels, la plateforme doit en avertir le fournisseur des données et des métadonnées (le producteur ou la plateforme source en fonction des cas) afin de définir de l'organisation à mettre en œuvre.



Démarche générale pour l'identification des doublons dans le SINP

2. Conformité et cohérence

Les données et les éléments de métadonnées sont concernés par les contrôles de conformité et de cohérence.

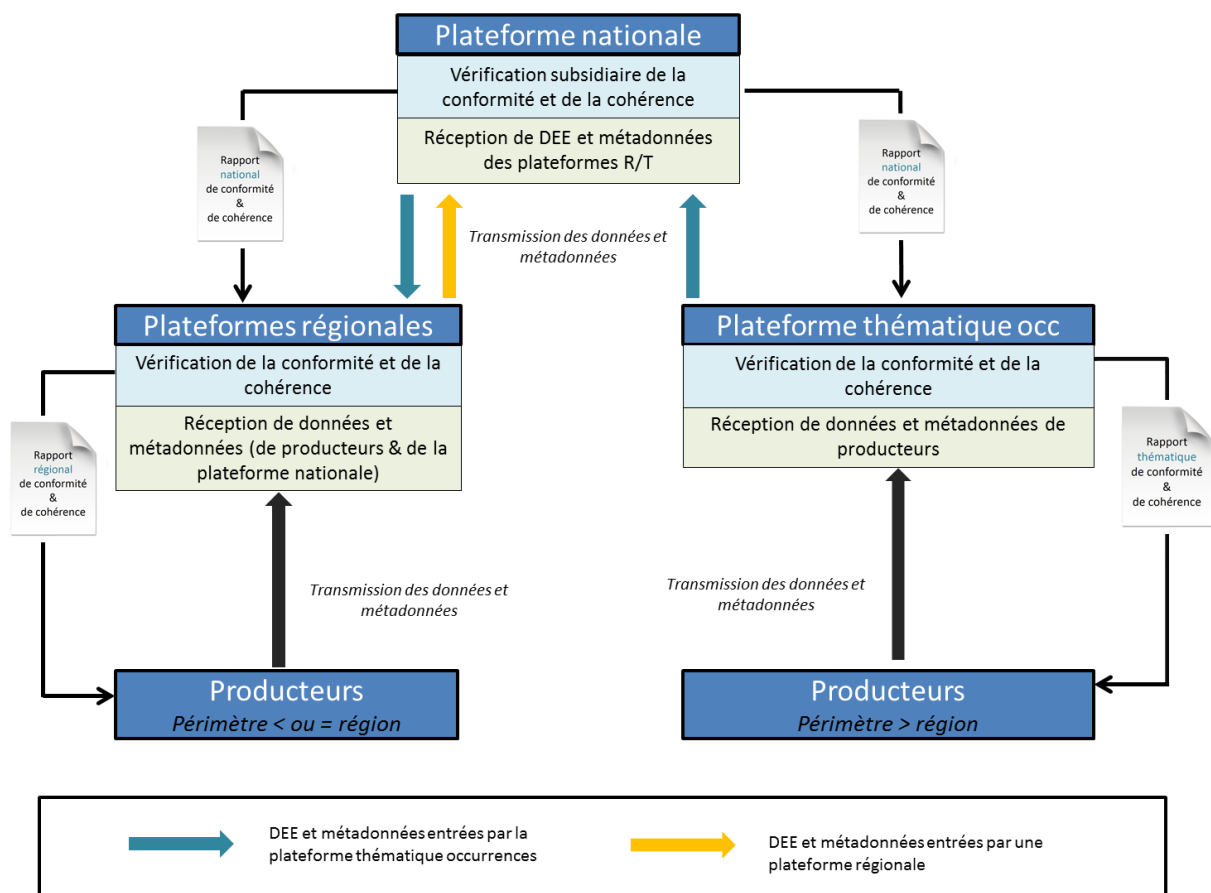
Les DEE et métadonnées associées circulant au sein du système doivent être conformes et cohérentes.

Les plateformes régionales et la plateforme thématique pour les occurrences ont notamment pour mission d'assurer la production des DEE à partir des données transmises par les producteurs. A ce titre, **elles portent la responsabilité de la conformité et de la cohérence des DEE et métadonnées associées qu'elles produisent.**

Les producteurs doivent être informés, via un rapport de conformité et de cohérence régional ou thématique, des erreurs et des manques le cas échéant ou du bon déroulement de la transformation. Ce rapport doit être transmis aux producteurs quel que soit le résultat de la transformation (conforme/non conforme – cohérent/incohérent).

Le terme de rapport est ici utilisé de manière générique, il peut consister en un document (.doc, .pdf, tableur ou tout autre format) ou encore en une page de consultation web sur la plateforme. Sa production peut être automatisée ou pas.

La plateforme nationale mettra également en place des contrôles de conformité et de cohérence et alertera les plateformes régionales et thématiques sur les éventuelles erreurs détectées par le biais d'un rapport national de conformité et cohérence.



Démarche générale pour les étapes de conformité et de cohérence.

Seules les données conformes et cohérentes circulent dans le SINP.

3. Validation scientifique

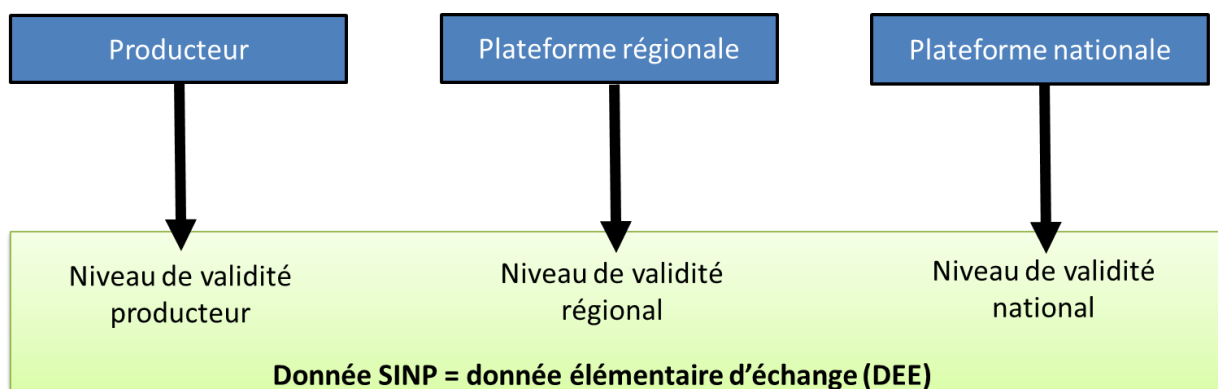
Seules les données sont concernées par la validation scientifique. (Les métadonnées ne sont pas concernées par cette opération.)

Les DEE circulant au sein du système portent trois niveaux de validité, résultats des validations scientifiques producteur, régionale et nationale. **Les différents niveaux de validité coexistent (c'est-à-dire que les trois niveaux circulent avec la DEE), sont indépendants les uns des autres et ne se substituent pas entre eux.**

La validation dite « producteur » permet au producteur d'autoévaluer les données qu'il a lui-même produit ou de transmettre le résultat d'une validation tierce réalisée sur ses données.

La validation régionale, coordonnée par la plateforme régionale, est réalisée au plus proche des systèmes producteurs et des experts régionaux. **Pour la mise en œuvre, les plateformes régionales peuvent mandater leurs réseaux partenaires, des pôles, organismes ou instances dédiés.**

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.**

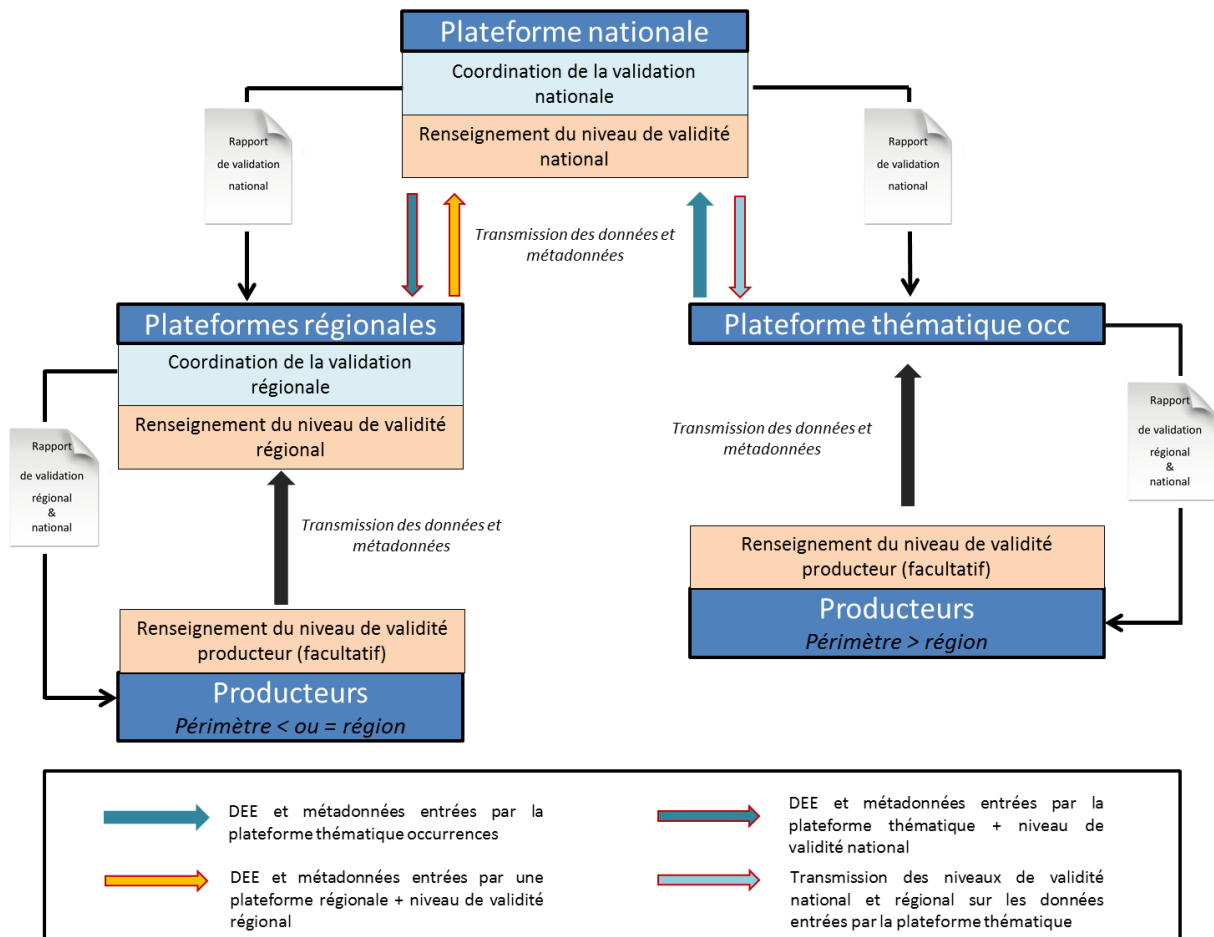


La procédure de validation scientifique ne doit en aucune mesure freiner la circulation des DEE entre les plateformes. Les données peuvent, de fait, circuler sans obligation de soumission à la validation scientifique.

Les producteurs sont informés des résultats de la validation régionale et de la validation nationale par le biais de la plateforme par laquelle la donnée est entrée dans le système.

La plateforme nationale transmet à chaque plateforme régionale et thématique le résultat de la validation nationale et l'ensemble des informations nécessaires à l'établissement du rapport de validation par ces dernières.

Le terme de rapport est ici utilisé de manière générique, il peut consister en un document (.doc, .pdf, tableur ou tout autre format) ou encore en une page de consultation web sur la plateforme. Sa production peut être automatisée ou pas.



Démarche générale pour l'étape de validation scientifique.

VI. Procédure pour l'identification des doublons, pour la conformité et la cohérence

1. Principes

a) **Doublons**

Cette étape consiste à vérifier, à l'entrée de chacune des plateformes, que la donnée ou métadonnée (jeu de données et dispositif de collecte) n'est pas déjà présente dans le système. Lorsque la donnée ou métadonnée est déjà présente dans le système avec une date différente de mise à jour, il peut s'agir d'une donnée ou métadonnée venant mettre à jour la précédente ou d'un doublon généré par des flux croisés au sein du SINP.

Dès lors que des doublons sont identifiés, les plateformes doivent en avertir le fournisseur afin de définir de la marche à suivre et clarifier l'organisation des flux de données avec les producteurs. Chaque plateforme dispose de toute liberté pour bancariser ces données doublonnées dans son propre système de gestion. Dans l'attente de la définition des règles de gestion et de mise en circulation de ces doublons par l'architecture du SINP, le GT Validation préconise que leur identification donne lieu à l'ajout d'un attribut permettant de tracer leur identification.

b) **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 les aspects physiques et conceptuels des standards, c'est-à-dire 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.

Si l'une de ces règles n'est pas respectée, la donnée ou la fiche de métadonnées est qualifiée de non conforme.

Les plateformes régionales et thématiques sont responsables de la mise en conformité des DEE et de leurs métadonnées qu'elles produisent à partir des données des producteurs. Seules les DEE et métadonnées associées conformes circulent dans le système.

Cas où un champ obligatoire des standards ne répond pas aux exigences de conformité :

La donnée ou l'élément de métadonnées est considéré comme non conforme et le producteur en est informé via le rapport de conformité. Il doit alors corriger¹⁰ les informations non conformes pour que les données puissent être véhiculées dans le SINP.

Cas où un champ obligatoire conditionnel ne répond pas aux exigences de conformité :

¹⁰ Voir IX. Mises à jour

La donnée ou élément de métadonnées est considéré comme non conforme et le producteur en est informé via un rapport de conformité. Il doit alors corriger¹¹ les informations non conformes pour que les données puissent être véhiculées dans le SINP.

Cas où un champ facultatif ne correspond pas aux exigences de conformité :

La donnée ou l'élément de métadonnées, tel quel, est considéré comme non conforme. Les attributs non conformes 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.

c) **Cohérence**

Cette étape consiste à vérifier la cohérence entre les informations transmises (intra ou inter fichier) pour les données et les métadonnées.

Le contrôle est relatif au contenu des informations d'un point de vue de la cohérence avec des règles préétablies par le groupe de travail. Si l'une de ces règles n'est pas respectée, la donnée ou la fiche de métadonnées est qualifiée de non-cohérente.

Les plateformes régionales et thématiques sont responsables de la cohérence des DEE et de leurs métadonnées qu'elles produisent à partir des données et des métadonnées des producteurs. Seules les DEE et métadonnées associées cohérentes entrent et circulent dans le système.

Cas où un champ obligatoire des standards ne répond pas aux exigences de cohérence :

La donnée ou l'élément de métadonnées est considéré comme incohérent et le producteur en est informé via un rapport de cohérence. Il doit alors corriger les informations incohérentes pour que les données puissent être véhiculées dans le SINP.

Cas où un champ obligatoire conditionnel ne répond pas aux exigences de cohérence :

La donnée ou l'élément de métadonnées est considéré comme incohérent et le producteur en est informé via un rapport de cohérence. Il doit alors corriger les informations incohérentes pour que les données puissent être véhiculées dans le SINP.

Cas où un champ facultatif ne correspond pas aux exigences de cohérence :

La donnée ou l'élément de métadonnées, tel quel, est considéré comme incohérent. Les attributs incohérents sont déportés dans un champ dédié, les champs incohérents sont mis à blanc ou avec la valeur par défaut requise. La donnée devient cohérente et peut circuler dans le système.

¹¹ Voir IX. Mises à jour

2. Protocole de doublons, de conformité et de cohérence

Un protocole commun à l'ensemble des plateformes sera établi nationalement (dans le cadre de la poursuite des travaux du GT) pour définir des règles précises à appliquer pour les contrôles de doublons, de conformité et de cohérence.

L'application d'une méthodologie commune a pour objectif d'éviter les blocages :

- ➔ à l'entrée de la plateforme nationale lorsque les données sont entrées dans le SINP par une plateforme régionale,
- ➔ à l'entrée d'une plateforme régionale lorsque les données sont entrées de le SINP par la plateforme thématique « Occurrences de taxon ».

L'établissement d'un protocole commun permet également d'alléger les travaux des plateformes en mutualisant les ressources, tout particulièrement en cas de mises à jour¹² lors de l'édition d'une nouvelle version d'un standard d'échange de données ou de métadonnées.

L'outil de plateforme GINCO¹³ pourra ainsi se baser sur ce protocole pour mettre en œuvre les développements nécessaires à l'automatisation de ces contrôles.

Ce document devra être versionné : toute modification devra faire l'objet d'une nouvelle publication avec un numéro de version différent.

Anticipant la publication de ce protocole, les règles pressenties pour le protocole (décliné par type de contrôle) sont explicitées ci-après. Ces règles sont fournies à titre d'exemples et ne sauraient être considérées comme exhaustives.

En cas de difficultés ou de demandes de précision sur ces règles :

- les plateformes régionales pourront se référer aux GT standard de données, GT standard de métadonnées ou directement auprès du GT Validation
- les producteurs régionaux sont invités à se rapprocher de leur plateforme régionale
- les producteurs supra-régionaux ou nationaux sont invités à se rapprocher de la plateforme thématique occurrences de taxons

¹² Voir IX. Mises à jour

¹³ Gestion d'Informations Naturalistes Collaborative Ouverte

a) Protocole d'identification des doublons

Les règles minimales de détection des doublons consistent à effectuer des vérifications sur l'identifiant unique et sur le couple identifiant unique/date de mise à jour de la DEE.

Il pourra être mis en place un contrôle de redondance plus large (incluant d'autres attributs), par exemple à partir de l'identifiant origine ou des attributs relatifs à la date de l'observation, au lieu, au taxon ou encore à l'observateur.

La gestion (suppression, gel, etc.) et la circulation des doublons au sein du SINP sont des questions importantes qui relèvent du GT Architecture.

b) Protocole de conformité

Les règles de conformité se basent sur les versions des standards de données et de métadonnées utilisées par la plateforme qui réalise le contrôle. **Ces versions doivent être celles publiées par les GT afférents.**

Les règles de conformité pour les métadonnées sont actuellement décrites dans le document produit par le GT Standard de métadonnées (Description du profil de métadonnées des ressources du Système d'Information sur le Nature et le Paysage (SINP)¹⁴.

Les règles de conformité pour les données sont décrites dans les documents « Standard de données SINP – Occurrence de taxon », « Identifiant permanent de la DEE » et « Le schéma GML »¹⁶.

c) Protocole de cohérence

Sur les attributs temporels :

Élément de :	Règle
Données	La date de début de l'observation doit être < ou = à la date de fin d'observation. Les dates de début et de fin d'observation doivent être inférieures ou égales à la date de création de la DEE.
Métadonnées	La date de début du jeu de données doit être < ou = à la date de fin du jeu de données. Les dates de début et de fin du jeu de données doivent être inférieures ou égales à la date du jour.
Métadonnées	La date de début du dispositif de collecte doit être < ou = à la date de fin du dispositif de collecte.

¹⁴ Voir site Nature France : <http://www.naturefrance.fr>

	La date de début du dispositif doit être inférieure ou égale à la date du jour.
Entre données et métadonnées	Les dates des données doivent être comprises dans l'emprise temporelle décrite pour le jeu de données.

Sur les attributs géographiques :

Élément de :	Règle
Données	Si la donnée véhicule un nom de commune et un code commune, alors les attributs doivent être cohérents entre eux. Cohérence entre codeCommuneValue et nomCommuneValue
Données	Si un code d'espace naturel est renseigné, alors ce code doit être retrouvé dans la liste des valeurs du référentiel spécifié (selon le type d'espace naturel). Cohérence entre TypeEN et CodeEN.
Données	Au moins une donnée du jeu de données transmis doit être située sur le territoire français (Métropole et Outre-Mer).
Données	L'objet géographique d'acquisition doit être cohérent avec la ou les communes associées (si commune(s) renseignée(s)).
Données	L'objet géographique d'acquisition doit être cohérent avec le ou les espace(s) naturel(s) mentionnés (si espace(s) renseigné(s)).
Données	L'objet géographique d'acquisition doit être cohérent avec la ou les maille(s) mentionnée(s) (si maille(s) renseigné(s))
Données	L'objet géographique d'acquisition doit être cohérent avec la ou les masse(s) d'eau mentionnée(s) (si masse(s) d'eau renseignée(s))
Données	Si transmission commune + Espace naturel : les deux informations doivent être cohérentes (test de contiguïté des objets).
Données	Si transmission commune + maille : les deux informations doivent être cohérentes (test de contiguïté des objets).
Données	Si transmission commune + masse d'eau : les deux informations doivent être cohérentes (test de contiguïté des objets).
Données	Si transmission espace naturel + maille : les deux informations doivent être cohérentes (test de contiguïté des objets).
Données	Si transmission espace naturel + masse d'eau : les deux informations doivent être cohérentes (test de contiguïté des objets).
Données	Si transmission maille + masse d'eau : les deux informations doivent être cohérentes (test de contiguïté des objets).
Métadonnées	L'emprise géographique du jeu de données doit intersecter le territoire français

Métadonnées	Le système de coordonnées ou de projection utilisé doit être cohérent avec l'emprise géographique (<i>Ex : si bounding box= métropole, alors ESPG in ('4326','27582', '2154')</i>)
Entre données et métadonnées	Les objets géographiques (incluant la prise en compte de la précision), les mailles, communes ou espaces transmis au sein des données doivent être inclus dans l'emprise géographique du jeu de données.

Sur les attributs taxonomiques :

Élément de :	Règle
Données	Le CdNom et le CdRef doivent être cohérents (si renseignés) et fidèles à la version de Taxref utilisée. C'est-à-dire que le CdNom renseigné est bien rattaché au CdRef cité dans la version du référentiel citée.
Métadonnées	L'emprise taxonomique pour le niveau du jeu de données doit être égale ou emboîtée inférieurement dans la cible taxonomique du dispositif de collecte. <i>Ex : Type de données « dispositif »= amphibiens, alors type de données « données » = amphibiens ou infra (anura, etc.)</i>
Entre données et métadonnées	Les CdRef des données doivent être inclus dans l'emprise taxonomique déclarée au niveau du jeu de données.

Autres attributs :

Élément de :	Règle
Données	Si les dénombrements sont renseignés, alors DenombrementMin doit être < ou = à DenombrementMax
Données	Si un code habitat est renseigné, alors ce code doit être retrouvé dans la liste des valeurs du référentiel Habitats cité. Cohérence entre codeRefHabitatValue et les valeurs de ce référentiel.

Élément de :	Règle
Données et annuaire ¹⁵	L'organisme gestionnaire des données du jeu doit être dans l'annuaire du SINP.
Données et annuaire	L'organisme responsable de la standardisation des données du jeu doit être dans l'annuaire du SINP.
Métadonnées et annuaire	Les acteurs déclarés dans les métadonnées (niveau jeu de données et dispositif de collecte) doivent être déclarés dans l'annuaire du SINP.

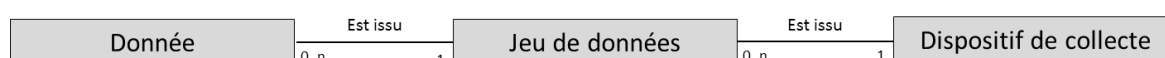
Élément de :	Règle
Entre données et métadonnées (jeu de données)	L'identifiant de jeu de données déclaré dans la donnée doit correspondre à un identifiant de jeux de données existants.

¹⁵ Lorsque l'annuaire sera réalisé et sous réserve que ces attributs soient inclus dans son périmètre/Définition : Voir XV. Glossaire

Métadonnées (entre jeu de données et dispositif de collecte)	L'identifiant du dispositif de collecte déclaré dans un jeu de données doit correspondre à un identifiant de dispositifs de collecte existants.
--	---

d) Dépendance entre données et métadonnées (jeux de données et dispositifs de collecte)

Pour circuler au sein du SINP, les DEE doivent être associées à une fiche de jeu de données qui doit être associée à une fiche de dispositif de collecte. Une fiche de dispositif de collecte peut être décrite sans fiche de jeu de données associée et une fiche de jeu de données peut être décrite sans données associées¹⁶.



De fait, lorsqu'une fiche de dispositif de collecte n'est pas cohérente/conforme, cela entraîne la non diffusion de la fiche de ce dispositif et des fiches de jeux de données ainsi que celle des données qui y sont rattachés.

De la même manière, lorsqu'une fiche de jeu de données est incohérente/non-conforme, la fiche du jeu de données et les données qui y sont attachées ne sont pas diffusées.

A contrario, si toutes les données d'un même jeu sont incohérentes/non-conformes, ces données ne sont pas diffusées mais la fiche de métadonnées du jeu ainsi que celle du dispositif de collecte associé peuvent circuler. De la même manière, si tous les jeux de données associées à un dispositif de collecte sont incohérents/non-conformes, les fiches des jeux de données ne sont pas diffusées mais la fiche du dispositif de collecte peut quant à elle être diffusée au sein du SINP.

3. Référentiels

Les contrôles de conformité et de cohérence vont nécessiter de disposer de l'ensemble des référentiels utilisés pour l'échange de données et de métadonnées, donc de tous les référentiels permis par les standards de données et de métadonnées. Par exemple (non exhaustif) :

- Maillages géographiques nationaux 10x10 Km pour l'ensemble des territoires français (métropole et outre-mer)
- Référentiels INSEE des communes et des départements et couches géographiques IGN (BD Topo)
- Référentiel des taxons TAXREF

¹⁶ Voir GT métadonnées du SINP

- Référentiel des habitats HABREF
- Référentiel des espaces naturels
- Annuaire des organismes et des plateformes
- Etc.

Tous ces référentiels sont ou seront diffusés par la plateforme nationale.

4. Rapport de doublons, de conformité et de cohérence

Les fournisseurs devront être informés de l'identification de doublons sur les éléments (données, jeux de données ou dispositifs de collecte) qu'ils ont transmis. Ils devront être en mesure d'identifier les éléments en question (par la transmission des identifiants des éléments concernés) et être informés de l'impact de ces doublons sur la circulation au sein du système (blocage ou non). La forme de ce retour au fournisseur (rapport spécifique, email, via site web, etc.) n'est pas imposée par le présent GT.

Les plateformes devront, avec les producteurs de données concernés, définir la marche à suivre et clarifier l'organisation des flux (à savoir une seule livraison au SINP : soit par la plateforme régionale, soit par la plateforme thématique occurrences de taxons).

Les contrôles de conformité et de cohérence doivent donner lieu à un rapport transmis au producteur et ce, quel que soit le résultat des contrôles.

Ce rapport transmis aux producteurs est établi par la plateforme dans laquelle le producteur a versé ses données ; il s'agit donc d'un rapport régional ou thématique. La forme de ce retour au fournisseur (rapport spécifique, email, via site web, etc.) n'est pas imposée par le présent GT.

Le rapport de conformité et de cohérence doit contenir, à minima, la mention des champs contenant des données non conformes ou incohérentes et la raison de la non-conformité ou de l'incohérence. Il doit renseigner le producteur sur l'impact de cette non-conformité ou incohérence pour la DEE (en cas notamment de mise à blanc des champs) et sur sa circulation (donnée bloquée qui ne circule donc pas dans le système). Il doit permettre au producteur de retrouver les données et les métadonnées non conformes ou incohérentes (en indiquant notamment l'identifiant producteur/origine des données et métadonnées concernées).

Suite à la réception de DEE issues des plateformes régionales et thématiques, la plateforme nationale établit un contrôle subsidiaire qui donne lieu à un rapport national de conformité et de cohérence transmis aux plateformes sources.

Ce rapport doit contenir, à minima, la mention des champs non conformes ou incohérents. Il doit renseigner la plateforme source sur l'impact de cette non-conformité pour la DEE (en cas notamment de mise à blanc des champs) et sur sa circulation (donnée bloquée qui ne circule donc pas dans le système). Il doit permettre à la plateforme de retrouver les données non

conformes (en indiquant notamment l'identifiant permanent des données et métadonnées concernées).

VII. Procédure pour la validation scientifique

1. Principes

La validation est un processus d'expertises visant à renseigner sur la fiabilité, dans le sens du degré de confiance que l'on peut accorder à la donnée. Ce processus fait intervenir des bases de connaissance et/ou directement des experts. Dans le cadre du SINP, cette étape ne s'applique pas aux métadonnées associées aux données.

La validation ne doit pas être une opération bloquante pour l'échange de DEE. L'échange entre la plateforme régionale ou thématique occurrences de taxons et la plateforme nationale au sein du SINP a lieu quel que soit le niveau de validité attribué aux données par les différentes plateformes et par le producteur. La procédure de validation est donc facultative pour la circulation des données dans le système.

Le résultat des opérations de validation scientifique (lorsqu'elles ont eu lieu) doit être véhiculé au sein de la DEE. Les opérations en elles-mêmes peuvent avoir lieu indifféremment sur une donnée source (DS) ou sur une DEE. Cela implique que toutes les informations de la DS ayant servi à la validation scientifique doivent être également présentes dans la DEE de manière à ce que les plateformes procèdent à la validation sur la même base d'informations.

Toutes les données peuvent être soumises à validation si l'organisation régionale permet la mise en place d'une validation avant transformation des DS en DEE mais il est rappelé que seules les données conformes et cohérentes, quel que soit le niveau de validation, seront échangées entre les plateformes SINP.

Le besoin en matière de « périmètre » de validation, à savoir la précision de la validation réalisée sur la donnée en termes d'informations considérées, sera différent en fonction de l'usage visé.

Exemple : la validation de la présence d'un taxon à une date sur un lieu suffira pour réaliser une étude de répartition mais ne suffira pas pour permettre une analyse d'abondance si les informations concernant les effectifs n'ont pas été validées.

Face à ce constat, le GT Validation identifie deux « périmètres » de validation scientifique possibles :

- ➔ Une validation de **périmètre minimal** qui consiste à appliquer une validation scientifique sur les informations relatives au taxon cité, à la date et au lieu sans

considérer la validation applicable aux champs tiers (effectifs, statut biologique, etc.).
C'est le niveau minimum de validation accepté dans le SINP.

- Une validation de **périmètre élargi** qui consiste à appliquer une validation scientifique sur les informations relatives au taxon cité, à la date et au lieu **et** à des champs tiers. Les champs tiers considérés pour la validation de périmètre élargie devront être décrits dans les protocoles de validation diffusés par la plateforme.

Ces deux périmètres (minimal et élargi) peuvent être portés communément sur la DEE pour les validations régionales et nationales.

Cette distinction entre les deux périmètres n'est pertinente que si le résultat de la validation est associé à un protocole de validation qui explicite les attributs considérés.

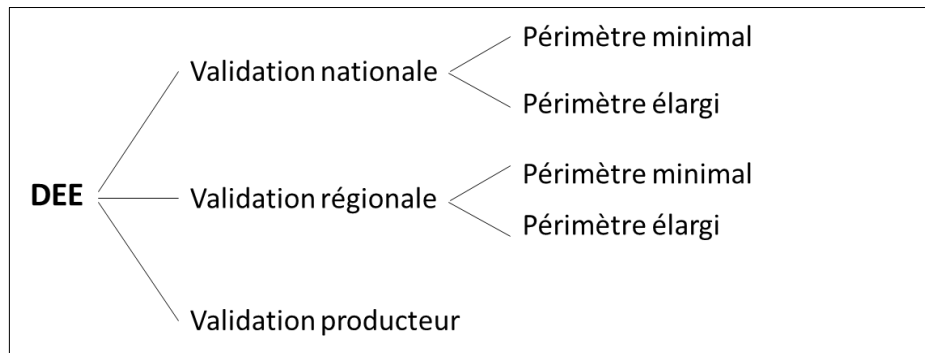
Afin de respecter le principe de non-intrusion dans les systèmes de validation des producteurs, la démarche SINP ne peut imposer à ces derniers de renseigner ces informations et de les rendre accessibles. Nous savons également d'expérience que les protocoles utilisés sont rarement disponibles et/ou exploitables (peu documentés, peu normalisés, parfois incomplets et, surtout, rarement bancarisés dans les bases sources). La distinction entre ces périmètres n'est donc actuellement pas applicable à la validation du producteur.

La validation de périmètre élargi permet notamment au SINP de s'appuyer sur des démarches de validation existantes dans le cadre de programmes spécifiques (inventaires nationaux, etc.). La mise en place d'une validation de périmètre élargi devra être décrite dans les protocoles de validation utilisés par la plateforme.

Dans le cas où une validation de périmètre élargi permettrait également de pré-renseigner le niveau de validité pour le périmètre minimal (exemple : une donnée certaine selon le protocole élargi engendrera un niveau de validité selon le protocole minimal de la même valeur), alors ces passerelles entre les périmètres de validation devront être décrites au sein des protocoles de la plateforme qui les met en œuvre.

2. Règles de validation

Les DEE circulant au sein du système peuvent porter trois attributs de niveaux de validité, résultats des validations producteur, régionale et nationale. Les résultats des validations régionales et nationales peuvent être déclinés selon s'il s'agit d'un périmètre minimal ou élargi.



La mise en œuvre de la validation « producteur » est de la responsabilité de la structure productrice ou d'un tiers (organisme ou personne) mandaté par elle. Elle est facultative. Ce niveau pourra être notamment utilisé pour mentionner un doute sur la fiabilité de la donnée transmise.

Les listes de valeurs pour le niveau de validité producteur sont laissées libres¹⁷. Le GT validation préconise cependant que les valeurs utilisées soient littérales et suffisamment explicites pour les usagers.

La mise en œuvre de la validation régionale est coordonnée par le responsable de la plateforme régionale qui s'appuie sur le réseau d'experts régional.

La mise en œuvre de la validation nationale est coordonnée par le MNHN porteur de la plateforme nationale; elle s'appuie pour cela sur des informations nationales de référence et sur des réseaux nationaux d'experts, définis selon les groupes taxonomiques et des aspects biogéographiques (en particulier pour l'outre-mer).

Seul le producteur (ou organisme mandaté par lui) dispose d'un droit d'écriture dans les champs de validation producteur.

Seule la plateforme régionale (ou organisme/pôle mandaté par elle) dispose d'un droit d'écriture dans les champs de validation régionale.

¹⁷ Voir X. Exigences fonctionnelles

Seule la plateforme nationale (ou organisme/pôle mandaté par elle) dispose d'un droit d'écriture dans les champs de validation nationale.

Les plateformes régionales et la plateforme nationale doivent mettre en place une procédure de validation scientifique, rendu accessible au public sous la forme d'un rapport documenté et versionné accessible sur le site de chaque plateforme. Les données ayant subi la procédure de validation porteront parallèlement l'url du protocole mis en œuvre.

Le choix de sectorisation des protocoles de validation selon, entre autres, des découpages taxonomiques, est laissé à la libre appréciation de chacune des plateformes.

Cette organisation permet de ne pas être intrusif dans les protocoles de validation régionaux. Cependant, il est recommandé aux plateformes régionales de se rapprocher des régions frontalières afin de rechercher une cohésion inter-régionale de manière à limiter les disparités et l'effet de « frontières » administratives. Il est également recommandé aux plateformes de rechercher une cohérence entre les protocoles mis en place au sein de chacune d'elles (en cas de sectorisation par groupe taxonomiques par exemple).

De la même manière, il sera utile de rechercher une cohérence entre les validations régionales et la validation nationale. Une analyse des disparités régionales et nationales à l'issue d'un premier « tour » de validation permettrait d'amorcer ce travail de mise en cohérence entre les différents protocoles de validation au sein du SINP (voir partie perspectives).

Le nombre d'experts et leur disponibilité étant limités au regard du nombre conséquent de données à valider, le GT Validation préconise aux plateformes d'organiser la validation en deux phases complémentaires : une première phase de validation automatique permettant de pré-valider les données selon des critères automatisables, puis, une validation manuelle portée par le réseau d'experts. Cela permet aux experts d'intervenir prioritairement sur les données ayant été jugées douteuses ou à risque par la validation automatique. Les experts doivent cependant avoir toute liberté pour intervenir sur les autres données quel que soit le résultat de la validation automatique.

Si le stockage du résultat des validations automatiques et manuelles est laissé à l'appréciation des plateformes en charge, seul le résultat retenu (final) de validation circulera au sein du système sur la DEE (sauf pour la validation producteur qui est libre et peut, de fait, porter plusieurs informations).

Le choix de primauté de l'un de ces contrôles sur l'autre (en général primauté du contrôle manuel quand il est réalisé après le contrôle automatisé ou d'une méthode combinatoire qui consiste en une validation qui fait intervenir, de manière combinée, une validation automatique et une validation manuelle) pour renseigner, in fine, le niveau de validité retenu dépend de la procédure choisie et relève donc de la responsabilité des plateformes en charge de la validation. Ces choix devront être explicités dans le protocole de validation scientifique de chaque plateforme.

a) Validation automatique

La validation automatique fait appel à des couches de référence, des modèles probabilistes (à la condition que ces modèles aient été publiés et calibrés avec des données indépendantes du jeu de donnée à valider), des bases de connaissance ou à des référentiels pré-existants ou spécialement dédiés à l'opération de validation¹⁸.

Ceux-ci doivent être réalisés à une échelle cohérente avec celle des données que l'on contrôle par cette voie (résolution taxonomique, temporelle, spatiale, et, le cas échéant, attributs complémentaires - statut biologique par exemple).

Exemple : ne pas chercher à valider une donnée d'oiseau nicheur de 2015 sur la base d'une couche de référence d'oiseau hivernant de 1950.

La documentation du protocole doit renseigner sur les couches de référence utilisées, les bases de connaissance ou les référentiels utilisés.

Les seuils de validation automatique (proximité géographique ou éventuellement temporelle avec les informations de références) de la donnée sont très variables en fonction des groupes taxonomiques considérés. La définition de ces seuils, pouvant être déclinés par groupes taxonomiques, doit donc reposer sur un choix d'expert.

Exemple : validation d'une donnée d'oiseau nicheur s'il est situé dans une maille de présence certaine ou probable du dernier Atlas des oiseaux nicheurs ; validation d'une donnée d'araignée si elle est dans un département où l'espèce a été signalée dans les 50 dernières années ; validation d'une donnée de mammifère si elle est située à moins de 5 km de la couche de répartition connue.

Cela doit se traduire par une explicitation et la définition des seuils retenus au sein de la documentation décrivant le protocole de validation réalisée par chaque plateforme.

La validation automatique peut également être réalisée par congruence par rapport à d'autres données du système. Il s'agit d'utiliser, par anticipation, des données qui n'auraient pas encore été prises en compte pour l'élaboration des couches de référence.

Auquel cas cette opération devra être explicitement décrite au sein de la documentation du protocole. Elle sera à manipuler avec précaution car une première erreur de validation pourra entraîner, de fait, d'autres erreurs de validation de manière exponentielle.

Une fois le protocole décrit, le GT Validation préconise que les modèles automatiques soient testés afin d'en évaluer la pertinence en mettant en place un ensemble d'indicateurs de performance. Ces modèles peuvent par exemple être testés à partir de données de jeux fictifs à erreurs connues.

¹⁸ Voir VII.3. Référentiels et bases de connaissance

b) Validation manuelle

La validation manuelle fait intervenir des experts mandatés par la plateforme en charge de la 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. Ils doivent avoir accès à l'ensemble des informations portées au niveau des données et au niveau des métadonnées nécessaires à assurer leur expertise.

Ils ont la possibilité de recontacter l'observateur, le déterminateur ou l'organisme producteur de manière à demander des éléments complémentaires.

Lorsqu'un contact avec le producteur de la donnée ou de la métadonnée est nécessaire, celui-ci est réalisé, dans la mesure du possible, au plus près de la production de données (par le premier point d'entrée de la donnée dans le SINP). Ce principe permet de limiter les sollicitations multiples des producteurs par des interlocuteurs multiples. Cependant, pour la validation scientifique, des sollicitations directes sont tolérées sous réserve que celles-ci n'interfèrent pas avec les sollicitations réalisées par la plateforme d'entrée dans le système.

De manière à éviter les sollicitations multiples, le contact avec le producteur (nom, prénom, organisme et date de contact) devra être retranscrit au sein des données¹⁹.

c) Critères et paramètres

Globalement, la validation scientifique est pratiquée en prenant en compte un ensemble de paramètres (éléments utilisés pour valider intrinsèquement la donnée) et de critères de sélection (éléments permettant d'orienter l'effort de validation sur des données spécifiques ou nécessitant une attention particulière). Certains éléments pouvant être à la fois critère et paramètre.

De manière synthétique, dans le cadre du GT Validation, les paramètres et critères suivants (non exhaustifs) ont été identifiés. Les plateformes SINP peuvent utiliser des paramètres et des critères de la présente liste pour mettre en place leurs protocoles de validation.

Libellé de l'élément	Commentaire	Paramètre ou critère
Congruence	La congruence comprend : -La validation par rapport à des couches de référence (distribution, phénologie, habitat...) validées par expertise ; -La validation par rapport à d'autres données avant intégration de celles-ci dans les couches de référence. Ces couches ou données peuvent être utilisées telles quelles ou après une étape de	Paramètre de validation et critère de sélection.

¹⁹ Voir X. Exigences fonctionnelles

	modélisation pour, par exemple, définir une probabilité de présence ou une enveloppe de distribution potentielle.	
Rareté des données (en termes de fréquence d'observation).	Un niveau de rareté peut être calculé automatiquement en fonction des données présentes dans le système. Il peut être calculé relativement à un groupe taxonomique, éventuellement sectorisé par année, etc.	Critère de sélection
Difficulté de détermination.	La difficulté de détermination tient en premier lieu au taxon, mais également à la zone géographique concernée, au protocole de collecte employé et à la méthode de détermination.	Paramètre de validation et critère de sélection.
Présence d'une preuve.	Disponibilité de la preuve en elle-même ou existence d'une preuve.	Paramètre de validation et critère de sélection.
Formation des observateurs.	Degré d'expertise lié à l'expérience ou aux formations dispensées aux observateurs dans le cadre d'un programme spécifique.	Paramètre de validation.
Validation producteur.	Présence d'une validation du producteur sur les données transmises. Paramètre à utiliser avec précaution pour garantir l'indépendance des validations.	Paramètre de validation.
Observateur et déterminateur distincts.	Détermination déléguée ou reprise par une personne dédiée.	Paramètre de validation.
Validation collective.	Concept de double validation ou/ et de validation collective. Plusieurs experts analysent les données pour statuer ou mise en place d'une validation par une communauté d'acteurs.	Paramètre de validation.

Critères et paramètres utilisables (non exhaustifs) dans les protocoles de validation mis en place par les plateformes

3. Référentiels et bases de connaissance

Le GT validation a identifié, via un questionnaire de besoin en matière de validation, un certain nombre de référentiels et de bases de connaissance utiles à l'élaboration des protocoles de validation des plateformes ou organismes en charge. Ce guide n'a pas vocation à imposer l'usage de tel ou tel référentiel ou base de connaissance de la liste proposée.

Ils sont énoncés ici de manière indicative, non-exhaustive et sans juger de leur disponibilité/existence (autant au niveau de l'emprise géographique (sur chacune des régions) que taxonomique (pour l'ensemble des taxons)) afin de proposer des pistes de travail aux plateformes qui n'auraient pas encore de procédures de validation ou qui souhaiteraient compléter une procédure déjà en œuvre.

Référentiels et bases de connaissances relatifs au taxon (notamment aux traits de vie) et à son observation

- TAXREF (statuts biogéographiques)
- Liste des taxons ayant fait l'objet de changements taxonomiques complexes et date de ces changements (*exemple : « splittage » d'une espèce en trois lors d'une révision publiée en 1998*)
- Liste des communes ayant fait l'objet de remaniements de type fusion ou splittage.
- Référentiels de difficulté de détermination
- Référentiels de rareté des données ou des espèces
- Référentiels des protocoles d'acquisition de données, des méthodes de terrain et du matériel utilisé
- Couches de référence de répartition des taxons pouvant être déclinées selon différentes résolutions géographiques (commune, maille, département, etc.) et pouvant être constituées des données d'observations ou enrichies par une approche experte ou statistique d'estimation de la niche et de la distribution potentielle (approche utilisant elle-même des variables issues d'autres référentiels comme le climat et l'occupation du sol par exemple)
- Bases de connaissances sur les erreurs et confusions fréquentes
- Bases de connaissances sur la phénologie et sur les altitudes limites des taxons
- Bases de connaissances sur les plantes-hôtes (ou autres types d'interactions) et couche de distribution de ces plantes-hôtes

Référentiels de contexte

- Référentiel des masses d'eau (SANDRE)
- Référentiels des espaces naturels protégés
- Couche d'occupation du sol, couche de végétation (*exemple : carte des séries de végétation du CNRS*), ou cartes d'habitats

- Couches climatiques (*brutes ou de synthèse comme les zonages bioclimatiques*)
- Couches pédologiques et géologiques
- Couches hydrographiques
- Couches de relief et de bathymétrie

4. Protocole de validation

Chaque plateforme SINP en charge de la validation (plateformes régionales et plateforme nationale) doit décrire le protocole de validation qu'elle applique sur les données au sein d'un document rendu accessible aux différents usagers du SINP.

Les CSRPN peuvent être mandatés par les responsables des plateformes régionales pour réaliser les protocoles de validation ou donner un avis scientifiques sur les protocoles mis en place. Le présent guide n'impose cependant pas la mobilisation des CSRPN mais donne toute liberté aux plateformes pour intégrer ces instances au sein de leur organisation interne.

De la même manière, les protocoles de validation établis au niveau national sont de la responsabilité du MNHN qui pourra s'appuyer le cas échéant sur des instances scientifiques nationales.

Ce document pourra être décliné par groupes taxonomiques (ou autres déclinaisons) selon la volonté de chaque plateforme. Ce ou ces protocoles devront être versionnés pour permettre le suivi des mises à jour²⁰.

Ce document doit décrire et renseigner sur :

- ➔ Les organismes ayant été mandatés par la plateforme le cas échéant pour réaliser ces opérations et leur périmètre d'action respectif (*exemple : flore pour un CBN, etc.*).
- ➔ La démarche de validation, qu'il s'agisse de la partie automatique ou de la partie manuelle.
- ➔ Les règles d'attribution du niveau de validité retenu en fonction des résultats des validations automatiques et manuelles.
- ➔ Les champs tiers considérés pour la validation de périmètre élargi.
- ➔ Les critères et les paramètres pris en compte et les seuils retenus pour ces derniers.
- ➔ Les outils techniques utilisés.
- ➔ Les référentiels, bases de connaissance ou couches de référence utilisés (incluant leurs résolutions (commune, maille, etc.)).
- ➔ La méthode d'attribution des niveaux de validité à partir de ces éléments.
- ➔ Les choix de gestion pour les données d'un regroupement²¹
- ➔ Les règles de mises à jour appliquées pour la révision des niveaux de validité.

²⁰ Voir IX. Mises à jour

²¹ Voir VIII. Regroupements de données

- La procédure mise en place pour la transmission des rapports aux producteurs et leur périodicité.

La description des référentiels, bases de connaissance ou couches de référence utilisés et la méthodologie de leur conception devront être rendues accessibles aux différents usagers du SINP.

Afin d'éviter la sortie d'une nouvelle version du protocole à chaque mise à jour d'un référentiel utilisé pour l'étape de validation, le GT Validation préconise de déporter les informations concernant les versions des référentiels en dehors du document de protocole sous la forme d'une table relationnelle rendue accessible aux usagers sur la plateforme.

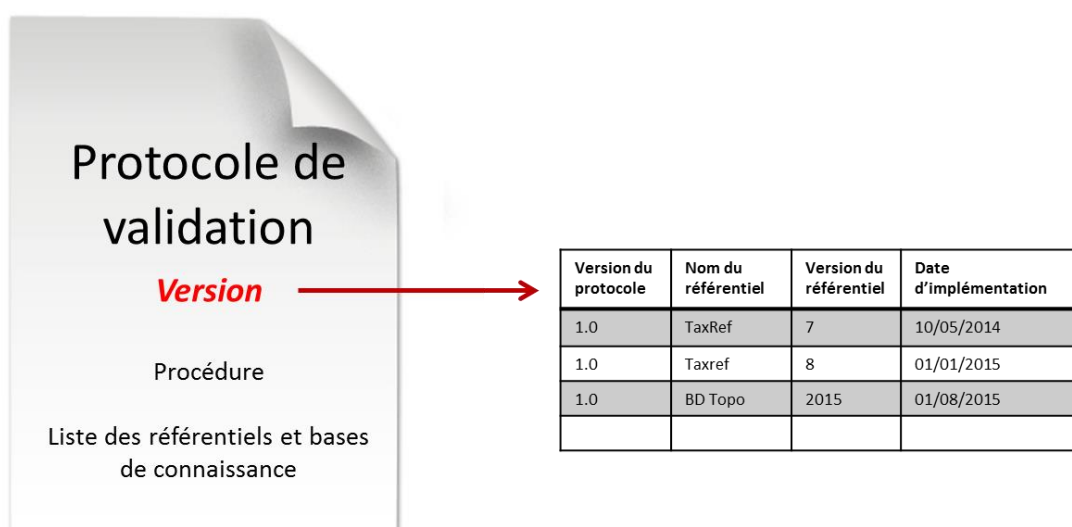


Schéma descriptif de l'association protocole/versions des référentiels

Cette table devra contenir à minima :

- un champ de lien entre le document et la table (version)
- le libellé du référentiel
- la version du référentiel
- la date d'implémentation de la nouvelle version du référentiel dans la mise en œuvre de l'étape de validation
- l'url de consultation du guide méthodologique de chaque référentiel ou base de connaissance ou sa référence bibliographique

5. Niveaux de validité

Les valeurs pour les niveaux de validité résultant de l'opération de validation scientifique du producteur sont laissées libres à celui-ci.

Cependant, il est proposé aux producteurs l'utilisation des mêmes libellés que ceux retenus pour la validation scientifique réalisée par les plateformes du SINP. Quel que soit le choix retenu par le producteur, les libellés utilisés devront être suffisamment explicites. Il est donc recommandé de ne pas utiliser de codage spécifique ou d'abréviation.

Les niveaux de validité retenus dans le SINP pour le résultat des opérations de validation scientifique régionale et nationale sont les suivants :

Valeurs ²²	Libellés
1	Certain - très probable
2	Probable
3	Douteux
4	Invalide
5	Non réalisable
	Non évalué

Chaque libellé correspond à une définition qui est fonction du processus utilisé, à savoir manuel/combinatoire ou automatique.

Pour mémoire:

La validation scientifique dite manuelle (nonobstant le fait qu'une interface informatique peut être utilisée) consiste en une validation des informations faisant appel à une expertise directe (avis d'expert suite à l'analyse des informations transmises). Ce processus est réalisé par examen « manuel » par un ou des experts dans le cadre du SINP.

La validation scientifique dite automatique consiste en une validation faisant appel à des résultats d'expertise préalables (des référentiels, des bases de connaissance, etc.). Ce processus est réalisé de manière automatique (informatique) dans le cadre du SINP.

*La validation scientifique dite combinatoire consiste en une validation faisant intervenir, de manière combinée, une validation automatique **et** une validation manuelle.*

²² Sous réserve des choix techniques d'implémentation du GT Standard de données

Libellés	Définitions	
	Processus automatique	Processus manuel ou combinatoire
Certain - très probable	<p>La donnée présente un haut niveau de vraisemblance (très majoritairement cohérente) selon le protocole automatique appliquée.</p> <p>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.</p>	<p>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.</p> <p>La validation a été réalisée <u>notamment</u> 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.</p>
Probable	<p>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.</p>	<p>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.</p>
Douteux	<p>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).</p>	<p>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.</p>
Invalide	<p>La donnée ne concorde pas selon le protocole automatique appliqué. Elle présente au moins une discordance majeure sur</p>	<p>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</p>

	<p>un des critères jugés les plus importants ou la majorité des critères déterminants sont discordants.</p> <p>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.</p>	<p>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.</p>
Non réalisable	<p>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.</p> <p>Notamment:</p> <ul style="list-style-type: none"> -Etat des connaissances du taxon insuffisantes -Informations insuffisantes sur l'observation 	
Non évalué	<p>Niveau initial ou temporaire.</p> <p>La donnée n'a pas été soumise à l'opération de validation ou l'opération n'est pas encore terminée (validation en cours). Elle n'est donc pas évaluée à un temps précis défini par la date de validation.</p>	

La définition plus fine des niveaux (seuils) est du ressort de chaque plateforme qui décrira alors les règles d'attribution de ces niveaux au sein de son protocole de validation. **Ces règles ne doivent cependant pas dénaturer la définition originale (générique) définie par le GT Validation de manière à ne pas engendrer de disparités notables entre les plateformes.**

Le niveau de validité correspond à un état des connaissances à un moment donné et doit donc toujours être associé à une date de validation pour la validation régionale et pour la nationale (voir exigences fonctionnelles). Cette date est facultative pour la validation du producteur mais son renseignement est cependant vivement recommandé.

6. Rapport de validation scientifique

L'étape de validation doit donner lieu à un rapport de validation transmis au producteur. Afin de ne pas multiplier les échanges et de privilégier la proximité avec le producteur, ce rapport est établi par la plateforme par laquelle la donnée est entrée dans le système. Il incombe donc aux plateformes régionales ou thématiques. La forme de ce retour au fournisseur (rapport spécifique, email, via site web, etc.) n'est pas imposée par le présent GT.

Le rapport doit informer le producteur du résultat de la validation régionale et de la validation nationale quel que soit le résultat des validations. Il peut s'agir d'un même rapport ou de rapports distincts selon le souhait de chaque plateforme et selon les délais intervenants entre les deux validations.

La périodicité de mise en œuvre des rapports (à chaque donnée ayant subi la procédure de validation, selon un pourcentage des données du jeu contrôlé ou à date fixe, etc.) est laissée au libre choix des plateformes. La procédure doit cependant être décrite dans le protocole et indiquée aux producteurs.

Le rapport de validation devra indiquer au producteur le résultat de la validation régionale et nationale. Il devra fournir un moyen au producteur pour revenir sur les données concernées et lui indiquer la démarche pour réaliser d'éventuelles corrections²³.

La plateforme nationale devra ainsi transmettre aux plateformes régionales et thématiques le résultat de la validation nationale et l'ensemble des informations nécessaires à l'établissement du rapport de validation par ces dernières.

²³ Voir IX. Mises à jour

VIII. Regroupements de données

Un regroupement de données (concept du standard de DEE) consiste en une possibilité laissée au producteur d'associer des observations entre elles, tout en gardant à chacune ses caractéristiques individuelles complètes. Il ne s'agit en aucun cas d'une ouverture sur la possibilité de synthèse ou d'agrégation mais de la possibilité de lier, ensemble, des données ayant des caractéristiques communes (en général des caractéristiques d'acquisition similaires).

Pour une donnée faisant partie d'un regroupement de données (exemple : ensemble de « prises » d'un même piège, relevé d'une communauté d'espèces, etc.), sa non-conformité, son incohérence ou son invalidation scientifique sont susceptibles d'impacter les autres données auxquelles elle est associée, la cohérence scientifique de l'ensemble du regroupement et son intégrité.

Le choix de gestion de ces données regroupées pour les contrôles de doublons, de conformité et de cohérence, sera discuté dans le cadre de l'élaboration du protocole commun pour ces aspects²⁴ afin de définir les cas qui nécessitent :

-de traiter unitairement la donnée,

-ou de traiter le regroupement dans son ensemble. Auquel cas :

-la non-conformité d'une donnée d'un regroupement entraîne la non-conformité de l'ensemble des données du regroupement. Cela a pour conséquence que l'ensemble des données du regroupement ne circule pas dans le SINP.

-la non-cohérence d'une donnée d'un regroupement entraîne l'incohérence de l'ensemble des données du regroupement. Cela a pour conséquence que l'ensemble du regroupement ne circule pas dans le SINP.

Le choix de gestion de ces données regroupées pour l'application de la validation scientifique devra être explicité au sein des protocoles de validation scientifique de chaque plateforme précisant les cas qui nécessitent :

-de traiter unitairement la donnée,

-ou de traiter le regroupement dans son ensemble. Auquel cas les données d'un même regroupement portent le même niveau de validité.

²⁴ Voir VI.2. Protocole de doublons, de conformité et de cohérence

IX. Mises à jour

Nous ne traitons pas ici les questions relatives à l'architecture du SINP pour lesquelles un groupe de travail dédié existe. De fait, ne seront pas abordées ici les questions relatives notamment à la périodicité des échanges entre plateformes.

Cependant, certaines mises à jour sont susceptibles (selon leur type) de nécessiter une révision du niveau de validité ou un nouveau contrôle de conformité ou de cohérence. Nous abordons donc ici l'identification des cas de figure pouvant avoir un impact sur la validation ainsi que la démarche méthodologique générale à suivre le cas échéant. La précision de la démarche technique restant quant à elle du ressort du GT architecture.

Afin de garantir la cohérence des informations diffusées au niveau régional et national, toutes les données ou éléments de métadonnées ayant subi une mise à jour (une correction, un ajout, la mise en conformité ou en cohérence des informations ou encore la mise à jour du niveau de validité ou des informations associées) doivent faire l'objet d'une nouvelle livraison par la plateforme source à la plateforme cible. Il faudra donc stocker²⁵ la date des modifications qui va permettre d'identifier automatiquement les données à relivrer.

Les producteurs doivent être tenu informés des mises à jour des contrôles réalisés sur leurs données et métadonnées le cas échéant.

La traçabilité des corrections et des mises à jour doit être assurée.

1. Cas pouvant nécessiter une mise à jour inhérente à la validation

a) **Correction et/ou ajout d'information**

Les producteurs sont responsables des données et des métadonnées qu'ils transmettent au SINP.

Tout producteur de données doit disposer d'un droit de rectification des informations qu'il transmet dans la limite du respect du protocole auquel ils adhèrent en alimentant le système (pas de dénaturation des données transmises a posteriori qui ne seraient pas en adéquation avec ces règles).

Aucune correction ou dénaturation (sauf mise à blanc des champs non conformes et incohérents et opérations inhérentes à la standardisation) ne doit intervenir sur les données sans l'aval préalable du producteur. En cas de correction, il sera alors nécessaire de stocker une date de mise à jour, les champs corrigés et éventuellement la raison de la correction.

²⁵ Voir XI. Autres besoins pour assurer la validation

Une mise à jour des niveaux de validité peut être mise en place suite à une correction ou un ajout d'information aux données précédemment transmises selon le caractère mineur ou majeur des modifications apportées²⁶.

b) Edition d'une nouvelle version d'un standard d'échange de données ou de métadonnées SINP

La conformité et la cohérence des informations doivent être revérifiées dès lors qu'une nouvelle version d'un standard d'échange de données ou de métadonnées est validée et mise en circulation. La validation scientifique pourrait également être impactée par ces évolutions (exemple dans le cas de l'ajout de nouveaux attributs normalisés par le standard).

c) Evolution des connaissances et de l'écologie des espèces

Les connaissances et les caractéristiques (chorologie, statut biologique, écologie, etc.) des espèces étant évolutives, il est nécessaire que les référentiels et les bases de connaissance utilisés pour mettre en œuvre la validation soient mis à jour et versionnés.

L'ampleur de ces évolutions étant variables selon les groupes taxonomiques concernés et les zones géographiques, la définition de la fréquence de mise à jour de chaque référentiel et base de connaissance est du ressort de chaque plateforme et soumise à la validation des différentes instances qui organisent leur production le cas échéant.

La table relationnelle associée au document de protocole (Cf. Protocole de validation) doit être complétée dès l'implémentation d'une nouvelle version d'un référentiel ou d'une base de connaissance.

Concernant la mise à jour du niveau de validité : une nouvelle version d'une couche de référence (pour une autre emprise que la précédente) ne nécessitera cependant pas de mise à jour des données de l'emprise précédente. Dans le cas des couches de référence, seules les corrections de ces couches pour une même emprise nécessiteront une correction éventuelle du niveau de validité. *Exemple : sortie d'un atlas faisant référence pour 2010-2015 : ne nécessite pas de revalider les données de 1950 à 2009.*

Le choix de révision du niveau de validité des données suite à la production d'un nouveau référentiel est du ressort des plateformes en charge de la validation pour les niveaux de validité qui leur incombent.

Dans le cas des mises à jour de référentiels, il serait nécessaire aux différents opérateurs de la validation de pouvoir identifier les évolutions impactant la validation, c'est à dire les cas pour lesquels la validation pourrait être réinitialisée automatiquement.

²⁶ Voir IX. Mises à jour

Exemple dans le cas de TAXREF : analyse des cas de « splittings » complexes à la sortie d'une nouvelle version du référentiel. Idem pour les évolutions INSEE (fusion ou scission (splittage) de communes, etc.).

Ce travail est jugé important par le groupe de travail et impacte la validation des données cependant la mise en pratique de ces travaux outrepassa le périmètre du présent GT.

d) Evolution du protocole de validation

Le document descriptif du protocole doit être mis à jour dès lors qu'il y a modification de la procédure, en tout ou partie et une nouvelle version doit être produite (VX.x+1 pour les versions avec modifications mineures et VX+1 pour les versions avec modifications majeures). La table relationnelle associée au document de protocole (Cf. Protocole de validation) doit être complétée dès l'implémentation d'une nouvelle version d'un référentiel ou d'une base de connaissance.

Les données portent la version du protocole faisant référence au moment de la mise en œuvre de l'étape de validation ainsi que la date de mise en œuvre de l'étape de validation.

2. Démarche de mises à jour des contrôles de conformité et de cohérence

La plateforme, pour les données dont elle est en charge pour les aspects de conformité et de cohérence, est responsable de la mise à jour des contrôles. Lors de la mise à jour de ces contrôles, les dates de réalisation sont mises à jour.

3. Démarche de mise à jour des niveaux de validité scientifiques et informations associées

Cas où la plateforme juge que la modification n'est pas de nature à modifier le niveau de validité (globalement ou par groupe taxonomique); la modification est dite mineure :

Exemple : dans la nouvelle version du référentiel taxonomique, des remaniements taxonomiques de type synonymie simple, sans splittage²⁷ dans l'aire géographique concernée.

-Le niveau de validité reste inchangé.

-La date de validité (pour le niveau de validité de la plateforme qui réalise l'opération) est mise à jour.

²⁷ Voir XV. Glossaire

**Cas où la plateforme juge que la modification est de nature à modifier le niveau de validité;
la modification est dite majeure :**

Exemple : dans la nouvelle version du référentiel taxonomique, des remaniements splittent une espèce en plusieurs taxons dans l'aire géographique concernée (donc la détermination antérieure est alors sujette à caution).

-Le niveau de validité (pour le niveau de validité de la plateforme qui réalise l'opération) est mis à zéro ou à la valeur par défaut (selon la normalisation arbitrée par le GT Standard de données).

-La date de validité (pour le niveau de validité de la plateforme qui réalise l'opération) est mise à jour.

-Les données repassent alors par le circuit de validation.

Dans le cas où la mise à jour d'un niveau de validité est réalisée par une plateforme régionale sur une donnée ayant déjà été soumise à une validation nationale et inversement.

Considérant que seule la plateforme régionale dispose d'un droit de rectification de la validation régionale et que seule la plateforme nationale dispose d'un droit de rectification de la validation nationale, les données doivent être re-livrées par la plateforme source à la plateforme cible qui peut alors choisir de mettre à jour la validation des données selon la démarche décrite ci-avant. La plateforme cible doit avoir les moyens d'identifier automatiquement les données ayant été modifiées qui nécessiterait une mise à jour du niveau de validité. C'est-à-dire qu'elle doit pouvoir identifier les cas de modifications majeures des cas de modifications mineures.

X. Exigences fonctionnelles

Les exigences fonctionnelles consistent à décrire les éléments nécessaires à faire porter sur la DEE et les métadonnées associées pour faire état des résultats des contrôles de doublons, de conformité, de cohérence et de la validation scientifique et assurer la traçabilité des opérations.

Ces exigences fonctionnelles seront portées à la connaissance des GT impliqués pour permettre leur implémentation dans les standards d'échange.

Conditionnalité et règles :

O : Obligatoire. Cette cardinalité s'applique lorsque le renseignement de l'attribut est toujours requis.

*OC : Obligatoire conditionnel. Cette cardinalité s'applique lorsque le renseignement de l'attribut dépend du renseignement d'un autre attribut **et que la conditionnalité est vérifiable à partir des éléments du fichier échangé.***

*R : Recommandé. Cette cardinalité, proche de l'OC, s'applique lorsque le renseignement de l'attribut dépend du renseignement d'un autre attribut **mais que la conditionnalité n'est pas vérifiable à partir des éléments du fichier échangé.***

F : Facultatif

1. Contrôle de doublons

Le concept de doublons est recommandé. Si des doublons ont été identifiés par la plateforme, alors, ce concept doit être renseigné et les éléments échangés devront porter les informations suivantes :

Niveau	Concepts	Proposition de conditionnalité et règle
Données	Identifiant unique de la donnée qui est identifiée comme un doublon de la donnée en question	O
Jeu de données	Identifiant unique du jeu de données qui est identifiée comme un doublon du jeu de données en question	O
Dispositif de collecte	Identifiant unique du dispositif de collecte qui est identifiée comme un doublon du dispositif de collecte en question	O

2. Contrôles de conformité et de cohérence

Ces contrôles étant obligatoires pour la mise en circulation des données et métadonnées au sein du SINP, **le concept de conformité et de cohérence est obligatoire**. Les éléments échangés devront porter les informations suivantes :

Niveau	Concepts	Proposition de conditionnalité et règle ²⁸
Données	Attributs de la donnée non conformes/incohérents et valeurs de ces attributs (uniquement pour les attributs facultatifs).	R Doit être renseigné dès lors qu'il y a eu mise à blanc (ou à la valeur par défaut) d'un champ non conforme/incohérent.
Jeu de données	Date de réalisation du contrôle de conformité/cohérence des données du jeu de données.	O
	Version du protocole de contrôle (doublons, conformité et cohérence) utilisé pour contrôler les données du jeu de données.	O
	Nom de l'organisme ou de la plateforme ayant réalisé le contrôle de conformité/cohérence sur les données du jeu de données.	O
Jeu de données	Attributs de la fiche du jeu de données non conformes/incohérents et valeurs de ces attributs.	R Doit être renseigné dès lors qu'il y a eu mise à blanc (ou à la valeur par défaut) d'un champ non conforme/incohérent de la fiche de jeu de données.
	Date de réalisation du contrôle de conformité et de cohérence de la fiche du jeu de données .	O
	Version du protocole de contrôle (doublons, conformité et cohérence) utilisé pour contrôler la fiche de jeu de données .	O
	Nom de l'organisme ou de la plateforme ayant réalisé le contrôle de conformité et de cohérence de la fiche de jeu de données .	O

²⁸ Les cardinalités indiquées sont la cible à atteindre. Dans un contexte de structuration et de mise en place progressive des fonctionnalités de plateforme, une tolérance devra être permise dans le renseignement de ces attributs.

Dispositif de collecte	Attributs de la fiche du dispositif de collecte non conformes/incohérents et valeurs de ces attributs.	R Doit être renseigné dès lors qu'il y a eu mise à blanc (ou à la valeur par défaut) d'un champ non conforme/incohérent de la fiche du dispositif de collecte.
	Date de réalisation du contrôle de conformité et de cohérence de la fiche du dispositif de collecte.	O
	Version du protocole de contrôle (doublons, conformité et cohérence) utilisé pour contrôler la fiche du dispositif de collecte.	O
	Nom de l'organisme ou de la plateforme ayant réalisé le contrôle de conformité et de cohérence de la fiche du dispositif de collecte.	O

3. Validation scientifique

Le choix de faire porter les résultats de la validation scientifique sur la donnée (et non au niveau du jeu de données) est lié au postulat selon lequel la structuration des données par jeux de données est liée à la collecte (protocole de collecte commun aux données d'un même jeu) et non à l'opération de validation scientifique qui intervient ultérieurement à la structuration des données en jeu de données.

a) Validation « producteur »

Le concept de validation producteur est facultatif.

Si le concept est renseigné, alors :

Niveau	Concepts	Conditionnalité et règle
Données	Date de réalisation de la validation chez le producteur.	F Souhaitable si un niveau de validité producteur est renseigné.
	Niveau de validité producteur*	O
	Nom, prénom et organisme de l'expert en charge de la validation chez le producteur.	F Souhaitable si un niveau de validité producteur est renseigné.
	Protocole de validation (url, référence bibliographique ou texte libre).	F Souhaitable si un niveau de validité producteur est renseigné.

*Le GT validation préconise que les valeurs pour le niveau de validité producteur soient laissées à son libre choix. Pour être utilisable, ce champ devra toutefois être suffisamment explicite. Il est donc recommandé aux producteurs de ne pas utiliser de codage spécifique ou d'abréviation.

b) Validation régionale et validation nationale

L'ensemble des informations citées ci-dessous (rubrique « concepts ») devront être déclinées :

- pour le niveau régional et pour le niveau national
- pour la validation de périmètre minimal et de périmètre élargi

Les concepts de validation régionale et nationale et leurs déclinaisons par périmètre sont facultatifs.

Si un concept est renseigné, alors :

Niveau	Concepts	Conditionnalité et règle ²⁹
Données	Date de réalisation de la validation	O
	Niveau de validité	O
	Type de validation retenu (automatique, manuelle ou combinatoire)	O
	Nom de la personne et de l'organisme ou de la plateforme ayant réalisé la validation	O
	Nom, prénom et organisme du producteur recontacté par l'expert pour la validation	R Doit être renseigné si le producteur a été recontacté pour obtenir des informations complémentaires.
	Date du contact avec le producteur par l'expert pour la validation	OC Doit être renseigné si le nom, le prénom ou l'organisme du producteur recontacté sont renseignés.
	Commentaire libre sur la validation réalisée	F
	Protocole de validation (url ou référence bibliographique) et version	O

Valeurs pour le niveau de validité:

Valeurs ³⁰	Libellés
1	Certain - très probable
2	Probable
3	Douteux
4	Invalide
5	Non réalisable
	Non évalué

Valeurs pour le type de validation retenue :

Valeur	Définition
A	Automatique
M	Manuel
C	Combinatoire (= automatique et manuel)

²⁹ Ces conditionnalités s'appliquent pour chaque concept de validation dès lors que l'un de ces concepts de validation (facultatif) est transmis.

³⁰ Sous réserve des choix techniques d'implémentation du GT Standard de données

XI. Autres besoins pour assurer la validation

Au-delà des exigences fonctionnelles indispensables pour véhiculer les résultats des étapes de validation et en assurer la traçabilité, le GT Validation a identifié des éléments complémentaires utiles à la validation et à ses mises à jour.

1. Echange d'informations entre les plateformes

Éléments relatifs aux mises à jour et aux corrections³¹ :

Le concept de mise à jour est de correction est recommandé. Il doit être renseigné dès lors qu'une mise à jour/correction a été réalisée sur une donnée préalablement entrée dans le système.

Si ce concept est utilisé, alors :

Niveau	Concept	Conditionnalité et règle
Données	Date de la dernière mise à jour ou correction.	O
	Champ(s) mis à jour/corrigés.	O
	Motif de la mise à jour/correction.	F Souhaitable.
	Nom et organisme de l'opérateur (de la mise à jour/correction).	O
	Caractère mineur ou majeur de la modification ³² .	O

Autres éléments nécessaires :

Niveau	Concept
Données	Permettre de véhiculer le nom scientifique interprété, en sus du nom cité.
Données	Permettre de véhiculer la méthode de détermination.
Données	Permettre de renseigner la référence de l'ouvrage (ou de la clé, ou du site web+date de visite) utilisé pour la détermination.
Métadonnées	Véhiculer l'emprise temporelle des jeux de données.
Métadonnées	Mettre en place un identifiant permanent pour le jeu de données et le dispositif de collecte.

³¹ Sous réserve des décisions issues du GT Architecture

³² Voir IX. Mises à jour

2. Gestion des plateformes

Concept
Gérer le contact chez le producteur permettant à l'expert en charge de la validation de demander des informations complémentaires (éventuellement adresse mail). Ce point de contact peut être différent de l'observateur ou du déterminateur.
Gérer la date de réception des données (DS pour les plateformes régionales et la plateforme thématique occurrences de taxons et DEE pour la plateforme nationale)
Gérer la date de livraison de la DEE à la plateforme cible.
Gérer la date de transformation de la DS en DEE.
Archiver l'ancien niveau de validité en cas d'une mise à jour (V-1)
Archiver les corrections apportées aux données (V-1)

3. Diffusion

Lors de la diffusion des données, il serait utile de renseigner les utilisateurs finaux sur l'échelle de validation de la donnée. L'échelle de validation de la donnée étant celle de la DEE « précise », cela revient à informer l'utilisateur de la résolution de la DEE « précise » dans le cas où celui-ci n'y a pas accès (droit restreint, etc.).

Des discussions seront probablement nécessaires pour définir la manière optimale pour présenter aux utilisateurs les résultats des validations aux différents niveaux (producteurs, régional et national) et selon leur périmètre (minimal ou élargi).

4. Outils

Outil naturaliste :

L'outil naturaliste développé dans le cadre du SINP (GINCO) pourrait inclure des contrôles de conformité et de cohérence dès la saisie des informations de manière à informer directement l'opérateur de saisie des manques ou des erreurs manifestes. L'outil pourrait également se doter d'un pré-contrôle de validation scientifique. Pour cela, il devrait alors utiliser un ensemble de couches de référence (selon les contrôles souhaités) et pouvoir interroger d'autres données du système. Le résultat de ce pré-contrôle serait ainsi directement visible pour l'opérateur de saisie qui pourrait, soit choisir de corriger les informations transmises, soit confirmer l'observation en toute connaissance de cause.

Outil de plateforme :

L'outil de plateforme développé dans le cadre du SINP (GINCO) pourrait inclure des contrôles de conformité et de cohérence, directement dans l'outil ou via l'appel à des web services³³ de conformité et de cohérence dédiés qui pourraient être mis en place et rendus accessibles sur la plateforme nationale.

Le GT Validation préconise que l'outil de plateforme « GINCO » (Gestion de l'Information Naturaliste Collaborative Ouverte), en cas d'intégration des contrôles dans l'outil, se base sur les règles nationales (protocole commun pour l'identification des doublons et pour les contrôles de conformité et de cohérence).

Outil de validation :

Un outil de validation serait utile aux acteurs du SINP en charge de ces opérations. Cet outil aurait pour objectif de faciliter le requêtage sur les données, le croisement avec des couches de référence et référentiels et l'attribution du niveau de validité et des champs associés.

XII. Bonnes pratiques pour améliorer la qualité des données et faciliter leur validation

La conformité, la cohérence, la qualité d'une donnée, l'aptitude à la valider scientifiquement ainsi que la facilité d'usage par les utilisateurs finaux sont dépendants des moyens engagés tout au long du cycle de vie de la donnée, de la phase de construction du programme d'inventaire de terrain jusqu'à la diffusion des données au public.

S'il est en général impossible de revenir sur les données déjà acquises (notamment s'il n'y a pas de preuves associées), il est en revanche possible d'intervenir sur le lancement de nouvelles phases de prospections. De fait, le GT Validation préconise que les règles de collecte³³ soient définies dès le lancement des inventaires de terrain et que soit mis en place un ensemble de mesures pour permettre à l'acteur de terrain de réaliser ses prospections dans les meilleures conditions. Cela passe par le renseignement précis d'un cahier des charges incluant les méthodes et protocoles de terrain à utiliser, les informations à récolter ou encore la manière de livrer ces informations au commanditaire (en faisant référence aux standards d'échange SINP ou aux standards déclinés par les plateformes). Une information nécessaire à la validation qui n'aurait pas été notée par l'observateur risquerait en effet de nuire à la validation future de la donnée.

L'utilisation de systèmes de saisie embarqués (sur smartphone, sur ordinateur de poche type PDA ou sur tout autre produit, etc.) plutôt que l'usage de bordereaux papiers limiteraient les erreurs possibles inhérentes à l'informatisation des données a posteriori.

³³ Voir XV. Glossaire

Dans la mesure du possible, sans mettre en péril la population du taxon concerné et en respectant la réglementation sur les espèces de faune et flore sauvage protégées en vigueur, il est préconisé de demander aux observateurs de rapporter une preuve (photo, échantillon, etc.) dès lors que l'observation en question est singulière (nouvelle observation du taxon pour le territoire, etc.).

Le lancement de campagnes de terrain pourrait judicieusement être accompagné de sessions de formation des observateurs ou de préconisations quant à l'usage d'ouvrages de référence (flores, faunes, monographies, guides, etc.) afin de limiter les erreurs de détermination. Les observateurs pourraient être alertés sur les risques de confusions fréquentes et les difficultés de détermination des taxons sur leurs zones de prospection.

Le GT Validation préconise également que l'ensemble des informations récoltées soient bancarisées dans les bases sources et transmises aux plateformes de manière à être véhiculées dans la DEE. La transmission partielle des informations est aujourd'hui un frein notable pour la mise en œuvre d'une validation scientifique optimale sur les données.

En outre, force est de constater que le concept de métadonnées est actuellement bien souvent méconnu ou considéré comme ne présentant que peu d'intérêt. Les fiches de métadonnées sont alors considérées comme des informations secondaires et chronophages à produire.

Pourtant, ces informations -décrivant les données transmises- présentent de nombreux intérêts. Au-delà de la simple réponse à des contraintes réglementaires (directive INSPIRE), elles permettent notamment de valoriser les différents acteurs (maitre d'œuvre, maitre d'ouvrage, financeur, gestionnaire/centralisateur, producteur, etc.) ou encore de fournir des éléments sur le contexte de l'acquisition de ces données et sur les protocoles mis en place (protocole de collecte, plan d'échantillonnage, etc.). Ces informations sont utiles aux experts dans le cadre de la mise en œuvre de la validation scientifique. Elles le sont également pour améliorer la sélection des données selon un usage/besoin particulier. Elles augmentent de fait, la possibilité de valorisation des données pour d'autres travaux de conservation ou de protection de la nature et permettent d'améliorer l'affichage des données et d'en faciliter la consultation. Leur renseignement standardisé permet d'améliorer les contrôles sur les données et de les automatiser.

Il est important que l'animation SINP réalisée par les plateformes promeuve l'utilité des métadonnées de manière à améliorer leur renseignement par les producteurs.

XIV. Perspectives

Ce guide est le résultat de la première année de travail du GT Validation. Les travaux mériteraient d'être poursuivis afin de compléter un certain nombre d'éléments abordés dans le présent document. Le GT validation, pour la poursuite des travaux et la mise en œuvre effective de la démarche décrite, identifie notamment les besoins suivants :

- Implémenter les exigences fonctionnelles dans les standards d'échange et décliner plus précisément les règles de conformité dans les standards (produire des FAQ³⁴ le cas échéant).
- Réaliser le protocole commun de contrôle de doublons, de conformité et de cohérence national en mobilisant les membres du GT Validation.
- Implémenter ce protocole commun dans l'outil GINCO et/ou mettre en place des services d'identification de doublons, de contrôle de conformité et de cohérence
- Réaliser un modèle (trame) pour les protocoles de validation scientifique.
- Rédiger les protocoles de validation scientifique (par chaque plateforme en charge).
- Décliner le présent guide aux autres thématiques du SINP.
- Réaliser les modèles de rapports pour les fournisseurs pour les doublons, la conformité, la cohérence et la validation scientifique.
- Evaluer, après l'implémentation de la méthode, les disparités en matière de validation dans le système (au sein des régions, entre régions et entre le niveau régional et le niveau national).
- En fonction des évolutions du système et de la mise en pratique opérationnelle de cette méthodologie : identifier les évolutions à faire porter dans une seconde version du présent guide.

³⁴ Foire Aux Questions

XV. Glossaire

Acteur SINP

Personnes morales ou physiques, publiques ou privées intervenant dans la production, la qualification (la validation), la gestion, le traitement, la valorisation et la diffusion des données sur la biodiversité et les paysages ou qui, sans être productrices ou exploitantes de données sur la nature et les paysages, assurent un rôle d'animation ou de soutien aux acteurs du SINP ou encore de valorisation et de promotion du SINP.

Référence: *Protocole SINP – article 6 / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Adhérent SINP

Acteur du SINP dont l'adhésion au SINP est effective, c'est-à-dire dont le processus d'adhésion a abouti positivement.

Référence: *Ateliers du GT définition de l'architecture/ diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Administrateur

Personne disposant des compétences et droits d'accès suffisants pour paramétrer le comportement d'un système d'information, gérer des utilisateurs et créer, modifier ou supprimer des données en dehors des processus définies par l'application informatique elle-même.

Référence: *Ateliers du GT définition de l'architecture / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Annuaire

Conteneur d'information organisé où sont stockés de manière hiérarchique les acteurs et les adhérents au SINP, par unités organisationnelles.

Référence: *Ateliers du GT définition de l'architecture / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Attributs

Ce qui est propre à un être ou à un objet. Les attributs (ou propriétés) sont des informations qui caractérisent une entité (ou objet). Ces informations sont partagées par toutes les instances d'une classe. Les dictionnaires de données comportent des entités caractérisées par des attributs. Un attribut peut notamment être traduit physiquement sous la forme d'un champ d'une base de données.

Référence: *D'après : SANDRE, Onema et OIEau / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Bancarisation

Les processus de bancarisation des données permettent de conserver les données dans le cadre organisé d'une base de données d'où il est aisé de les extraire au moyen de requêtes.

Référence: *Glossaire IDCNP / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Cohérence

La cohérence désigne le respect de la logique combinatoire des informations transmises au sein des données, au sein des métadonnées ou entre données et métadonnées.

Référence: *Ateliers du GT « Validation »*

Collecte de données

Processus se rapportant à la transmission des données produites par différents acteurs et moyens vers un dispositif de conservation. Le plus souvent la collecte des données est assurée par l'observateur lui-même, qui effectue la sélection, la mise en forme et la transmission des données. Il peut arriver qu'un intermédiaire assure la collecte d'un ou plusieurs observateurs.

Référence: *Sandre / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Conformité

La conformité désigne le respect des règles fixées dans le cadre de la mise en œuvre des formats standards de données et de métadonnées autant sur les aspects physiques que conceptuels (renseignement des champs obligatoires, format, utilisation des référentiels et des listes de valeurs/nomenclatures).

Référence: *Ateliers du GT « Validation »*

Déterminateur

Personne physique publique ou privée chargée d'identifier et nommer à partir d'une preuve de l'observation (échantillon, photographie, etc.), un individu du vivant ou un objet géologique comme faisant partie d'un taxon ou d'un objet déjà décrit dans la littérature scientifique (en utilisant notamment les « clés » dichotomiques proposées par exemple dans les flores et les faunes pour le vivant).

Référence : *Ateliers du GT « Validation »*

Diffusion de données

Activité qui consiste pour un producteur de contenu informatif à organiser un système actif pour porter à la connaissance d'un public prédéfini l'existence et le contenu de l'information. Dans ce cas, celle-ci est dite portable. La diffusion publique est une forme de réutilisation des données

Référence: *Diffusion des données publiques et révolution numérique / Dieudonné MANDELKERN, Commissariat général du Plan : 1999 / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Dispositif de collecte sur la biodiversité

Tout dispositif qui permet, par mesure, observation ou toute autre méthode, d'acquérir des données de connaissance sur : les milieux, les ressources et leurs usages, les pressions (et impacts associés) qui s'exercent sur les milieux et ressources, et les données économiques afférentes.

Référence: *D'après le glossaire IDCNP ET SANDRE, d'après Onema et OIEau / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Donnée d'observation

Les données d'observations [naturalistes] correspondent généralement à des données issues des relevés de terrain lors des inventaires, des suivis ou des études d'un groupe taxonomique, d'un espace, ou même d'un paysage. Elles peuvent également être issues de dépouillement de bibliographies ou d'informations issues de capteurs (radar, satellite, ...).

Référence: *D'après le GT Standard de métadonnées / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Donnée élémentaire d'échange SINP

Ce sont des données standardisées inter-opérables. Elles sont élaborées à partir des données-source selon un format standard national propre à chaque thématique du SINP (observations de biodiversité, paysages, espaces protégés, etc.). Elles peuvent correspondre à une ou plusieurs données-sources sous réserve d'assurer la traçabilité entre données-sources et DEE. Le format standard des DEE comprend des informations obligatoires correspondant à des utilisations nationales strictement listées à l'article 10.3.6 du présent protocole et des informations facultatives.

Référence: *Projet de protocole SINP V2 - 29/10/2012 / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Doublon

Un doublon est un enregistrement que l'on retrouve plusieurs fois (au moins deux fois) au sein d'un fichier, d'une table, d'une base ou d'un système. Un doublon peut s'appliquer à une, à plusieurs, ou à toutes les colonnes (champs).

Référence : *Ateliers du GT « Validation »*

Echange de données informatisé

Terme générique définissant un échange d'informations sur des réseaux d'informations d'origines diverses selon des spécifications ou des normes.

Référence: *D'après Onema et OIEau / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Expertise

Analyse faite par un spécialiste mandaté

Référence: *Centre national de ressources textuelles et lexicales du CNRS / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Fiche de métadonnées

Une fiche de métadonnées décrit une série ou un jeu de données. Une fiche contient l'adresse sur Internet (URL) de cette série ou de ce jeu.

Référence: Règlement européen n°1205/2008 du 3 décembre 2008, repris dans : *La Directive Inspire pour les néophytes*, F. Merrien et M. Léobet, 2ème édition, février 2011 / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)

Fournisseur

Personne ou organisme fournissant des données ou métadonnées à une entité cible. Le fournisseur peut, selon les cas, être également le producteur, le gestionnaire ou encore le validateur des informations qu'il transmet.

Référence: Ateliers du GT « Validation »

Format

1. Agencement structuré d'un support de données. 2. Disposition des données elles-mêmes.

Référence: *franceterme.culture.fr* - parution au JO du 22/09/2000 / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)

Gestionnaire de données

Personne disposant des compétences pour manipuler un système d'information (créer, intégrer, supprimer, modifier des données, ...).

Référence: Ateliers du GT « Validation »

Identifiant unique SINP

Attribut spécifique qui identifie de manière unique un objet au sein d'un ensemble de données.

Référence: Sandre / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)

Jeu de données d'observations

Ensemble de données d'observations homogènes et identifiables

Référence: *Note sur les données de référence et données de synthèse*, MNHN / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)

Métadonnées

Information décrivant les séries et services de données géographiques et rendant possible leur recherche, leur inventaire et leur utilisation.

Référence: *Code de l'environnement (L127-1)* / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)

Observateur :

Personne physique publique ou privé qui collecte des informations des éléments naturels (faune, flore) ou géologiques qu'il a observés sur le terrain.

Référence: Ateliers du GT « Validation »

Observation naturaliste

Ensemble des critères relevés ou mesurés sur le terrain caractérisant un objet physique (être vivant, habitat, autre) ou une association d'objets.

Référence: *Compte-rendu du GT outil de saisie naturaliste - modèle de données / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Pôle régional

Un pôle régional du SINP est une structure ou un organisme mandaté par l'animateur politique en région pour mettre en œuvre des missions de plateforme régionale (animation du réseau, gestion de données, contrôles, validation scientifique, etc.)

Référence : *Ateliers du GT « Validation »*

Producteur SINP

Il s'agit de la personne physique ou morale, privée ou publique qui produit, administre, vérifie, éventuellement diffuse des données-source à l'origine des métadonnées, données élémentaires d'échange et données de synthèse du SINP.

Référence: *Projet de protocole SINP V2 - 29/10/2012 / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Protocole

Ensemble des conventions nécessaires pour faire coopérer des entités distantes, en particulier pour établir et entretenir des échanges d'informations entre ces entités.

Référence: *Ateliers du GT architecture / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Protocole d'acquisition

Ensemble de règles et méthodes à respecter lors de la collecte de données permettant de constituer des jeux homogènes et normés de données (protocoles d'observation, de comptage, etc.).

Référence: *diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Qualification

La qualification est l'aptitude à répondre aux usages. Elle est le résultat d'une sélection selon des critères inhérents à l'objectif visé et est donc variable en fonction de cet objectif.

Référence: *Ateliers du GT « Validation »*

Regroupement

Association de plusieurs observations partageant des caractéristiques communes (notamment relatives à l'acquisition de la donnée) pour des cas spécifiques (par exemple pour une strate dans un relevé phytosociologique, un passage pour un navire de recherche, un prélèvement...).

Référence: *Ateliers du GT « Validation » / à partir du standard de données d'occurrences de taxon.*

Réutilisateur

Personne physique ou morale qui procède à une réutilisation des données du SINP.

Référence: *Projet de protocole SINP V3 – 10/12/2012 / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Services web

Technologie qui permet à des applications de dialoguer à distance via l'Internet indépendamment des plates-formes et des langages sur lesquelles elles reposent (cf. interopérabilité). Le Sandre spécifie les services web par des scénarios techniques et diffuse ses jeux de données de référence par des services web.

Référence: *Document de réflexion pour l'architecture technique du SINP / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Splittage (de taxons existants)

Un taxon est subdivisé en plusieurs, l'un d'eux gardant éventuellement le nom du taxon initial. Le « splittage » s'opère selon au moins trois cas : l'ancien nom désigne toujours une espèce de France : il faut retourner au spécimen s'il existe, et changer le taxon au niveau de l'observation ; l'ancien nom n'est plus d'actualité en France, mais toutes les populations de France sont désignées par le même nom : Taxref assure alors la transition en changeant l'auteur en "auct." ; l'ancien nom n'est plus d'actualité en France, mais il existe plusieurs noms qui s'appliquent aux populations de France : il faut retourner au spécimen s'il existe, et changer le taxon au niveau de l'observation.

Référence: *Compte-rendu du GT référentiels / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Standard

Dans le cadre des ateliers d'architecture SINP, la notion de "standard" fait référence au format standard de données élémentaires d'échanges ou de métadonnées SINP.

Référence: *Ateliers du GT définition de l'architecture / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Système d'information

Ensemble des moyens (organisation, acteurs, procédures, systèmes informatiques) nécessaires à la collecte, à la bancarisation, au traitement et à l'exploitation et la diffusion d'informations.

Référence: *SANDRE, d'après Onema et OIEau / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Thématique SINP

Domaine de connaissance du SINP présentant des caractéristiques homogènes en termes de compétences, de métier, de réseau, ou d'expertise ou encore en termes de structuration de la donnée. Les principales thématiques SINP sont : zones biogéographiques, espaces protégés, mesures de gestion de la biodiversité, observation d'espèces, paysages, habitats/végétation, sites géologiques, génétique.

Référence: *protocole SINP / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Traçabilité (éléments de)

Ensemble des informations nécessaires et suffisantes pour connaître soit les différents états successifs pris par une donnée dans les différents processus auxquelles elle est soumise (notion d'historique), soit la nature des processus ou des opérations qui l'ont affectés (création, modifications, suppression, par qui, quand, etc.).

Référence: *Ateliers du GT définition de l'architecture / diffusé sur le glossaire du site Nature France (consulté le 06/01/2016)*

Utilisateur de données et métadonnées du SINP

Désigne toute personne ou organisme utilisant (en consultation, en téléchargement ou selon tout mode de mise à disposition existant) tout ou partie des informations diffusées au sein du SINP qui lui sont rendues accessibles.

Référence: *Ateliers du GT « Validation »*

Valideur

Il s'agit de la personne physique ou morale, publique ou privée, en charge des opérations de validation.

Référence: *Ateliers du GT « Validation »*

Validation scientifique

La validation scientifique 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.

On distingue :

- La validation scientifique dite automatique qui consiste en une validation faisant appel à des résultats d'expertises préalables (des référentiels, des bases de connaissance, etc.). Ce processus est réalisé de manière automatique (informatique) dans le cadre du SINP.

- La validation scientifique dite manuelle (nonobstant le fait qu'une interface informatique peut être utilisée) qui consiste en une validation des informations faisant appelle à une expertise directe (avis d'expert suite à l'analyse des informations transmises). Ce processus est réalisé par examen « manuel » par un ou des experts dans le cadre du SINP.

Référence: *Ateliers du GT « Validation »*



SINP

Système d'Information
sur la Nature et le Paysage



Résumé :

Une multitude de jeux de données sur la nature et les paysages est produite chaque année par de nombreux acteurs sur les territoires français. Dans le cadre du SINP, les différents utilisateurs des données doivent être en mesure de connaître le niveau de confiance associé à celles-ci.

Ce document a été produit dans le cadre du Groupe de Travail sur la validation des données d'occurrence du SINP. Il décrit la méthodologie générale de validation dans le cadre du SINP. Il définit les éléments de terminologie et les principes pour l'identification des doublons, la conformité, la cohérence et la validation scientifique.

Il liste les propositions d'amendement pour les standards d'échange de données et de métadonnées sous la forme d'une liste d'exigences fonctionnelles et ouvre la voie sur les perspectives de poursuite des travaux sur cette thématique pour améliorer son efficacité et permettre sa mise en œuvre.



MUSÉUM
NATIONAL D'HISTOIRE NATURELLE