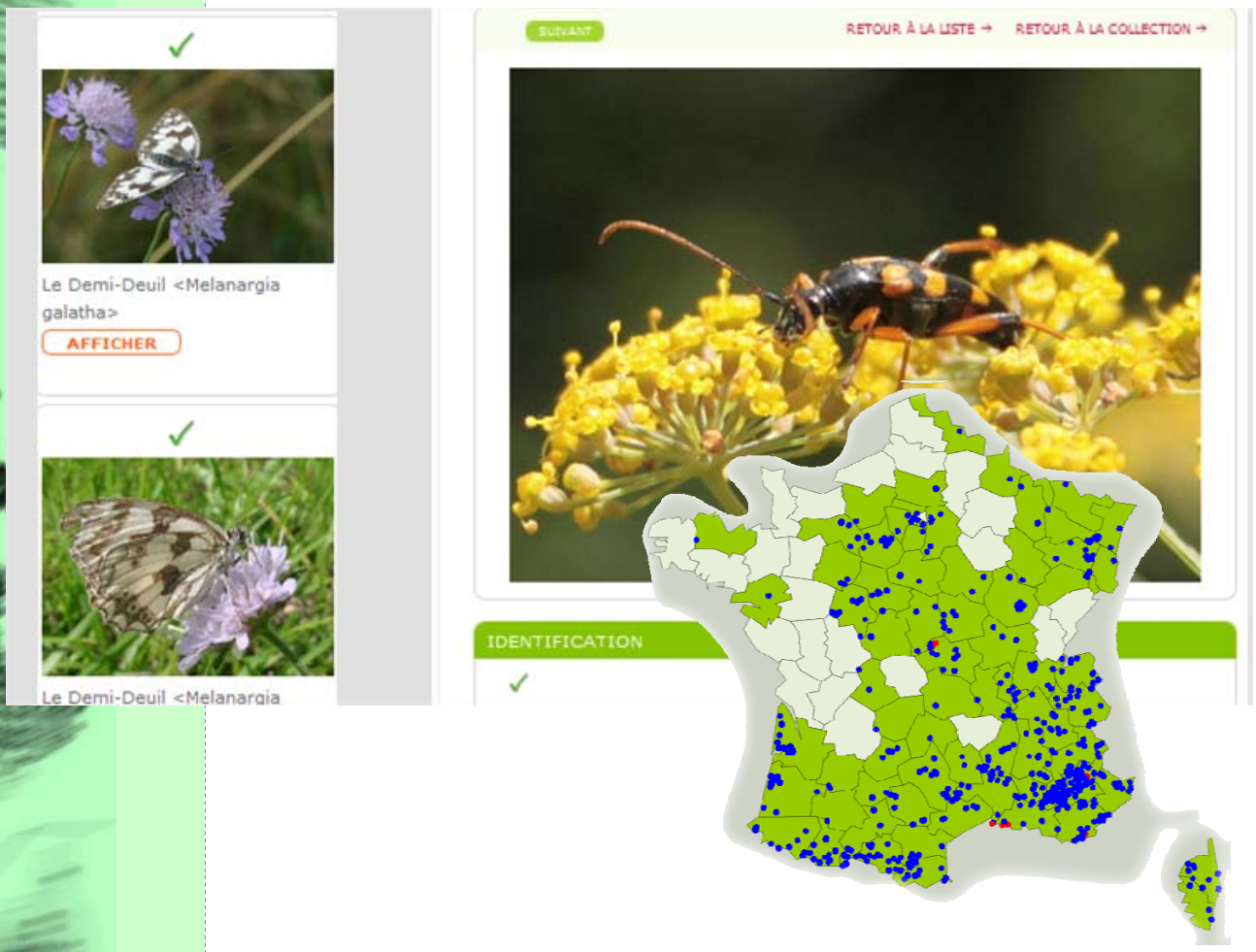
Direction de la Recherche, de l'Expertise et de la Valorisation
Direction Déléguée au Développement Durable, à la Conservation de la Nature et à l'Expertise

Service du Patrimoine Naturel

Luisa Vieira **Bogéa Soares**
Encadrement et co-rédaction : Julien **Touroult** & Laurent **Poncet**



Réflexions sur la validation des données naturalistes : le cas des erreurs d'occurrence dans la distribution des espèces

Rapport SPN 2014-38

Date : août 2013

Ce document s'appuie sur un stage de 5 mois au MNHN, réalisé en 2012 (Luisa Soares, étudiante en licence III de biologie à l'Université de la Rochelle) qui a suivi le modèle de réalisation d'une revue bibliographique, complété par des entretiens avec des spécialistes. Le présent document est une version revue et complétée par les responsables du stage.
L'ouvrage d'Arthur D. Chapman (2005), « **Les principes de qualité des données** », une publication très complète d'échelle internationale, a été pris comme guide pour la réalisation de ce travail.

Rapport rédigé en 2012
Relu et repris en 2013
Mis en forme et finalisé en 2014

Relecture :

Jeanne de Mazière
Julie Chataigner

Avertissement

Ce document explore le sujet complexe de la « validation » des données d'occurrence de taxons. Les recommandations n'ont pas l'ambition d'être directement applicables dans le cadre du Système d'information sur la nature et les paysages (SINP) ni dans l'Inventaire national du Patrimoine naturel (INPN).

L'angle adopté concerne les fausses présences (signalement erroné d'un taxon à un endroit et une date donnée). D'autres aspects de qualité sont fondamentaux pour l'usage d'une donnée biodiversité mais ne sont pas traités ici. On peut notamment attirer l'attention sur l'importance des « fausses absences » (non détection ou non prospection) et sur la qualité et l'adéquation des protocoles utilisés en fonction des objectifs d'utilisation des données.

Il faut aussi garder à l'esprit que les usages finaux ne nécessitent pas tous la même qualité en termes d'identification taxonomique.

Citation conseillée :

Bogéa Soares L. V., Touroult J. & Poncet L. 2013. Réflexions sur la validation des données naturalistes : le cas des erreurs d'occurrence dans la distribution des espèces. Rapport SPN-2014-38, 25 p.

Sommaire

1. Introduction

Les recherches bibliographiques faites au cours de ce travail ont permis de conclure que la validation des données naturalistes est peu mentionnée dans les publications scientifiques et rarement traitée comme thème principal d'étude. Pourtant son importance est incontestable. Comme l'ont argumenté Wheeler & Cracraft (1997), la gestion de la biodiversité ne peut être assurée que si elle est basée sur des informations fiables sur les espèces, habitats et ressources naturelles.

Les ouvrages traitant la problématique de la validation des données sont rares. Toutefois il y a plusieurs auteurs qui indiquent dans leurs descriptions des « matériels et méthodes » les principes de validation utilisés lors de leurs programmes de recherche. De même, quelques inventaires informatisés et issus des sciences participatives, ainsi que les bases des données partagées en ligne telles que celles du Global Biodiversity Information Facility (GBIF) fixent clairement les principes de qualité de données qu'ils appliquent. En effet, avant même de discuter précisément de la validation, **le fait de documenter la méthodologie de validation constitue déjà une bonne pratique** recommandée lors de la réalisation d'un programme d'acquisition de connaissance (Morin *et al*, 2009). C'est le meilleur moyen pour que les utilisateurs et gestionnaires puissent juger correctement de la qualité des informations et ainsi de pouvoir les utiliser de la manière la plus appropriée. La qualité des données traitées lors de la réalisation d'un programme d'acquisition de connaissance est assurée par la conception du protocole dont la mise en place d'un processus de validation.

Le présent travail s'applique aux démarches d'inventaires nationaux d'espèces[1] (Touroult *et al.*, 2012) pour assurer leur qualité sur le couple taxonomie/géographie. Dans ce cadre, les atlas issus des inventaires ont **deux défauts de nature bien différente** : le problème des fausses-absences et celui des fausses-présences (Rondinini *et al.*, 2006).

- Les **fausses-présences**. Objet de ce travail, cette erreur se produit lorsqu'une espèce est enregistrée en tant que présente dans un lieu alors qu'elle n'y est pas. Ce type d'erreur **fausse la distribution**[2] des espèce peut déboucher sur des prises de décisions biaisées, par exemple, la désignation d'aires protégées pour des espèces qui ne sont pas présentes (Rondonini *et al.*, 2006). Au-delà des inventaires, la question des fausses présences est importante car elle peut être propagée jusqu'à des utilisateurs incapables d'isoler ces erreurs de détermination.
- Les **fausses-absences**. Il ne s'agit pas à proprement parler d'une erreur mais d'un problème de synthèse, d'utilisation et d'interprétation des données. C'est le cas où, bien que présente, une espèce n'est pas enregistrée dans un lieu donné parce qu'elle n'a pas pu y être observée (problème de détectabilité) et/ou parce que le site n'a pas été échantillonné (*cf.* Sastre & Lobo, 2009). Ce problème **affecte l'aire de répartition**[3] de l'espèce. Ce problème, s'il n'est pas correctement évalué, peut conduire à des mauvaises interprétations comme par exemple de négliger indument certaines zones pour la création d'un réseau d'aires protégées (Rondonini *et al.*, 2006).

Toute aussi importante que les fausses-présences, la question des fausses-absences est un problème plus dépendant de l'utilisation qui sera faite des données. Pour ce problème, que nous ne traiterons pas en détail dans cette note, une bonne pratique serait d'associer systématiquement **une carte de l'incertitude** à une carte de répartition, qu'elle soit issue de données de distribution, d'expertise ou de modélisation (Rocchini *et al.*, 2011). La qualification des fausses-absences renvoie à une question de maille d'analyse (grain) et de quantification de l'effort de prospection (Robertson *et al.*, 2010) et à des approches statistiques (par ex : Botts *et al.*, 2011 ; Wintle *et al.*, 2012) ou pragmatiques. Ces informations doivent relever des méta-données associées à chaque jeu de données.

[1] Les questions soulevées sont cependant génériques et peuvent s'appliquer à tout niveau depuis l'inventaire d'un espace, une commune, jusqu'à une consolidation internationale.

[2] Cf. glossaire page 5-6.

[3] Cf. glossaire page 5-6

Classiquement, les cinq informations ci-dessous sont indispensables pour valider les données d'un programme de connaissance :

Quoi ? Définir les objets naturalistes (espèces, habitats patrimoine géologique)
Pourquoi ? Définir les objectifs du programme d'acquisition de connaissance est essentiel pour fixer des aspects tels que la précision, l'exactitude exigée, la complétude et les compétences attendues des opérateurs.
Qui ? Les personnes intervenant dans la réalisation de l'inventaire et de la validation doivent être bien définies. Les experts sont les opérateurs clés pour la fiabilité des données. Par contre, ils ne sont pas nombreux et leur participation rend plus coûteux le processus de validation. De ce fait, ils seront choisis selon les objectifs du programme, dans un compromis d'optimisation entre la fiabilité et les moyens de travail. À la suite de discussion avec différents experts, il ressort qu'un inventaire, pour qu'il soit bien géré, comporte trois groupes de personnes participant à sa réalisation ; chaque groupe étant responsable d'une ou plusieurs étapes.

- Le(s) **coordinateur**(s) : ils ont la mission de définir la méthodologie ou le protocole à utiliser dans la recherche, nommer les opérateurs de terrain et définir l'aire de prospection.
- Les **opérateurs** de terrain : chargés de l'observation des espèces, selon le protocole
- Les **experts naturalistes** : responsables pour la validation des données, ils peuvent aussi éventuellement participer aux prospections de terrain.

Où ? La délimitation géographique qui comprendra les données sur une aire délimitée *a priori*.
Quand ? Les dates de prospection de terrain sont aussi à définir d'après l'écologie de l'espèce et la difficulté de prospection.

L'objectif est de proposer un cadre méthodologique général de validation des données dans une perspective déclinable d'optimisation des moyens.

Ce cadre est fondé sur un comparatif des méthodologies actuellement employées et des propositions méthodologiques trouvées dans la littérature. De ce fait, il exprime la synthèse des méthodologies étudiées, permettant d'avoir une vision globale du processus et favorisant la mise en place d'une validation opérationnelle. De même, ce travail constitue un guide pour la détection des erreurs taxonomiques/géographiques (fausses-présences) dans les bases des données et un accent a été donné à l'étape semi-automatique en tant qu'outil de modernisation du processus de validation d'inventaires.

2. Enjeux des définitions

Compte tenu de la difficulté sujets et des multiples usages et pratiques, il semble important de définir les termes qui seront utilisés par la suite. Le sens précis des concepts peut changer la compréhension du rapport.

2.1. Définitions de base

Donnée naturaliste : Un évènement (observation, capture…), définit au minimum par : un taxon, une date, un lieu. Pour des questions de traçabilités permettant notamment la validation : un observateur doit accompagner une donnée naturaliste.

Jeu de données : Regroupement de données selon une entité de gestion d'information.

Protocole. Descriptif technique pour la collecte de données et leur agrégation, avec notamment les techniques de prospection à utiliser, les données à collecter, les précautions taxonomiques ou méthodologiques, les formats de transmissions, les techniques de contrôle et de validation.

Exactitude. C'est une mesure de l'erreur (taxonomique, géographique) entre la donnée d'observation et la réalité. Une valeur sera le plus exacte lorsqu'elle s'approche des valeurs réelles (Chapman, 2005), en ce qui concerne l'enregistrement des caractéristiques déterminantes d'une espèce, la localisation géographique, la date etc. En taxonomie, une valeur est **exacte** quand sa détermination est correcte.

Précision. Selon Chapman (2005), on a deux types : la précision statistique est « la proximité des observations répétées » ; et la précision numérique est le nombre de chiffres significatifs dans l'enregistrement de l'observation.

Vraisemblance. Caractère plausible (crédible) de l'observation, compte des éléments déjà connus par ailleurs, comme par exemple l'aire de répartition, la période d'activité, l'habitat d'espèce etc.

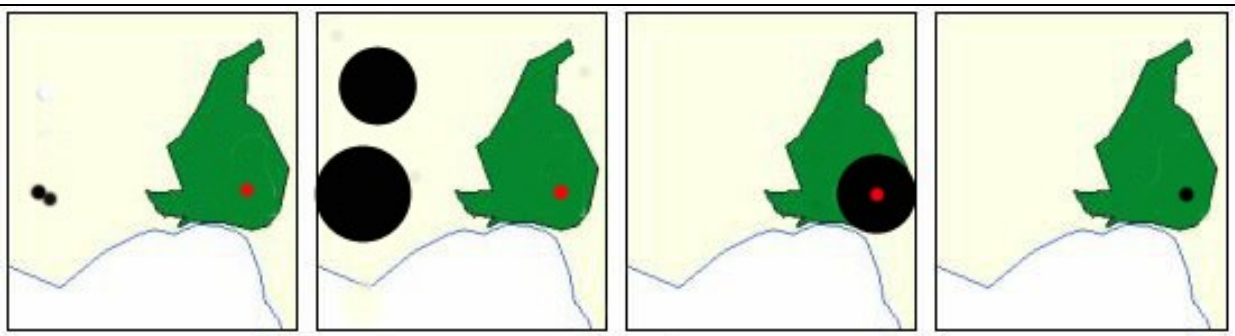


Fig. 1. Différence entre l'exactitude et la précision d'après Chapman (2005). Les points rouges montrent la vrai station d'observation, le point noir la station telle qu'indiquée par l'observateur.

Inventaire national d'espèces. C'est un processus organisé d'acquisition de données de répartition d'espèces dans le temps et dans l'espace, caractérisé au minimum par les 5 éléments suivants : un ensemble défini de taxon(s), une couverture géographique, une étendue temporelle, un processus de validation des données, un ou plusieurs niveaux de synthèse géographique(s) ou administratif(s). *cf.* fig. 2.

Détermination. Processus d'interprétation d'un ensemble de caractéristiques de spécimens pour l'affecter à une entité d'une typologie (taxon, habitat…). Dans la pratique, une détermination fiable d'un taxon repose sur une connaissance des critères à observer pour tous les taxons qui pourraient se ressembler, dans la zone d'étude et à proximité.

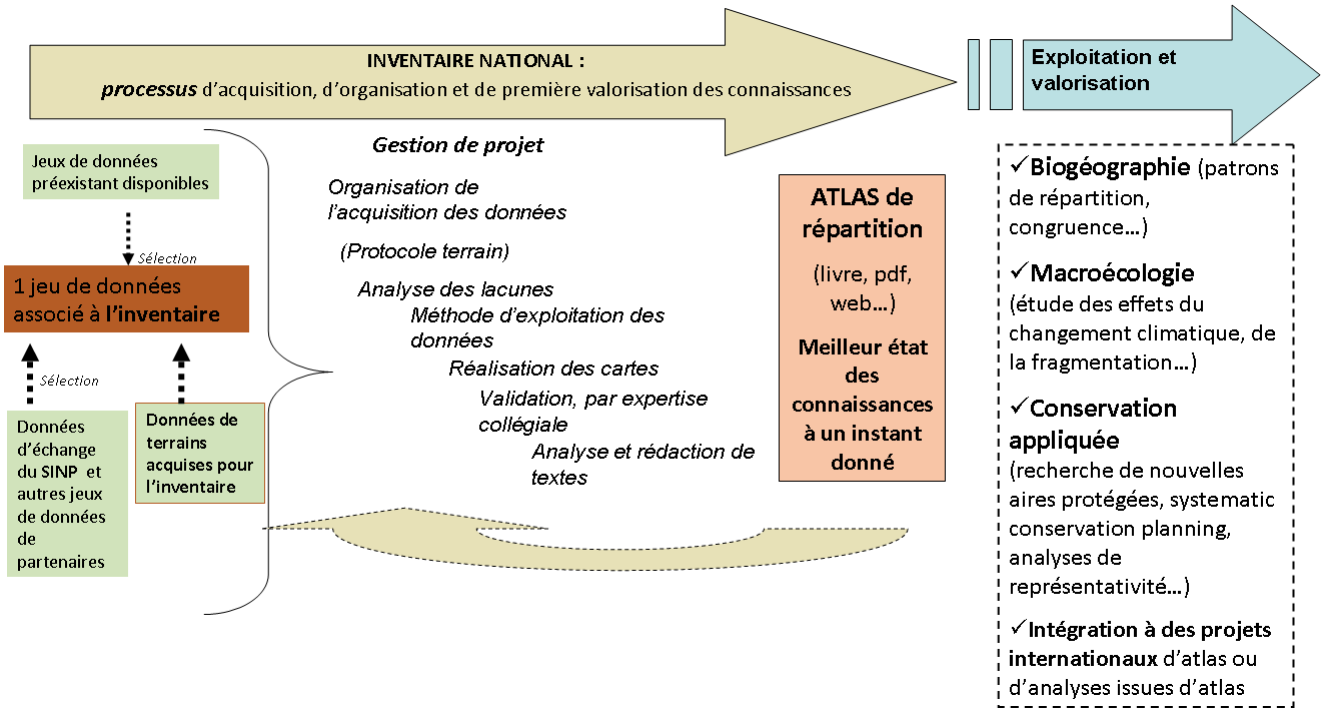


Fig. 2. Schéma global du processus d'inventaire national (d'après Touroult *et al.*, 2012). La qualification de la donnée pour son usage dans un atlas de répartition se fait à plusieurs niveaux, à minima au niveau de la sélection des données à prendre en compte dans l'inventaire et lors de la validation de la carte finale, par expertise collégiale.

2.2. Définitions géographiques

Atlas de répartition. C'est le produit de la démarche d'inventaire qui « fige » la meilleure connaissance disponible, compilée et validée dans son ensemble à un moment donné. Sous forme de publication (« papier » traditionnelle ou mise en ligne sur internet) qui synthétise des connaissances sur la répartition des espèces.

Distribution. Ensemble des zones d'occurrence avérée d'une espèce, rapporté ou non à une maille standard. Par nature il n'y a pas de notion d'absence dans la distribution.

Répartition. Aire de répartition : enveloppe (limite externe) des surfaces réellement occupées, pouvant contenir des zones non occupées. C'est une synthèse interprétée. Il faut définir une règle (seuil) pour considérer s'il y a ou non une discontinuité dans l'aire de répartition. Par définition, il y a un enjeu à définir autant la présence que l'absence.

2.3. Définitions concernant les techniques de contrôle « qualité »

Les définitions proposées sont adaptées à partir de multiples sources qui ne sont pas toujours cohérentes entre elles (*cf.* James, 2011 par rapport à Chapman, 2005). Ces définitions restent sujettes à débat et devraient faire l'objet d'un travail collégial (dans le cadre du SINP), afin d'arriver à des définitions stabilisées et partagées.

Cohérence interne. Quand les différentes informations liées et partiellement redondantes à une données ne présentent pas de contradictions. [Intéressant à évaluer uniquement si les sources sont indépendantes]. Exemple : quand on dispose à la fois du CD_nom et du nom cité, cela permet de détecter un éventuel problème de rattachement au référentiel taxonomique si les deux ne correspondent pas. Autre exemple : point GPS ne correspondant pas à la commune citée.

Vérification. Processus technique de contrôle de conformité. Vise à détecter les données qui ne sont pas conformes à des critères pré-établis. Ces critères sont définis dans une méthode et peuvent porter sur le standard de données, la cohérence interne, des aspects écologiques ou chorologiques. En théorie, cette vérification ne préjuge pas de l'usage, elle se fait par rapport à des critères génériques. Ces critères doivent être explicites.

Validation. Attribution d'un jugement d'expert sur la donnée, prenant en compte les résultats de la vérification, une connaissance approfondie du sujet (critères implicites) et l'usage prévu.

Qualification. Processus de classement **des jeux de données** en fonction de critères. Ce classement peut être selon l'usage ou la question posée et selon l'existence et l'organisation des vérifications et/ou validations appliqués aux données qui le composent.

Si le but de la validation est d'estimer l'exactitude, dans la pratique, on recherche surtout à détecter les inexactitudes visibles.

3. Les sources de données et les points clés

La figure 3 introduit les notions basiques de validation des données d'un inventaire, qui seront ensuite expliquées pour la compréhension du cadre méthodologique. La réalisation d'un atlas s'appuie pratiquement toujours sur de la mobilisation des données existantes en parallèle de l'acquisition de nouvelles informations.

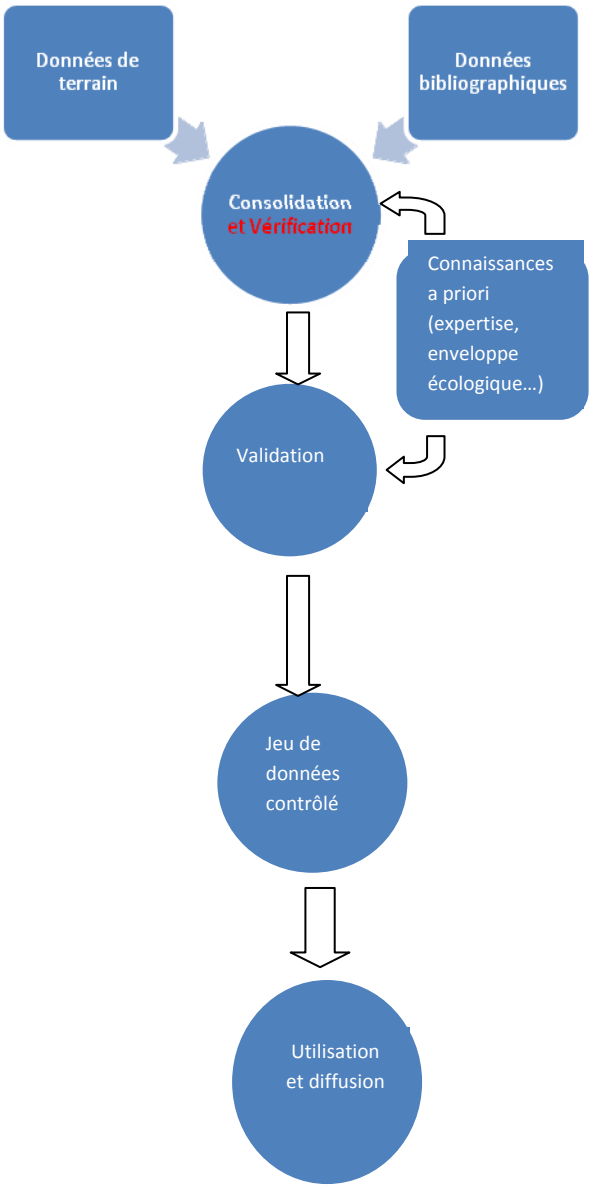


Fig. 3. Représentation simplifiée de la consolidation et de la validation des données naturalistes (schéma dans une perspective de réalisation d'atlas de répartition).

3.1. Forces et faiblesses des principales sources de données

Type de source	Point fort	Point faible	Remarques
Données de terrain selon protocole dédié au projet	o Adaptation aux questions du programme o Effort de prospection quantifié et estimation possible de la probabilité d'absence o Homogénéité des données et maîtrise des sources d'erreur	o Coût d'acquisition, même avec des bénévoles o Temps nécessaire à l'acquisition (généralement plusieurs années)	Il faut que le protocole ait été bien conçu (*cf.* qualité « à priori »).
Données de terrain opportunistes (c'est-à-dire non dédiées au projet)	o Coût d'acquisition réduit o Beaucoup de données sur certains groupes ou régions.	o Exactitude difficile à vérifier, sauf documentation associée o Protocole rarement documenté, difficulté à associer ces données avec celles obtenues selon un protocole (Robertson *et al.*, 2010) o Hétérogénéité selon les sources ou observateurs	Catégorie vaste : rapport d'étude, base de données d'association, programme de sciences participatives etc. Il est possible de sélectionner les sources à utiliser (*cf.* critères développés dans ce rapport)
Données bibliographiques	o Par le processus de publication, ces données ont fait l'objet d'une validation. o Le protocole est souvent documenté dans la publication. o Informations contenues dans la publication peuvent servir à la validation des données. o Souvent des informations sur les espèces rares.	o Il n'est pas toujours possible de revenir au spécimen o Biais taxonomique : concerne souvent les espèces rares ou en marge de répartition, et de plus en plus les espèces menacées (Boakes *et al.*, 2010) o L'informatisation des données peut être chronophage et coûteuse	
Données de collection ou d'herbier	o Possibilité de revenir au spécimen pour vérifier une détermination (Reutter *et al.*, 2003) o Couverture temporelle large o Une étude à large échelle montre que la couverture géographique est moins biaisée que dans les deux sources précédentes (Boakes *et al.*, 2010)	o Absence de protocole o Précision géographique souvent réduite (Reutter *et al.*, 2003) o Souvent moins de données récentes ; o Données encore peu informatisées (Krishtalka & Humphrey, 2000) ou peu mises en réseau (Ponder *et al.*, 2001)	

Tableau 1. Synthèse des points forts et points faibles des différents types de données d'observation d'espèces, en particulier dans une perspective d'inventaire national de répartition d'espèces

Les moyens liés à l'acquisition des données de terrain et le statut des observateurs ont beaucoup évolué ces deux dernières dizaines d'années et les publications sur le sujet sont en augmentation (McDade *et al.*, 2011). On peut citer par exemple, les clés d'identification en ligne (Penev *et al.*, 2009) et les sciences participatives.

Les naturalistes bénévoles sont nécessaires, par leur compétence et leur investissement, dans quasiment tous les inventaires menés et, ces dernières années, même le grand public joue un rôle important dans plusieurs programmes d'acquisition de connaissance (sciences citoyennes).

3.2. La consolidation

C'est le processus chargé de mettre dans le même format et la même logique (structures, référentiel taxonomique, géographique…) toutes les données d'un inventaire. Ce point est particulièrement critique quand les sources n'ont pas toutes le même référentiel, voire quand certaines sources n'ont pas réellement de référentiel. Des méthodes automatisées se développent pour réconcilier les listes taxonomiques de référence (exemple

TAXmatch du MNHN) et des méthodologies de standardisation sont en cours de formalisation dans le cadre de l'INPN (MNHN)

3.3. Les erreurs : inévitables mais à maîtriser

Toute récolte d'information contient des erreurs. Celles-ci sont en fonction de l'expérience de l'observateur, la difficulté d'identification de l'espèce, l'effort d'échantillonnage, les outils de saisie et la documentation technique fournie aux opérateurs.

Partant des définitions de Chapman (2005) des erreurs possibles dans la réalisation d'un inventaire, les quatre erreurs traitées dans la méthode actuellement proposée sont expliquées ci-dessous :

- les doublons :
  Il s'agit du cas ou deux ou plusieurs codes ou noms représentent la même donnée. Ils sont dus à des évolutions de codification ou à la consolidation de données de source hétérogène utilisant des référentiels différents. Les doublons n'affectent pas nécessairement la qualité d'un atlas mais ils peuvent être gênants pour certaines valorisation (par exemple ceci peut faire croire à une station importante d'une espèce rare par la présence de plusieurs observations qui n'en sont qu'une en réalité).

- les données incorrectes :
  Portant une ou plusieurs informations erronées, les données incorrectes comprennent surtout les erreurs d'identification. Ce sont les plus courantes et elles peuvent avoir plusieurs causes. Les principales sources sont l'erreur humaine de détermination (liée à la compétence/formation/information de l'observateur ou aux conditions d'observation …), l'erreur de saisie, ou l'erreur de conversion (lors de la consolidation par exemple, quand les référentiels initiaux ne sont pas les mêmes).

- les données incomplètes :
  Une ou plusieurs informations de la donnée indispensables à l'identification de l'espèce sont manquantes. Exemple : non-identification de l'observateur.

- les incohérences :
  Des informations de la métadonnée ou du jeu de données sont discordantes entre elles.

Selon Robertson *et al.* (2010) l'erreur la plus gênante serait la mauvaise identification de l'espèce, qui aurait tendance à être plus importante lors d'un inventaire avec participation des non-scientifiques. De la même façon, Chapman (2005) affirme qu'une détermination fiable ne peut être se faire qu'avec une implication constante des spécialistes puisque l'identification automatisée ne semble pas opérationnelle.

L'apport de nouvelles techniques de détermination

Des techniques de détermination semi-automatique ont été testées et se sont avérées efficaces dans certains cas. Motta *et al.* (2001), par exemple, ont mis en œuvre un outil de semi-automatisation de l'identification et du comptage de protozoaires. Le but était d'éviter la traditionnelle procédure manuelle, qui demande du temps et la participation d'un ou plusieurs experts. Leur stratégie est assez simple : à partir de l'analyse photographique des individus, 10 paramètres de taille et forme sont définis pour chaque espèce, avec l'aide d'un expert ou à partir des données de la littérature. A partir de ces informations, chaque individu est automatiquement déterminé. Comme résultat, le processus de validation de l'étude a prouvé que la plupart des espèces ont eu plus de 70% de fiabilité dans la détermination ; taux supérieur à celui obtenu manuellement. La détermination semi-automatique simplifie le travail des spécialistes de façon qu'ils ne soient demandés que lors de l'apparition d'une espèce-type ou si on manque d'information sur une espèce.

Cette approche reste marginale pour l'instant mais pourrait se développer dans un avenir proche, tout comme des techniques prometteuses de **détermination indirecte par l'ADN laissé par les espèc**es dans le milieu, en particulier dans l'élément liquide (ADN environnemental, exemple amphibiens dans les mares : Dejean *et al.*, 2012 ; poissons récifaux : Hubert *et al.*, 2011).

Encart 1. Les nouvelles techniques de détermination

3.4. L'expertise : une ressource rare

Les experts sont indispensables dans un processus de validation car ils sont la principale garantie de fiabilité des informations. Leur intervention se fait principalement dans la conception du protocole, notamment pour apporter des solutions aux problèmes de déterminations qui risquent de se présenter (notion de qualité *a priori*), dans la définition des vérifications à effectuer (par exemple, contrôle de l'aire de répartition, contrôle des dates d'observation par rapport à la phénologie, contrôle de la commune par rapport au point GPS) puis dans validation finale, c'est-à-dire l'acception ou non de la donnée.

Selon Oliver *et al.* (2000), le nombre de taxonomistes professionnels décline depuis des années et le nombre de taxonomistes actifs pour un groupe d'espèces ne correspond pas à la richesse spécifique du groupe C'est un point critique pour l'efficacité des méthodologies de validation.

C'est pour cette raison que les méthodes mises en place doivent simplifier et optimiser le travail des spécialistes pour valider les données d'inventaires qui le nécessitent. Il s'agit donc de concentrer le travail des experts sur les activités où leur valeur ajoutée est la plus importante. Deux pistes pour cela :

1. **Renforcer les dispositifs permettant d'assurer la qualité à priori c'est-à-dire en amont de la collecte ou de la saisie des informations ;**
2. **Trier les données de façon automatisée, selon des critères définis par les experts, pour ne soumettre à la vérification de l'expert qu'une faible proportion des données, celles délicates qui ne sont ni manifestement à rejeter, ni très vraisemblables.**

4. Proposition et structuration d'un cadre méthodologique de validation

Il s'agit du processus de vérification/validation qu'on peut schématiser en neuf points clés qui accompagnent les étapes de réalisation d'un inventaire (Fig. 7). Chaque étape comprend un aspect important pour la qualité finale de la donnée et du jeu de données.

4.1. La phase amont : prévenir plutôt que corriger les erreurs

Cette phase est particulièrement importante pour limiter structurellement les erreurs et donc faciliter la suite du processus. Cette phase n'est cependant pas maîtrisable dans le cas de récupération des données tierces collectées dans le cadre d'un autre projet, sauf par le biais des métadonnées, et le choix de ne pas prendre les jeux de données qui ne correspondent pas aux exigences.

(point 1) Définition des objectifs du programme et connaissance des espèces inventoriées

Avant d'initier un inventaire, il faut bien entendu formaliser la ou les questions auxquelles le programme doit répondre, gardant à l'esprit qu'une donnée de biodiversité recueillie pourra être réutilisées dans le cadre d'une autre étude. Cela permet aux observateurs et gestionnaires de l'orienter de façon à ne pas manquer d'informations nécessaires.

Il est aussi important d'avoir des connaissances préalables sur l'espèce qui doit être inventoriée : sa répartition géographique connue, ainsi que l'écologie et l'habitat typique vont déterminer le suivi géographique de la recherche. **Ces connaissances sont obtenues par un état de l'art approfondi avant le début du travail pratique**. En fonction de ces éléments, les responsables du programme peuvent estimer le temps nécessaire pour réaliser l'inventaire. Ces connaissances sur l'espèce sont nécessaires pour mettre en place des outils d'aide pour les opérateurs puis pour pouvoir mettre en place des vérifications semi-automatiques (point n°6).

(point 2) Définir le niveau de précision et la structuration requis de l'information

Une fois que les objectifs du programme ont été formulés, il faut définir la précision des éléments de l'étude : des données obligatoires (la précision taxonomique, géographique, de date, d'expérience de l'observateur) et des données complémentaires traitées dans le programme.

La précision doit être adaptée aux opérateurs ou alors il faut choisir les opérateurs compétents pour la précision voulue, ce qui peut être onéreux. Dans la pratique, c'est généralement la recherche du meilleur compromis.

Des outils d'aide (comme par exemple des enregistreurs GPS de terrain) permettent d'augmenter l'ambition en termes de précision (cf. point 4).

(point 3) Élaboration d'un plan d'acquisition et d'une méthodologie

Le protocole doit définir les pratiques nécessaires pour que l'inventaire puisse répondre à toutes les questions formalisées précédemment.

Les méthodologies de vérification de terrain, de récolte des données, de la saisie et lors de la saisie doivent être précisées. Par exemple, dans le protocole d'inventaire et récolte des données pour l'atlas de Lépidoptères Rhopalocères du Poitou-Charentes (PCN, 2008), cinq méthodologies sont proposées, selon chaque type d'opérateur ou selon la vulnérabilité des espèces :

- L'inventaire coordonné avec des bénévoles des groupes Lépidoptères départementaux ;
- La recherche d'espèces rares et/ou localisées ;
- Les données du réseau de naturalistes français ;
- Les données du grand public et
- Les missions d'expertise.

Ayant déjà une bonne connaissance du taxon étudié, on va aussi définir, décrire et expliquer les «coefficients de qualité» qui seront utilisées pour vérifier les données (cf. parties suivantes).

4.2. La qualité « a priori »

(point 4) Outils, formations et exigences adaptées aux opérateurs

La qualité *a priori* désigne les dispositifs d'aides aux opérateurs, inclus dans le protocole, qui permettent de maîtriser à la source les problèmes de qualité.

On peut citer :

- Un travail sur la liste d'espèces, en signalant les **risques de confusion**, des méthodes de détermination à utiliser selon ces risques (exemple pour les papillons : certains peuvent se déterminer en vol, d'autres sur photos, d'autres être capturés, d'autre encore doivent être disséqués pour études des pièces génitales…).
- L'adaptation des informations demandées par rapport au type d'observateur, et à ses compétences.
- Des **fiches d'information** sur les espèces et leur reconnaissance. Il est particulièrement important de bien montrer toutes les espèces qui pourraient se ressembler dans la zone d'étude, condition indispensables à une identification fiable (si on ne connaît pas toutes les espèces possibles, on ne peut évaluer s'il s'agit de telle espèce, puisqu'on ne connaît pas les alternatives ressemblantes). Des informations écologiques peuvent également aider l'observateur dans sa détermination.
- La fourniture de bibliographie et de clés de détermination.
- Des **formations, sessions de terrain** de mise en pratique du protocole. Rarement mise en place pour des raisons de moyens, la formation à la détermination et à l'application du protocole est certainement la meilleure façon de renforcer la qualité des observations.
- L'utilisation de **référentiels** qui contraignent la saisie et limite ainsi les erreurs. En particulier, un référentiel taxonomique.
- Des systèmes informatiques experts pouvant aider lors de la **saisie**. Exemples :
    - Lors de la saisie d'une espèce qui n'est pas connue du département, alerte de l'utilisateur pour lui demander s'il est sûr.
    - Lors de la saisie d'une espèce d'un groupe complexe, signalement des espèces proches avec les critères de distinction.

- Un système de retour d'information vers l'observateur quand une de ses données n'a pas été jugée valide, en vue d'une constante amélioration.
- Utilisation d'un formulaire standard pour la collecte de données ?

4.3. Après la saisie : détecter les erreurs

De façon générale, quand un problème bloquant est détecté, la donnée n'est pas supprimée mais elle est « flaguée » pour ne pas être utilisée. Selon l'enjeu de la donnée et les possibilités pratiques, un retour est fait au producteur pour qu'il identifie si le problème détecté peut être corrigé.

(point 5) Réconciliation : une étape intermédiaire à risque

Ce processus est une comparaison de toutes les données qui arrivent de sources multiples extérieures qui n'utilisent pas ou pas forcément les mêmes référentiels géographiques et/ou taxonomiques.

Le but de la réconciliation est de faire correspondre les données à partir des mêmes référentiels. Ceci permet de détecter ensuite des doublons et les données peuvent être soumises à des comparaisons ou analyses communes

Elle peut aussi être source d'erreur dans l'interprétation des données qui ne correspondent pas au référentiel. Par exemple, l'outil TAXmatch du MNHN propose 7 catégories de résultats selon la concordance du nom avec un nom du référentiel taxonomique TAXREF (Gargominy *et al.*, 2013). Quand le nom ne correspond que très partiellement, l'intervention d'un expert est nécessaire pour décider à quel nom rattacher l'observation.

Le cas échéant, l'observation peut être écartée si elle n'est pas rattachable au référentiel.

(point 6) Vérification scientifique semi-automatique

La « validation scientifique » actuelle s'appuie encore très souvent sur l'avis d'un expert pour chaque donnée. L'objectif présent est de limiter l'intervention des experts dans la vérification des lots de données, pour la concentrer sur les données qui nécessitent une intervention. Leur travail sera nécessaire en amont : les espèces analysées seront qualifiées par un spécialiste, qui déterminera les risques de confusion dans la détermination ainsi que les tris automatiques possibles (selon la géographie, la phénologie, l'écologie etc.). Ensuite, à partir de ces connaissances, le système informatique déterminera la présence potentielle des erreurs et des codes de fiabilité seront attribués. Selon l'objectif du programme, le système pourra directement « exclure » les données les plus douteuses ou, et c'est la solution à privilégier, les renvoyer aux fournisseurs[4].

À titre d'exemple de mise en place de ces pratiques, deux institutions britanniques sont en train de mettre en œuvre des processus semi-automatiques de validation des données : la Société Botanique des Iles Britanniques (Botanical Society of the British Isles - BSBI) et la Société Britannique de Lichens (British Lichen Society - BLS). Ce travail est également en cours dans le cadre de l'INPN. La BSBI a développé un système de stockage et partage de données – the distribution database (DDB) et le BLS accorde une importance majeure à l'exactitude de ses données. Dans les deux systèmes, une vérification semi-automatique est effectuée en début de processus et se déroule de la même façon dans cette phase initiale. Le développement consiste à des vérifications de **vraisemblance géographique en ce qui concerne la répartition connue des taxa**, ainsi qu'une réconciliation des différentes échelles géographiques et de dates pour les données arrivant d'origines différentes.

Le but de la vérification semi-automatique est d'alléger l'implication des experts. Par exemple, on pourrait viser une proportion de moins de 20% des observations devant être validées par un spécialiste contre 80% de validation ou rejet automatisé. **Ce principe est d'autant plus important que l'inventaire est ambitieux** : pour un petit projet, il reste envisageable que l'expert vérifie toutes les données ; pour des projets avec plusieurs dizaines de milliers de données, l'automatisation est nécessaire.

[4] Ceci met en avant l'importance de la traçabilité du fournisseur primaire, pas seulement son nom mais aussi par exemple un courriel permettant de le contacter facilement.

Pour définir les paramètres de validation semi-automatique le spécialiste se base sur les deux questions ci-dessous. Pour chaque espèce analysée :

- Quels sont les **risques de confusion** dans la détermination ?
- Quelle est la **vraisemblance des informations de la donnée** ? (selon l'écologie, la distribution géographique, la phénologie, l'habitat etc.)

Les réponses doivent constituer des caractéristiques suffisantes pour estimer le risque d'erreur d'une observation élémentaire.

Attention : il ne s'agit de rejeter « sèchement » une donnée s'écartant de l'écologie ou de la biogéographie connue du taxon mais de l'extraire des autres données pour la soumettre à expertise et échanger si possible avec le producteur. Le cas échéant, cela peut être au producteur de ré-expertiser la donnée qu'il a fournie.

4.4. Un codage de la qualité

Lorsqu'un jeu de données arrive au gestionnaire d'un inventaire, il faut classer le **potentiel de validité** de chaque donnée[5] afin de simplifier et d'orientation la suite du processus. Cette classification peut être faite avec l'attribution des codes de qualité. Il est évident que c'est un travail long et laborieux de devoir traiter toutes les données une par une. Pour cette raison l'attribution de codes de qualité est l'aspect majeur abordé lors de l'automatisation de la validation.
Pour constituer ce code, on considère, comme le fait implicitement tout spécialiste, deux aspects fondamentaux : le risque d'erreur de détermination et la vraisemblance de la donnée.
Ces aspects se caractérisent par des paramètres plus spécifiques, ceux-ci étant les paramètres utiles pour la vérification semi-automatique. Chaque paramètre sera évalué selon deux ou trois modalités utiles pour détailler la crédibilité des informations apportées (tableau 2).

[5] A noter que les métadonnées associées au jeu de données peuvent donner une indication précieuse sur la qualité et la confiance générale à avoir dans les données transmises, et servir directement à attribuer une qualification au jeu de données.

	Paramètre	Modalités	Remarques
Risque d'erreur ?	**A.** L'expérience de l'observateur	1 : Connu et compétent ; 2 : Autres.	Des informations complémentaires comme quel **ouvrage (clé) de détermination** est utilisé peuvent servir à estimer le risque d'erreur. Ce paramètre peut être évalué pour l'ensemble du jeu de données. Dans certains cas, c'est l'identité du déterminateur (différent de l'observateur) qu'il faut examiner.
	B. La difficulté de l'identification (pour le public visé par l'inventaire, s'il est bien défini)	1 : Facile ; 2 : Moyenne ; 3 : Difficile.	La difficulté d'identification de l'espèce est relative et dépend des participants de l'inventaire (grand public ou naturalistes). Il faudrait en fait définir un couple Expérience/difficulté ; voir un triptique : expérience / difficulté / méthode (ex : papillon avec ou sans capture du spécimen).
	C. L'existence d'un premier niveau de vérification dans le jeu de données transmis (exemple d'une validation lors de la saisie chez un partenaire).	1 : Validation par un expert ; 2 : contrôle technique (référentiel, cohérence interne) ; 3 : Pas de vérification connue.	Exemple de programme de sciences citoyennes avec système de validation des photos par des experts, type SPIPOLL (niveau 1)
Vraisemblance ?	**D.** Vraisemblable géographiquement par rapport à un état des connaissances	1 : Correspond à la répartition connue pour l'espèce 2 : En marge de la répartition connue pour l'espèce ou dans les zones ; 3 : Hors répartition connue pour l'espèce.	Peut être effectué à différents grains selon les données disponibles : région, département, maille 10 x 10 km… L'approche raisonnable actuellement nous semble le **département** ou secteur marin. Selon les informations disponibles au préalable, on peut envisager de croiser avec l'abondance dans le département, pour affiner la détection des observations intéressantes, à vérifier.
	E. Vraisemblable écologiquement (phénologie, habitat, enveloppe écologique etc.)	1 : Correspond à l'écologie connue pour l'espèce (95 % de la niche écologique) ; 2 : En marge de l'écologie connue pour l'espèce ; 3 : Ne correspond pas du tout à l'écologie connue pour l'espèce.	Le nombre et les types de caractères écologiques évalués vont varier selon l'espèce, d'après le choix du spécialiste, selon les critères définis préalablement par l'usage connu et sur base bibliographique.
Vérifiable ?	**F.** La possibilité de vérification de la donnée.	1 : Echantillon prélevé et accessible 2 : photographie, observateur pouvant être contacté 3 : pas de moyen de vérification ni de contact de l'observateur	Elément informatif pour savoir ce qu'il est possible de faire de la donnée. Un effort de contact avec l'observateur pourra être effectué pour une observation d'espèce à enjeux.

Tableau 2. Synthèse des paramètres clés pour la validation d'une donnée

L'ensemble des codes accordés à chaque donnée pourrait constituer une sorte de « code-barres », qui permettra l'attribution d'un **statut à la donnée**. Finalement, un statut serait associé à chaque code-barre. Le plus simple est par exemple : validée, possible ou rejetée. Cette classification selon 3 statuts (tableau) utilisée par l'ONF (Bouix, 2009) est efficace et simple mais peu s'avérer réductrice dans certains cas de figure rencontrés. **La méthodologie d'attribution d'un statut pour chaque combinaison de code-barres n'est pas fixe** (c'est-à-dire que le même code barre n'aboutit pas forcément à la même conclusion selon le groupe taxonomique voir selon le taxon)**. L'interprétation du code-barres est définie par espèce ou groupe d'espèces, selon les exigences du programme. Cette interprétation constitue un des éléments méthodologique du programme d'acquisition de données qui devrait être documenté.**

Statut de la donnée à l'issue de la vérification semi-automatique	Devenir de la donnée	Remarques
Vérifiée et validée automatiquement	Passage à l'étape suivante : utilisation de la donnée, selon les exigences du programme	Ne préjuge pas de la validation pour certains usages (par exemple pour certaines modélisations qui utiliseraient l'effort de prospection).
rejetée	Retour au producteur et mise de côté de la donnée	Aucune donnée n'est supprimée, elle est simplement « cochée » comme rejetée. Elle n'est pas rediffusée ni utilisée dans les analyses et cartes.
possible (à valider)	Doit être examinée par un expert qui la fera basculer dans une des 2 catégories ci-dessus. Ceci peut se faire en consultant le fournisseur de données qui assurera lui-même une partie de cette expertise.	**Statut provisoire.** Le fait que la donnée soit « à vérifier » ne porte pas de jugement sur la valeur de l'observation. Notamment toutes les observations originales de nouvelles stations à l'écart de la répartition connue rentrent dans cette catégorie, qui alerte sur le besoin d'un examen attentif de la donnée

Tableau 3. Exemple de classification d'une donnée selon son stade de validation

0. A statuer

0.1 Aucun examen [la donnée n'a pas été examinée]

0.2 En cours (dans le cas où la validation nécessite une consultation d'un expert externe)

1. Statuée :

1.0 Invalide [c'est-à-dire jugée fausse, erronée ou trop douteuse, hautement improbable compte tenu de critères écologiques etc.; la donnée ne doit être utilisée dans aucun projet]. Les cas de fort doute doivent entrer dans cette catégorie.

1.1 Validé [sans préjuger du projet]

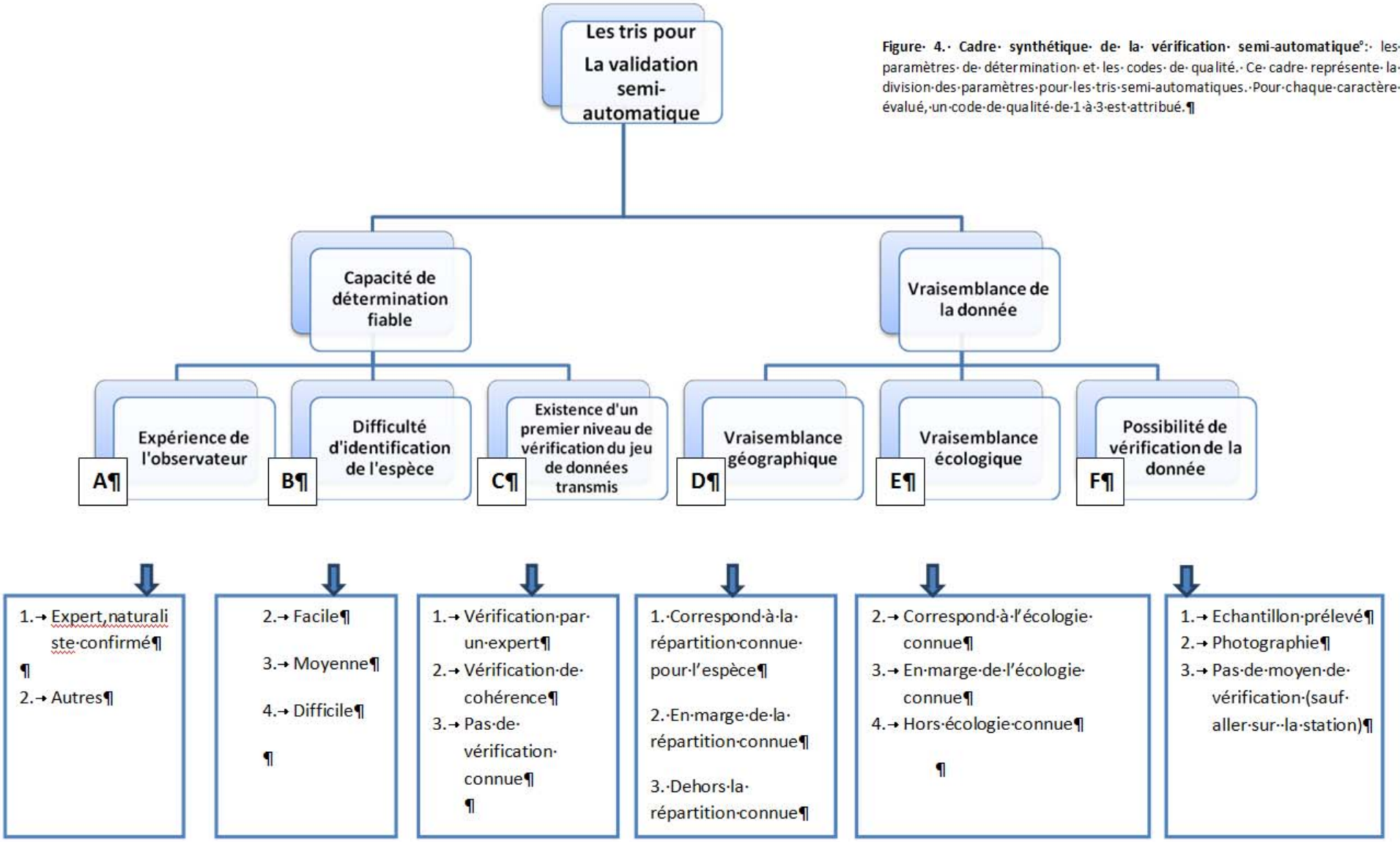
1.1.0 Possible [formulation positive dans des cas où il n'y a pas de raison formelle d'exclure mais où il n'y a pas non plus un taux de fiabilité fort, « douteux »]. Catégorie à limiter le plus possible.

1.1.1 Fiable-cohérent [tous les indices sont « positifs »]. Catégorie correspondant à la plupart des cas. « Probable »

1.1.2 Certain-vérifié [cas particulier de spécimen de collection, d'une photo vérifiable, d'un échange spécifique avec l'observateur etc.]

Encart 2. Autre exemple de typologie de statut de validation (inspiré de l'inventaire national des Rhopalocères).

Dans le système présenté dans l'encart 2, les programmes utilisant les données peuvent choisir le niveau de validation selon leurs propres exigences et objectifs.



Figure· 4.· Cadre· synthétique· de· la· vérification· semi-automatique°:· les· paramètres· de· détermination· et· les· codes· de· qualité.· Ce· cadre· représente· la· division·des·paramètres·pour·les·tris·semi-automatiques.·Pour·chaque·caractère· évalué,·un·code·de·qualité·de·1·à·3·est·attribué.¶

Espèces	***Lucanus cervus*** (Linnaeus, 1758) 	**_Cerambyx scopolii_ Fuessly, 1775** 
Programmes	*Insectes des Jardins*	*SPIPOLL : http://www.spipoll.org/galeries*
Observation utilisée comme exemple	Commune : 33 550 Mois : Octobre 2011 Pas d'autre information disponible	Geovreisset (01171). 20/05/2012 Photo, condition d'observation disponible ainsi qu'un historique de détermination de la photo
A. L'expérience de l'observateur	coefficient :2. Justification : le jeu de données vient des sciences participatives et l'observateur n'est pas connu comme expert.	coefficient : 2. Justification: le jeu de donnée vient des sciences participatives et l'observateur n'est pas connu comme expert.
B. La difficulté d'identification de l'espèce	coefficient : 2. Justification: Les *lucanus cervus* mâles peuvent être confondus en Provence avec les *Lucanus tetraodon* Thunb. Ainsi, les femelles de *Lucanus cervus* peuvent être confondues avec *Dorcus parallelipipedus* (la petite biche). La distinction est facile pour un expert mais l'expérience montre que le grand public confond facilement les deux taxa.	coefficient : 1. Justification: l'espèce se discerne des autres par sa petite taille, mais il existe des espèces qui peuvent sembler proches dans d'autres genres (*Monochamus* par exemple, pour un observateur non averti). Cependant s'agissant d'un programme sur les pollinisateurs, C. scopolii est pratiquement la seule espèce floricole parmi celles qui pourraient ressembler.
C. Existence d'un premier niveau de validation	Coefficient : 3 Justification : les observations venant de la science participative sont directement saisies dans le jeu de données qui sera géré.	coefficient : 1. Le programme SPIPOLL permet à des experts de valider (vérifier) les données sur photos. Les données vérifiées sont clairement identifiées
D. Vraisemblance géographique	Coefficient : 1. Justification : l'espèce est trouvée dans presque tous les départements de la France (http://inpn.mnhn.fr/espece/cd_nom/10502), y compris le 33, commune 33550.	Coef 1 Justification : d'après les connaissances, *Cerambyx scopolii* est trouvé dans toute la France.
E. Vraisemblance écologique	Coefficient : 3. Justification : ayant choisi la date d'apparition comme caractère écologique pertinent, la détermination est hors de la date connue d'apparition, qui est de juin à août.	Coef 1. Justification : la date d'observation est vraisemblable avec la période connue d'apparition de cette espèce printanière (90 % des observations en mai et juin ; Gouverneur & Guérard, 2011).
F. Possibilité de vérification de la donnée	Coefficient : 3. Justification: la détermination a été envoyée sans photo ni spécimen photographie.	Coefficient : 2. Justification la détermination est accompagnée d'une photographie accessible en ligne (SPIPOLL)
Code-barres pour la donnée et statut associé	323133 - possible (à valider).	311112 vérifiée
Explication / commentaire	*Cette observation ne peut être automatiquement valide pour une prise en compte pour une répartition.* *Dans ce cas, malgré le manque d'élément de vérification l'expert pourra prendre en compte cette donnée, considérant qu'il doit s'agir d'un lucane mort observé en octobre.*	*Cette donnée est « automatiquement » considérée comme valide pour une prise en compte pour une répartition.* *Elle est effectivement très vraisemblable et à fait l'objet d'un premier niveau de validation.*

Tableau 4. Illustration du mécanisme possible de validation semi-automatique de données issues de programmes différents.

Exemple de vérification de vraisemblance géographique

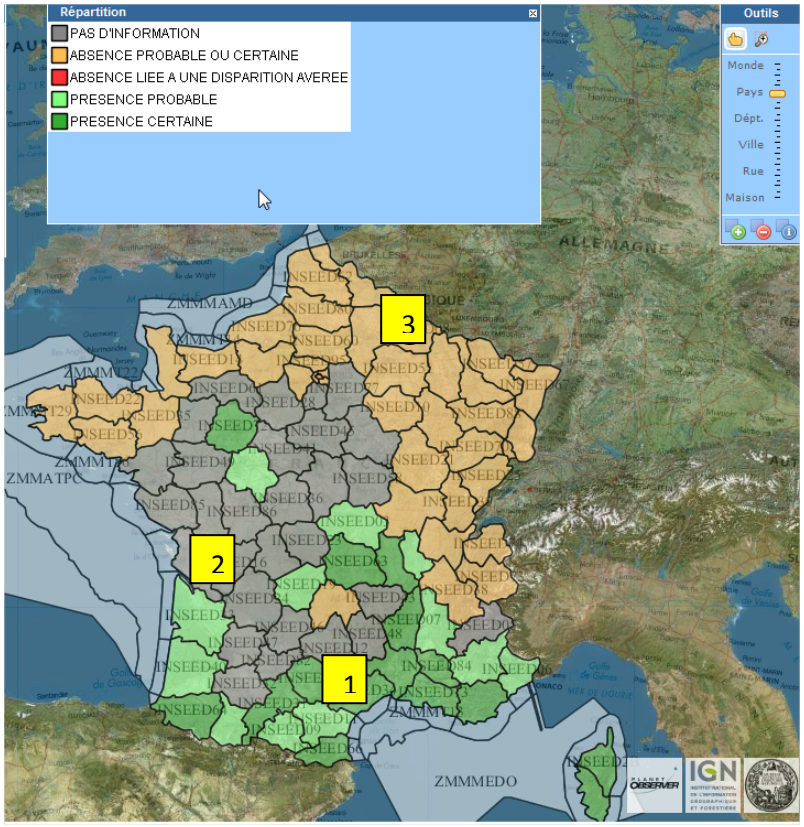


Fig. 5. Carte de répartition départementale de *Deroplia genei* (Coleoptera, Cerambycidae ; source SPN/J. Touroult). Source INPN, mars 2013.

Interprétation du paramètre D, vraisemblance de la répartition

Les codes prévus dans le programme atlas de la biodiversité départementale et des secteurs marins[6] sont pensés pour permettre de faciliter une vérification de vraisemblance

Dans le cadre proposé Fig. 4, une donnée de ce petit coléoptère peu détectable :

- des Ardennes prendraient le code 3 (donnée hors répartition) ;
- de Charente ou des Landes le code 2 (en marge de répartition) ;
- de l'Hérault, le code 1 (donnée située dans la répartition connue).

Il ne s'agit évidemment pas de rejeter directement les données codées 2 et 3 mais d'attirer l'attention de l'expert (ou en amont, de celui qui saisit).

Un exemple de codage est donné dans le tableau 4 à partir d'observations d'espèces de Coléoptères issues de deux programmes de science participative pilotés par le MNHN. Il s'agit du lucane cerf-volant et du petit capricorne, espèces plutôt communes et simples à reconnaître, y compris pour le grand public. La figure 5 illustre une des vérifications de la vraisemblance de l'observation la plus utilisée.

(point 7) Validation scientifique des données « possible, à valider »

Dans cette étape qui dans ce processus ne concerne que les données ayant un statut « possible, à valider » à l'issue de l'étape de vérification semi-automatique, une validation de la cohérence sera faite pour associer un statut définitif à la donnée, pour aboutir à uniquement **2 statuts** par rapport au programme visé : **« vérifiée » ou bien « rejetée »**.

(point 8) pour un Atlas : validation de la répartition

Les inventaires menés par le Service du Patrimoine Naturel du MNHN ont comme objectif premier la réalisation d'atlas de répartition géographique des espèces. Une fois la carte de répartition élaborée sur la base des données ayant le statut « vérifiée », une dernière et importante **validation est faite sur**

[6] http://inpn.mnhn.fr/programme/inventaire-abdsm

l'ensemble des données. Pour un atlas de répartition, cette dernière étape est la seule obligatoire et peut être couplée à la vérification évoquée précédemment.

Elle consiste à qualifier le jeu de données et la répartition qui en est issue, par une expertise collective qui vise à **valider** l'atlas de répartition.

Dans cette étape, la validation sera faite sur l'ensemble du jeu des données (on ne considère plus les données séparément), selon une évaluation de la **complétude de la répartition.** Pour cela, on réalise une analyse qui se base sur deux démarches principales.

La première consiste à estimer la pression de prospection. A partir d'une connaissance de la détectabilité de l'espèce (probabilité de détection sur terrain), on estime si le nombre d'observations par maille est cohérent et conséquemment, si on peut faire confiance aux absences[7].

La deuxième démarche part aussi de la détectabilité de l'espèce, ainsi que de ses caractéristiques écologiques pour évaluer si la méthodologie de terrain utilisée a été suffisante pour connaître l'aire de répartition de l'espèce. Par exemple, une espèce connue comme étant de détection difficile, mais qui est bien représentée dans l'échantillon (même si le nombre d'observations n'est pas étendu) ; et qu'en plus a eu une méthodologie de prospection appropriée est considérée comme potentiellement présente et peut avoir sa carte de répartition publiée.

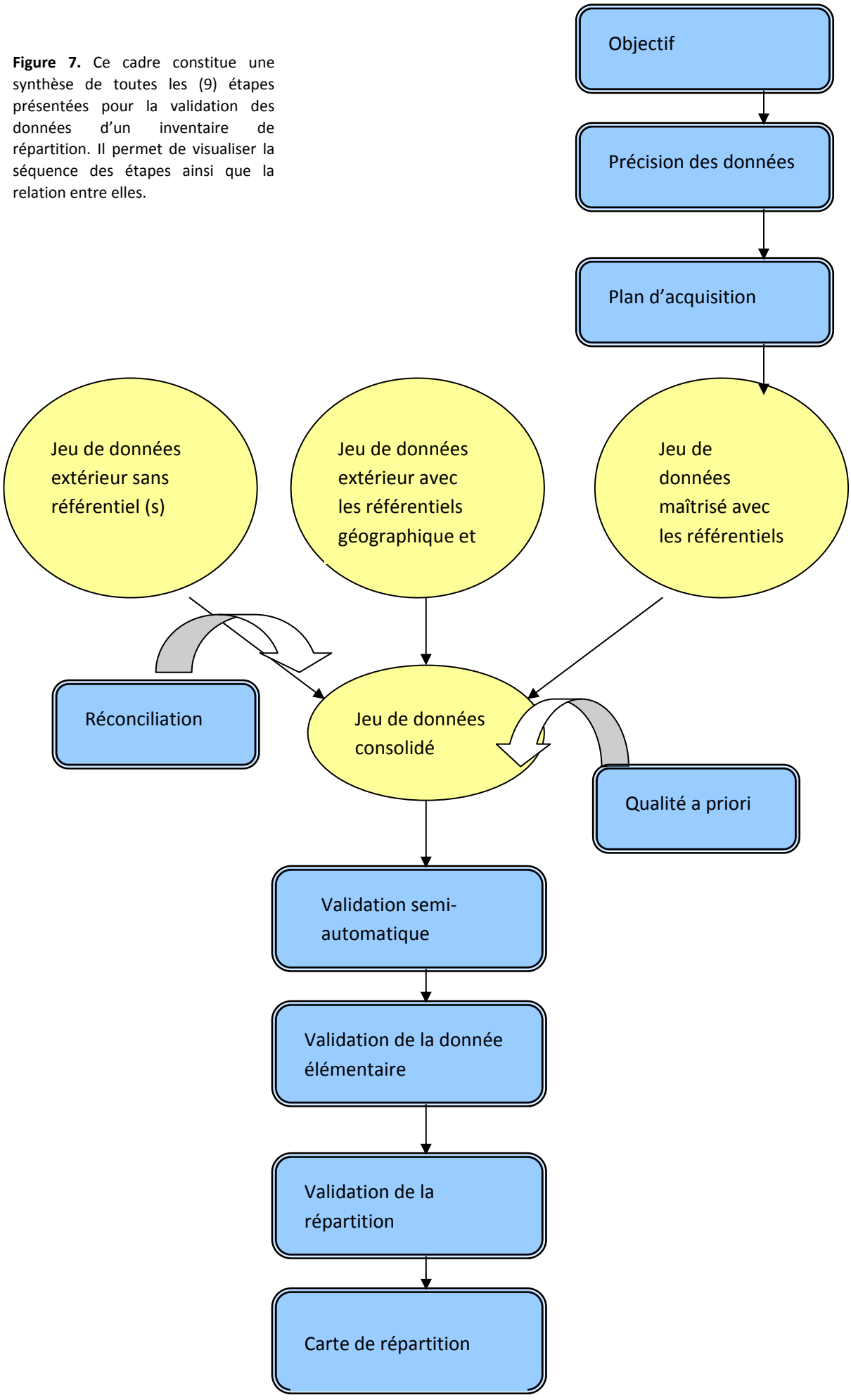
Cette étape permet également une **dernière détection de données aberrantes ou douteuses** passées au travers des étapes précédentes.

(point 9) Carte de répartition

Sous forme d'atlas, c'est le produit final de la démarche d'un inventaire. Elle représente la répartition géographique d'une espèce. Avec la documentation des éventuelles limites (biais de prospection, taxons cryptiques etc.), elle peut être utilisée de façon fiable pour définir et suivre les politiques de conservation. Suivant les recommandations de Rocchini *et al.* (2011), les atlas devraient être accompagnés de cartes d'incertitude, globales et par groupes d'espèces, voir par espèce. Il peut s'agir par exemple de la richesse spécifique observé divisée par la richesse attendue par un modèle prenant en compte l'occupation du sol et la position géographique.

[7] Si le protocole le permet, on peut aller jusqu'à estimer une probabilité d'absence basée sur le taux de détectabilité du taxon et sur un taux d'occupation estimé (cf. Wintle *et al.*, 2012)

Figure 7. Ce cadre constitue une synthèse de toutes les (9) étapes présentées pour la validation des données d'un inventaire de répartition. Il permet de visualiser la séquence des étapes ainsi que la relation entre elles.



5. Discussion et perspectives : comment mettre en pratique un tel dispositif pour les flux de données naturalistes ?

Notre proposition, issue de la synthèse documentaire et de rencontres d'experts, s'inspire des principes de la qualité au sens de la norme ISO 9001 : limiter les erreurs en amont, les détecter au plus près de la source, dès leurs premières constatations, ce qui évite l'émergence des futures confusions et inexactitudes et diminue l'effort et les coûts de détection et de nettoyage dans les étapes finales de validation (Dalcin, 2004 ; Chapman, 2005).

Cette proposition est ambitieuse par rapport à la situation actuelle. Elle nécessite des outils applicatifs et des sources de connaissances pas encore totalement constituées (exemple des répartitions par département, pas encore disponible pour de nombreux groupes taxonomiques). On peut distinguer 2 cas :

Un inventaire (régional, national). Dans ce cadre il y a une animation et un nombre restreint de taxon. Le dispositif proposé précédemment correspond peu ou prou à des bonnes pratiques existantes qu'il faudrait standardiser, rationaliser, automatiser et généraliser.

Des flux de données entre producteurs, sur tous groupes taxonomiques et pour tous usages. C'est l'enjeu des flux de données (données sources et données élémentaires d'échanges) au sein du SINP[8]. La difficulté tient alors au grand nombre d'espèces potentielles et à la disponibilité d'informations standardisées permettant la vérification semi-automatique.

Les développements nécessaires consisteraient à

- Définir ce qui est strictement nécessaire dans le code-barres (par exemple, le critère de vraisemblance géographique est plus simple et généralisable que la vraisemblance écologique qui va dépendre fortement des taxons).
- Développer rapidement les **cartes de répartition par département et secteur marin**, point clé de la vérification de vraisemblance d'une donnée et prévoir d'autres bases de connaissances permettant des filtres et alertes (par exemple, une information sur la facilité de détermination/risque de confusion).
- **Développer un outil de « vérification/validation »** pour l'attribution des codes-barres, la visualisation cartographique de jeux de données, la vérification et l'attribution d'un statut aux données. Idéalement cet outil devrait pouvoir fournir des graphiques de répartition des observations (niche écologique) : par altitude ou profondeur, occupation de sol, mois etc. et permettre la sélection des données extrêmes par l'expert (Exemple en Fig. 9).
- Intégrer au **standard de données d'échange** les champs du code-barres de vérification semi-automatique et le champ de statut de la donnée selon une typologie partagée (par exemple, à minima : rejetée, à vérifier, vérifiée).
- Organiser et fluidifier, dans la cadre du SINP, les échanges entre le producteur de la donnée source, la vérification régionale et la vérification nationale.

[8] Système d'Information Nature et Paysage

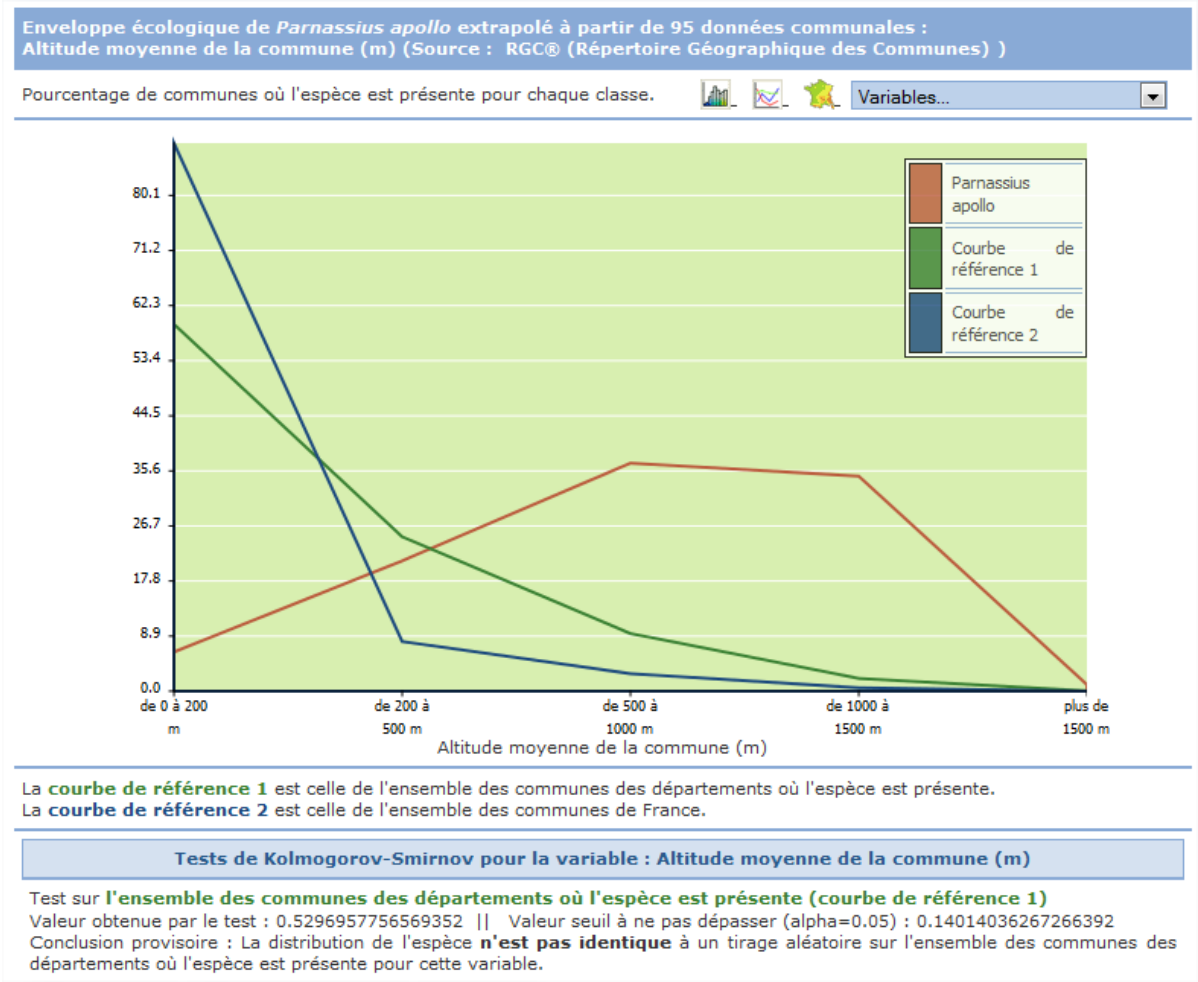


Fig. 9. Exemple d'information fournie de façon standard par l'INPN (onglet « enveloppe écologique », extrait septembre 2012). Cas d'un papillon de montagne, l'Apollon (Lépidoptère : *Parnassius apollo*).Pour le paramètre E. Vraisemblance écologique, les données extrêmes 0-500m, seraient codées 3 entraînant un statut « à vérifier », celles au dessus de 1000 m codée 1.

6. Références

Amaral, A. L., Baptiste, C., Pons, M. N., Nicolau, A., Lima, N., Ferreira, E. C. & Mota, M. 1999. Semi-automated recognition of protozoa by image analysis. *Biotechnology*. Techniques 13: 111–118.

Beaufort F. & Maurin H. 1988. *Le Secrétariat de la Faune et de la Flore et l'Inventaire du Patrimoine Naturel : Objectifs, Méthodes et Fonctionnement*. Edité par le Secrétariat de la Faune et de la Flore, Muséum d'Histoire Naturelle. 1ère édition.

Boakes EH, McGowan PJK, Fuller RA, Chang-qing D, Clark NE, *et al.* (2010) Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. *PLoS Biol* 8(6): e1000385. doi:10.1371/journal.pbio.1000385

Bouix T. *Validation des Données*. 2009. Document relatif au projet Bases des Données Naturalistes (BDN) ; Office National des Forêts (ONF), document interne non publié.

Chapman, A. D. 2005. *Principles of Data Quality*, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

Conflor. Consultation en foresterie. http://www.conflor.com.br

Corona P., Chirici G. et Marchetti M. 2002. Forest ecosystem inventory and monitoring as a framework for terrestrial natural renewable resource survey programmes. *Plant Biosystems*, 136:1,69-82.

Dalcin E. C. 2004. *Data Quality Concepts and Techniques Applied to Taxonomic Databases*. Thesis for the degree of Doctor of Philosophy, School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton. http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf.

Dahdul W.M., Balhoff JP, Engeman J, Grande T, Hilton EJ, *et al.* 2010. Evolutionary Characters, Phenotypes and Ontologies: Curating Data from the Systematic Biology Literature. *PLoS ONE* 5(5):e10708.doi:10.1371/journal.pone.0010708.

Dejean T., Valentini A., Miquel C., Taberlet P., Bellemain E. & Miaud, C. 2012. Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog *Lithobates catesbeianus*. *Journal of Applied Ecology*, 49 : 953-959. doi: 10.1111/j.1365-2664.2012.02171.x

Dommanget J.-L. 2002. *Protocole de l'Inventaire cartographique des Odonates de France*. 2002. Muséum National d'Histoire Naturelle, Service du Patrimoine Naturel & Société française d'odonatologie. 3ème édition.

Gargominy, O., Tercerie, S., Régnier, C., Ramage, T., Dupont, P., Vandel, E., Daszkiewicz, P.& Poncet,L. 2013. TAXR EF v7.0, référentiel taxonomique pour la France. Méthodologie, mise en œuvre et diffusion. Muséum national d'Histoire naturelle, Paris. Rapport SPN 2013–22.104pp

Garrard G. E., Bekessy S. A., Mccarthy M. A. & Wintle B. A. 2008. When have we looked hard enough? A novel method for setting minimum survey effort protocols for flora surveys. *Austral Ecology*, 33: 986-998. doi: 10.1111/j.1442-9993.2008.01869.x

Gougeon F., Labrecque P., Guérin M., Leckie, D., Dawson A. *Détection du pin blanc dans l 'Outaouais à partir d'images satellitaires à haute résolution IKONOS*. Presented at the 23rd Canadian Symposium on Remote Sensing / 10e Congrès de l'Association québécoise de télédétection, Sainte- Foy, Québec, Canada, August 21-24, 2001.

Gouverneur X. & Guérard P. 2011. *Les longicornes armoricains* – Atlas des coléoptères Cerambycidae des départements du Massif Armoricain. Invertébrés armoricains, les cahiers du GRETIA, 7. 224 pp.

Harvey D. J., Gange, A. C., Hawes, C. J. *et al.* 2011. Bionomics and distribution of the stag beetle, Lucanus cervus (L.) across Europe. *Insect Conservation and Diversity:* 4: 23–38.

Hubert *et al.*, 2011

Identify trees. Natural History Museum of London. http://www.nhm.ac.uk/nature-online/british-natural-history/urban-tree-survey/index.html).

James T. 2011. *Improving wildlife data quality: guidance on data verification, validation and their application in biological recording.* Rapport pour le NBN Trust. website: http://www.nbn.org.uk/getdoc/940177b0-c0fb-4604-8ca6-00c1d4a1fcfb/Promoting-data-quality.aspx.

Krishtalka L. & Humphrey P.S. 2000.*Can Natural History Museums Capture the Future? BioScience* , 50 (7) : 611-617.

Langlois D., Gilg, O. 2007. *Méthode de suivi des milieux ouverts par les Rhopalocères dans les Réserves Naturelles de France. Révision de la proposition de protocole 2002 de David DEMERGES et de Philippe BACHELARD*. Réserves Naturelles de France : BP100, 21803 Quetigny.

McDade L.A., Maddison D.R., Guralnick R., Piwowar H.A., Jameson M.L., Helgen K.M., Herendeen, P.S., Hill A. & Vis M. L. 2011. Biology Needs a Modern Assessment System for Professional Productivity. *BioScience* , 61 (8): 619-625.

Morin P.A., Martien K.K., Archer F.I., Cipriano F., Steel D., Jackson J. & Taylor B.L. 2010. Applied Conservation Genetics and the Need for Quality Control and Reporting of Genetic Data Used in Fisheries and Wildlife Management. *Journal of Heredity*, 101 (1): 1–10, doi:10.1093/jhered/esp107.

Motta M D A, Pons M N, Vivier H, Roche N., Amaral L. P., Ferreira E. C. & Mota M. Reconnaissance semi-automatique de la microfaune des boues activées des stations d'épuration des eaux usées : Protorec V 2.0. 2001.

Noël P. *Validité d'un signalement*. 1993. Texte adapté selon les pages 10 et 11 de la publication : Atlas des Crustacés Décapodes de France (espèces marines et d'eaux saumâtres). Muséum National d'Histoire Naturelle. Document non publié

Penev L, Sharkey M, Erwin T, van Noort S, Buffington M, Seltmann K, Johnson N, Taylor M, Thompson FC, Dallwitz MJ. 2009. Data publication and dissemination of interactive keys under the open access model. *ZooKeys,* 21: 1–17. doi: 10.3897/zookeys.21.274.

Ponder W.F., Carter G.A., Flemons P. & Chapman R.R. 2001. Evaluation of Museum Collection Data for Use in Biodiversity Conservation Biology, 15 (3): 648-657. Stable URL: http://www.jstor.org/stable/3061445.

Oliver I., Pik A., Briton D., Dangerfield J.M., Colwell R. & Beattier A. J. 2000. Virtual Biodiversity Assessment Systems. *BioScience*, 50 (5): 441-450.

Reutter A. B., Helfer V., Hirzel A. H. & Vogel P. Modelling habitat-suitability using museum collections: an example with three sympatric Apodemus species from the Alps. 2003. *Journal of Biogeography*, 30: 581–590.

Robertson M. P., Cumming G. S. & Erasmus B.F.N. 2010. Getting the most out of data. *Diversity and Distributions*; 16: 363-375.

Rocchini D., Hortal J., Lengyel S., Lobo JM, Jiménez-Valverde A.,Ricotta C.,Bacaro G. & Chiarucci A. 2011. Accounting for uncertainty when mapping species distributions: The need for maps of ignorance *Progress in Physical Geography, 35: 211-226, doi:10.1177/0309133311399491*

Rondinini C; Wilson KA, Boitani L, Grantham, H & Possingham HP. 2006. Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecology Letters*, 9 (10) : 1136-1145. DOI: 10.1111/j.1461-0248.2006.00970.x

Sastre P. & Lobo J. M. 2009. Taxonomist survey biases and the unveiling of biodiversity patterns. *Biological Conservation,* 142 (2) : 462-467. ISSN 0006-3207, 10.1016/j.biocon.2008.11.002

Shaffer H. B., Fischer R. N. & Davidson C. The role of natural history collections in documenting species declines. 1998. *Trends in ecology & evolution*, 13 (1): 27-30.

Suarez A.V. & Tsutsui N. D. 2004. The Value of Museum Collections for Research and Society. *BioScience*, 54 (1): 66-74.

Tingley M. W & Beissinger, S. R. 2009. Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends in ecology & evolution*, 24 (11) : 625-633.

Touroult J., Haffner P., Poncet L., Gargominy O., Noël P., Dupont P. & Siblet J.-P. 2012. *Inventaires nationaux d'espèces: définitions, concepts, organisation et points clés.* Rapport méthodologique –version 1. 2012. Rapport SPN 2012-24. http://www.mnhn.fr/spn/docs/rapports/

Tremblay R.R. et Perrier Y. 2006.*Les méthodes d'Investigation.* Complément à l'ouvrage *Savoir plus*, 2e éd. 2006, Les Éditions de la Chenelière inc.

Urban Tree Survey. National History Museum. http://www.nhm.ac.uk/nature-online/british-natural-history/urban-tree-survey/index.html.

Vigie Nature. Muséum National d'Histoire Naturelle de Paris. http://vigienature.mnhn.fr/.

Wheeler Q. D. & Cracraft J. 1997. *Biodiversity II : Understanding and Protecting Our Biological Resources.* Taxonomic Preparedness : Are We Ready to Meet the Biodiversity Challenge ? . Chapter 28: 435-446

Wintle, B. A., Walshe, T. V., Parris, K. M. & McCarthy, M. A. 2012. Designing occupancy surveys and interpreting non-detection when observations are imperfect. *Diversity and Distributions*, 18: 417–424. doi: 10.1111/j.1472-4642.2011.00874.x

Witt M., Carlson J., Brandt D .S. & Cragin, M.H. 2009.Constructing Data Curation Profiles. The *International Journal of Digital Curation,* Issue 3 (4) : 93-103.