PROPOSITION D'UNE ECHELLE POUR QUALIFIER LA VALIDITE DES DONNEES ISSUES DE PROGRAMMES DE SCIENCES PARTICIPATIVES

Groupe de travail « Poissons osseux, requins, raies »
Septembre 2022

VIGIEMER
COLLECTIF


© F.Soppelsa

Cette synthèse présente une partie des travaux et réflexions menés depuis septembre 2021 par les membres du groupe « poissons » du Collectif Vigie Mer (CVM) sur la thématique de l'interopérabilité des données, l'homogénéisation des variables et la création d'un système commun de validation des données pour les programmes de sciences participatives (SP). Le but est notamment d'améliorer la qualité des métadonnées lors de la remontée des données au niveau national.

La priorité est donc de travailler sur la valorisation des données des programmes de SP "marines" (a priori non prioritaires en ce qui concerne les politiques publiques actuelles pour la majorité des programmes existants), la validation de la qualité des données remontées au SINP n'est pas une fin en soi. Pour cela, il faut à minima que les données soient au préalable harmonisées, d'où l'importance des réflexions menées.

Dans ce document sont présentés la démarche et le résultat du travail réalisé dans le cadre d'un atelier visant à répondre à la question :

Comment évaluer la validité et la qualité des données issues de programmes de SP (sur la base de la thématique "poissons") et tendre vers une approche commune ?

Ce travail a été réalisé en collaboration avec Jeanne de Mazières (PatriNat), cheffe de projet "Connaissance Biodiversité Marine" et en charge des données « marines » pour le système d'information de l'inventaire du patrimoine naturel (SINP).

Le groupe de travail du Collectif Vigie Mer
« Poissons osseux, requins, raies »

Rédacteurs
Damien ELOIRE – NaturDive, Patrick LOUISY - Peau-Bleue, Alexandra ROHR - APECS

Remerciements aux membres du Collectif Vigie Mer ayant participé aux réflexions
Laura BARTH - Septentrion Environnement, Jeanne DE MAZIERES - PatriNat, Matthieu LAPINSKI - AILERONS, Pascal MONESTIEZ - Groupe de travail « données » du CVM, Morgane PERRI - Al Lark, François SICHEL – Les Amis de BioObs, et Paul TACHON - Réserve Grand Connétable

Mise en page
Pauline LOUBAT, animatrice du Collectif Vigie Mer

Citation du document
Eloire D., Louisy P. et Rohr A. (2022). Proposition d'une échelle pour qualifier la validité des données issues de programmes de sciences participatives. Groupe de travail « Poissons osseux, requins et raies » - Collectif Vigie Mer, Septembre 2022, 11 p.

Le SINP procède à une vérification systématique de l'ensemble des jeux de données qui lui est soumis avant la diffusion sur le portail de l'INPN. Ce processus cherche à identifier les erreurs de saisie mais aussi la vraisemblance des données. Il s'agit d'un système de validation automatique qui peut être dans certains cas complété par un processus de validation manuelle

D'un autre côté, beaucoup de temps est investi par les structures porteuses d'un programme de SP pour valider en interne les données collectées puis soumises au SINP.

Il y a en moyenne 10% des données marines soumises qui ne sont pas diffusées (sciences participatives ou non) pour cause de non-conformité aux standards de données ou de vraisemblance des observations. Pour des données validées en amont par les structures contributrices, cette "perte" d'environ 10% pose problème car elle entraîne, dans le meilleur des cas, un délai et des complications pour la prise en compte des données validées, et jusqu'à la perte totale d'informations qui, par leur « invraisemblance » même, constituent des avancées scientifiques importantes.

Le système de validation du SINP pour la plateforme nationale est aujourd'hui peut-être, et dans certains cas, moins précis que la validation que les producteurs des données peuvent proposer. Cette "perte" de fiabilité est due à la méthodologie de validation automatique qui est dépendante de la complétude et de la précision des données de référence utilisées pour les contrôles. Par exemple, la validation se base sur la vraisemblance par rapport à la répartition déjà connue des espèces disponible dans le référentiel TAXREF.

De plus, il n'y a actuellement pas de variable indiquant les cas où le système de validation du producteur de la donnée est plus précis que le système national. Cela pourrait être prochainement mentionné lors du téléchargement des données par les utilisateurs (ajout d'une colonne avec un indice de fiabilité du contributeur de la donnée). Néanmoins, lorsque les données sont intégrées dans le cadre d'un inventaire national, la validation est aussi réalisée manuellement par des experts et devient donc plus robuste.

Le système est évolutif et des mises à jour sont faites régulièrement sur la méthodologie de validation des données au niveau du SINP.

Mérou brun © P. Louisy

Le tableau de référence du SINP qui a servi de support pour le travail est présenté ci-après *(Tableau 1)* et est également accessible via ce lien : https://inpn.mnhn.fr/programme/donnees-observations-especes/references/validation

Tableau 1 : Niveaux de validité du SINP pour le résultat des opérations de validation scientifique

Code SINP	Libellé SINP	Définition SINP
1	Certain (très probable)	La donnée est exacte. Il n'y a pas de doute notable et significatif quant à l'exactitude de l'observation ou de la détermination du taxon. La validation a été réalisée notamment en présence d'une preuve de l'observation qui confirme la détermination du producteur ou après vérification auprès de ce dernier.
2	Probable	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.
3	Douteux	La donnée est peu vraisemblable ou surprenante mais on ne dispose pas d'éléments suffisants pour attester une erreur manifeste. NB C'est souvent la catégorie où atterrissent les nouvelles signalisations géographiques si l'on n'y prend garde !
4	Invalide	La donnée a été infirmée (erreur manifeste/avérée) ou présente un trop bas niveau de fiabilité. Elle est considérée comme trop improbable (aberrante notamment au regard de l'aire de répartition connue, des paramètres biotiques et abiotiques de la niche écologique du taxon, la preuve révèle une erreur de détermination).
5	Non réalisable	La donnée a été soumise à l'opération de validation mais l'opérateur (humain ou machine) n'a pas pu statuer sur le niveau de fiabilité. Notamment : état des connaissances du taxon insuffisantes (par l'opérateur ou en général) ; informations insuffisantes sur l'observation. NB C'est la valeur par défaut (niveau initial ou temporaire).
6	Non évalué	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.

La correspondance avec le standard SINP a été réalisée lors du processus de mise en place des différentes catégories proposées par le CVM *(Tableau 2)*.

Les discussions au cours de l'élaboration de ce tableau ont conclu qu'il n'était pas souhaitable d'attribuer le niveau de fiabilité le plus élevé à des données simplement parce qu'elles nous semblent cohérentes vis-à-vis de nos connaissances des espèces concernées : facilement identifiables, aire de répartition bien connue, etc. Ce "surclassement" pourrait avoir pour conséquence de surestimer les populations pour ces espèces, ce qui pourrait être dommageable dans certain cas (e.g. renégociation des moratoires mérou et corb).
Ainsi, une donnée considérée comme très probable ne doit pas être classée comme "certaine" sur la seule base d'une cohérence avec ce que nous connaissons de l'espèce en question *(se référer aux définitions et critères des catégories 1b et 2 du Tableau 2)*.

Une double validation avec présence de preuve est à encourager dans la mesure du possible *(catégorie 1a du Tableau 2)*. Cette preuve permettra, à tout moment, à un utilisateur ultérieur des données de valider l'identification de l'espèce, et dans certains cas, par exemple pour les espèces "rares" ou "peu communes", d'aller au-delà du standard SINP qui les exclut car absentes de leurs listes de répartition : si on a la preuve, la donnée pourra être acceptée plus facilement et utilisée.

Dans les catégories SINP, pour les observations douteuses et invalides *(catégories 3 et 4 du Tableau 2)*, les définitions ne sont pas assez objectives du point de vue du CVM. Le besoin de clairement différencier ces deux catégories a été travaillé. Le doute de l'observateur n'existe que si l'on offre à celui-ci la possibilité d'indiquer son degré de certitude lors de la saisie de sa donnée, ce qui n'est pas le cas de tous les programmes. Le CVM propose une formulation moins ambiguë de ces deux catégories et des exemples ont été ajoutés pour aider à la compréhension de chacune des catégories.

L'intérêt des données peu fiables a été discuté *(catégorie 3 du Tableau 2)*, ainsi que la manière dont elles peuvent être prises en compte dans des analyses. En général, une logique de bancarisation du type "on garde ou non la donnée" (on garde seulement si on est certain, sinon on supprime du tableau de données) est à proscrire, car on risque de perdre des données intéressantes à terme. Le schéma de pondération (plus ouvert et souple) des données en fonction de leur indice de fiabilité (niveau de validation) est recommandé. On garde ainsi toutes les données dans la base, mais avec des niveaux de fiabilité différents. Selon ses objectifs d'analyse, un statisticien peut ainsi donner plus ou moins de poids à certaines données en fonction de l'usage qu'il veut en faire, sur la base de leur indice de fiabilité.
Des données peu fiables mais concordantes peuvent ainsi apporter des éléments scientifiques intéressants (par exemple dans le cas des modifications des aires de répartition de certaines espèces).

Il est donc important que les producteurs des données transmettent l'ensemble de leurs données acquises par les observateurs (toutes les catégories même si les données ne sont pas valides - *catégorie 4 du Tableau 2*) pour permettre aux utilisateurs des données de choisir ce qu'ils veulent conserver en fonction du sujet de recherche. Par exemple, des chercheurs en SHS qui vont étudier l'intérêt, l'évolution et l'efficacité des programmes de SP marines.

Tableau 2 : Tableau final de référence pour la validation des données de SP du Collectif Vigie Mer (CVM) avec la correspondance au standard INPN

Code SINP	***Libellé SINP***	***Définition SINP***	**Code CVM**	**Libellé CVM**	**Définition CVM proposée**	**Critères CVM d'attribution à la catégorie**	**Remarques CVM**	**Exemples**
1	*Certain (très probable)*	*La donnée est exacte. Il n'y a pas de doute notable et significatif quant à l'exactitude de l'observation ou de la détermination du taxon. La validation a été réalisée notamment en présence d'une preuve de l'observation qui confirme la détermination du producteur ou après vérification auprès de ce dernier*	1a	Certain (et prouvé)	La donnée est exacte. La validation a été réalisée en présence d'une preuve de l'observation qui confirme la détermination de l'observateur et qui peut être consultable par un tiers	Existence de preuves matérielles bancarisées : images (photo, vidéo), sons, prélèvements (échantillon ADN, spécimen, etc.)	Proposition de mettre en place une double-validation. Importance de la bancarisation des images, sons, prélèvements, etc.	Observation d'un mérou avec une photo prise par l'observateur
			1b	Quasi certain	La donnée présente un très bon niveau de fiabilité. Il n'y a pas de doute notable et significatif quant à l'exactitude de l'observation ou de la détermination du taxon **MAIS** il n'y a pas de preuve de l'observation consultable par un tiers pour en attester	1/ Observateur fiable (expert, ou ayant reçu une formation adaptée) **OU** 2/ donnée validée par un expert suite à un échange avec l'observateur sur la base de critères précis		1/ Observation d'un mérou en Méditerranée par un observateur formé/habitué à l'identification de cette espèce **MAIS** sans photo 2/ Observation d'un requin pèlerin par un observateur qui en voit un pour la première fois **MAIS** qui décrit précisément l'animal au validateur

Code SINP	***Libellé SINP***	***Définition SINP***	**Code CVM**	**Libellé CVM**	**Définition CVM proposée**	**Critères CVM d'attribution à la catégorie**	**Remarques CVM**	**Exemples**
2	*Probable*	*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*	2	Probable	La donnée est vraisemblable et crédible au regard des connaissances sur le taxon **MAIS** il n'est pas possible d'obtenir des informations supplémentaires permettant d'attester positivement sa validité	Conditions de l'observation totalement compatibles avec la répartition connue et l'écologie du taxon identifié		Observation d'un mérou en Méditerranée par un observateur dont on ne connaît pas le niveau d'expertise pour l'identification de cette espèce **ET** avec qui le validateur ne peut pas échanger

Code SINP	***Libellé SINP***	***Définition SINP***	**Code CVM**	**Libellé CVM**	**Définition CVM proposée**	**Critères CVM d'attribution à la catégorie**	**Remarques CVM**	**Exemples**
3	*Douteux*	*La donnée est peu vraisemblable ou surprenante mais on ne dispose pas d'éléments suffisants pour attester une erreur manifeste*	3	Très incertain	La donnée est douteuse/très improbable, mais on ne dispose pas d'éléments suffisants pour attester d'une erreur manifeste. Il n'est donc pas possible de l'invalider		C'est souvent la catégorie où atterrissent les nouvelles signalisations géographiques	Observation d'un mérou en Mer du Nord
4	*Invalide*	*La donnée a été infirmée (erreur manifeste/avérée) ou présente un trop bas niveau de fiabilité. Elle est considérée comme trop improbable (aberrante notamment au regard de l'aire de répartition connue, des paramètres biotiques et abiotiques de la niche écologique du taxon, la preuve révèle une erreur de détermination)*	4	Invalide	La donnée a été infirmée. Il existe une preuve qui révèle une erreur de détermination et/ou la donnée est aberrante	Erreur manifeste d'identification ou erreur dans les informations associées	Attention, toute observation, même très douteuse, dont on ne peut pas prouver l'invalidité devrait être dans la catégorie 3 "Très incertain"	1/ Observation d'un mérou dans un lac d'altitude 2/ Signalement d'un requin pèlerin mais photo d'un poisson lune

Code SINP	***Libellé SINP***	***Définition SINP***	**Code CVM**	**Libellé CVM**	**Définition CVM proposée**	**Critères CVM d'attribution à la catégorie**	**Remarques CVM**	**Exemples**
5	*Non réalisable*	*La donnée a été soumise à l'opération de validation mais l'opérateur (humain ou machine) n'a pas pu statuer sur le niveau de fiabilité. Notamment : état des connaissances du taxon insuffisantes (par l'opérateur ou en général), informations insuffisantes sur l'observation.* *NB C'est la valeur par défaut (niveau initial ou temporaire)*	5	Non réalisable	La donnée a été soumise à l'opération de validation mais l'opérateur (humain ou machine) n'a pas pu la classer dans l'une des catégories de validation précédentes		L'utilisation de cette catégorie doit rester exceptionnelle	1/ On manque de critères visuels (morphologie et coloration) pour différencier le gobie *Zebrus pallaoroi* (récemment décrit) de *Zebrus zebrus*, même sur la base d'une bonne photo sous-marine. En l'attente d'un progrès des connaissances sur leur aspect in vivo, les observations de ces deux espèces entrent dans cette catégorie 2/ Echange avec un observateur sur une possible observation d'un requin pèlerin mais impossible de savoir si c'est bien cette espèce qu'il a observée. Il décrit bien un requin mais l'observation a été trop brève 3/ Photo d'un requin observé depuis la côte, silhouette déformée par les reflets de l'eau, il est impossible de déterminer l'espèce

Code SINP	***Libellé SINP***	***Définition SINP***	**Code CVM**	**Libellé CVM**	**Définition CVM proposée**	**Critères CVM d'attribution à la catégorie**	**Remarques CVM**	**Exemples**
6	*Non évalué*	*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*	6	Non évalué	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			

En plus de ce travail, la dernière version mise à jour des variables du SINP est présentée en *Annexe 1*. Il s'agit des variables disponibles dans le standard principal, qui peut être associé à des extensions contenant d'autres variables

Annexe 1 : variables du SINP

Légende couleurs
Obligatoire
Obligatoire conditionnel
Recommandé
Facultatif

CHAMP	TYPE	DESCRIPTION
cleObs	VARCHAR	Attribut technique servant de clé primaire de l'observation
cleGrp	VARCHAR	Attribut technique servant à faire le lien avec un regroupement
statSource	VARCHAR	Indique si la donnée source (DS) de l'observation provient directement du terrain (via un document informatisé ou une base de données), d'une collection ou de la littérature
refBiblio	VARCHAR	Référence de préférence au format ISO690 de la source de l'observation
idJdd	VARCHAR	Code de la fiche de métadonnées associée au jeu de données
idOrigine	VARCHAR	Identifiant unique de la Donnée source de l'observation dans la base de données du producteur (où est stockée et initialement gérée la Donnée Source).
orgGestDat	VARCHAR	Nom de l'organisme qui détient la Donnée Source et qui en a la responsabilité
idSINPOcc	VARCHAR	Identifiant unique de la Donnée Élémentaire d'Échange de l'observation dans le SINP.
dSPublique	VARCHAR	Indique explicitement si la Donnée Source (DS) est publique ou privée
statObs	VARCHAR	Indique si le taxon a été observé (directement ou indirectement), ou non observé
cdNom	NUMBER	Code du référentiel taxonomique TAXREF pour chaque taxon
nomCite	VARCHAR	Nom du taxon cité à l'origine par l'observateur et géré dans la source
denbrMin	NUMBER	Nombre minimum d'individus du taxon composant l'observation
denbrMax	NUMBER	Nombre maximum d'individus du taxon composant l'observation
objDenbr	VARCHAR	Indique l'objet du dénombrement.
comment	VARCHAR	Champ libre pour informations complémentaires indicatives
dateDebut	DATE	Date du jour de l'observation dans le système grégorien (JJ/MM/AAAA)
dateFin	DATE	Date du jour de l'observation dans le système grégorien (JJ/MM/AAAA)
heureDebut	DATE	Heure et minute dans le système local auxquelles l'observation du taxon a débuté (hh:mm)
heureFin	DATE	Heure et minute dans le système local auxquelles l'observation du taxon a pris fin (hh:mm)
dateDet	DATE	Date de la dernière détermination du taxon de l'observation dans le système grégorien (JJ/MM/AAAA)
altMin	NUMBER	Altitude minimum de l'observation en mètre
altMoy	NUMBER	Altitude moyenne de l'observation en mètre
altMax	NUMBER	Altitude maximum de l'observation en mètre

CHAMP	TYPE	DESCRIPTION
profMin	NUMBER	Profondeur minimum de l'observation en mètre
profMoy	NUMBER	Profondeur moyenne de l'observation en mètre
profMax	NUMBER	Profondeur maximale de l'observation en mètre
nomLieu	VARCHAR	Nom propre du lieu où a été effectuée l'observation
cleObjet	VARCHAR	Attribut technique permettant de faire le lien avec l'objet géographique du fichier SIG « St_SIG »
x	NUMBER	Longitude, coordonnée X de l'observation
y	NUMBER	Latitude, coordonnée Y de l'observation
projection	VARCHAR	Code de la projection utilisée pour les coordonnées x et y.
X_PREC	NUMBER	Estimation en mètres du rayon d'une zone tampon autour de l'objet géographique suivant l'axe X (longitude). Cette précision peut inclure la précision du moyen technique d'acquisition des coordonnées (GPS,…) et/ou du protocole naturaliste.
Y_PREC	NUMBER	Estimation en mètres du rayon d'une zone tampon autour de l'objet géographique suivant l'axe Y (latitude). Cette précision peut inclure la précision du moyen technique d'acquisition des coordonnées (GPS,…) et/ou du protocole naturaliste.
natObjGeo	VARCHAR	Nature de la localisation transmise
floutage	VARCHAR	Indique si la donnée a été dégradée ou non
identObs	VARCHAR	Nom et prénom de la ou les personnes ayant réalisées l'observation
orgObs	VARCHAR	Nom de l'organisme ou des organismes du ou des observateurs dans le cadre du/desquels ils ont réalisé l'observation
detminer	VARCHAR	Nom, prénom et organisme de la ou des personnes ayant réalisé la détermination taxonomique de l'observation
orgDeterm	VARCHAR	Nom de l'organisme ou des organismes du ou des observateurs dans le cadre du/desquels ils ont réalisé l'observation.
nivValid	VARCHAR	Niveau de validité évalué par le producteur à la suite d'une validation formelle de la donnée.
datValid	DATE	Date à laquelle le niveau de validité a été affecté
validateur	VARCHAR	Nom, prénom et/ou organisme de la personne ayant réalisé la validation scientifique de l'observation
orgValid	VARCHAR	Nom de l'organisme ou des organismes du validateur.
protValid	VARCHAR	Protocole de validation utilisé : Adresse web à laquelle on pourra le trouver, ou référence bibliographique.