Université de Toulouse

MASTER 2 GÉOMATIQUE

« **Sc**I**e**nces **G**éomatiques en environne**M**ent et **A**ménagement » (**SIGMA**)

http://sigma.univ-toulouse.fr

RAPPORT DE STAGE

Conception d'une méthode de validation automatique de données d'observations naturalistes

CHEVALLIER Luc

Observatoire Aquitain de la Faune Sauvage

Maître de stage : Paul FROMAGE, Chargé de projets SI (OAFS)
Tuteur-enseignant : David SHEEREN, Maître de conférence (UMR DYNAFOR)

Septembre 2019

RÉSUMÉ

La connaissance de la validité des données est un enjeu majeur de la gestion des données naturalistes. Une donnée erronée peut en effet avoir de graves conséquences sur la crédibilité des démarches engagées en faveur de la biodiversité. Fort de ce constat, le Muséum National d'Histoire Naturelle (MNHN), garant du Système d'Information sur la Nature et les Paysages (SINP) et donc en charge de la bancarisation de l'ensemble des données naturalistes en France, cherche à mettre en place un système de validation permettant d'attribuer à chaque donnée un degré de fiabilité. L'Observatoire Aquitain de la Faune Sauvage (OAFS), en tant que plateforme thématique régionale SINP, suit la même logique pour la Nouvelle-Aquitaine et a fait le choix de mettre en place un processus de validation automatique. Ce stage, réalisé au sein de l'OAFS, a donc pour objectif l'élaboration et la mise en place de ce système.

Afin de mener à bien ce projet, une analyse de l'ensemble des méthodes de validation existantes a permis l'identification des critères nécessaires à l'évaluation du niveau de validité d'une donnée. Des bases de connaissances ont ensuite été imaginées, sélectionnées ou construites pour permettre le contrôle de ces critères, puis ces derniers ont été regroupés afin d'obtenir la note de validité globale de la donnée. La méthode construite a été transcrite sous la forme d'une chaîne de traitement avec le logiciel FME permettant son lancement en routine. Enfin, la chaîne de traitement a été testée sur les Mammifères d'Aquitaine, menant à des résultats satisfaisants.

Mots clés : Validation des données, Automatisation, Données naturalistes, Chaîne de traitement, FME

ABSTRACT

Knowledge of the validity of data is a major issue in the management of naturalistic data. Erroneous data can indeed have serious consequences for the credibility of the actions in favor of biodiversity. In light of this, the National Museum of Natural History (MNHN), the structure responsible for the Information System on Nature and Landscapes (SINP) and therefore in charge of the storage of all naturalist data in France, seeks to set up a validation system allowing to attribute to each data a degree of reliability. The Aquitaine Wildlife Observatory (OAFS), as a SINP regional thematic platform, follows the same logic for New Aquitaine and has chosen to set up an automatic validation process. This internship, realized within the OAFS, aimed to develop and install this system.

In order to carry out this project, an analysis of all the existing validation methods allowed the identification of the criteria required to evaluate the validity level of a given data. Reference documents were then imagined, selected or built to allow the control of these criteria. Then, they were grouped together to obtain the overall validity score of the data. The constructed method has been transcribed as a processing chain with the FME software, thereby allowing it to be launched on a routine basis. Finally, the treatment chain has been tested on the Mammals of Aquitaine, leading to satisfactory results.

Key words : Data validation, Automating, Naturalistic data, Processing chain, FME

REMERCIEMENTS

Je souhaite remercier toutes les personnes ayant participé à la réalisation de ce stage et toutes les personnes qui m'ont soutenu au cours de ces six derniers mois afin de le mener à terme :

- Paul FROMAGE et Marie BARNEIX, pour leur encadrement tout au long du stage, pour leurs conseils, pour leur aide, pour leur confiance et pour leur disponibilité. Je les remercie de m'avoir permis de réaliser ce stage qui aura été extrêmement enrichissant.

- Monsieur David SHEEREN, mon tuteur-enseignant, pour avoir répondu à mes questions et pour avoir suivi l'avancée de mon stage.

- L'ensemble de l'équipe de l'OAFS pour leur accueil et pour avoir participé à la bonne humeur qui aura régné durant la totalité du stage. Merci également pour leur aide et pour les réponses qu'ils ont apporté à mes questions sur le versant « biodiversité » durant le stage.

- Mes collègues stagiaires de l'UMR Biogéco, pour nos échanges qui auront également contribué à rendre ce stage agréable.

- Et enfin, merci à l'ensemble des personnes qui ont accepté de relire ce rapport afin d'en extraire la moindre coquille (tout du moins je l'espère !).

SOMMAIRE

TABLE DES FIGURES

TABLE DES TABLEAUX

LISTE DES SIGLES ET ABRÉVIATIONS

ABDSM	Atlas de la Biodiversité Départementale et des Secteurs Marins
BDD	Base de données
CarHAB	Cartographie nationale des habitats terrestres
CBNSA	Conservatoire Botanique Nationale Sud-Atlantique
CEN	Conservatoire des Espaces Naturels
CSRPN	Conseil Scientifique Régional du Patrimoine Naturel
DEE	Donnée Élémentaire d'Échange
DREAL	Direction Régionale de l'Environnement de l'Aménagement et du Logement
DS	Donnée Source
DSR	Donnée Source Régionale
ENS	Espace Naturel Sensible
ETL	Extract-Trasform-Load
FME	Feature Manipulation Engine
GBIF	Global Biodiversity Information Facility
INPN	Inventaire National du Patrimoine Naturel
INSPIRE	Infrastructure d'information géographique dans la Communauté européenne
IPBES	Plateforme intergouvernementale sur la biodiversité et les services écosystémiques
MEDDTL	Ministère de l'Écologie, du Développement durable, des Transports et du Logement
MNHN	Muséum National d'Histoire Naturelle
OAFS	Observatoire Aquitain de la Faune Sauvage
ORGFH	Orientations régionales de gestion et de conservation de la faune sauvage et de ses habitats
PLU	Plan Local d'Urbanisme
RCOAFS	Réseau des Contributeurs de l'OAFS
SI	Système d'Information
SI Faune	Système d'Information sur la Faune de l'OAFS
SINP	Système d'Information sur la Nature et les Paysages
SQL	Structured Query Language
TAXREF	Référentiel Taxonomique national
UMR	Unité Mixte de Recherche
UMS	Unité Mixte de Service
ZNIEFF	Zone Naturelle d'Intérêt Écologique, Faunistique et Floristique

INTRODUCTION

La biodiversité décline actuellement à un rythme sans précédent dans l'histoire humaine. La Plateforme intergouvernementale sur la biodiversité et les services écosystémiques (IPBES) estime en effet que près d'un million d'espèces sont aujourd'hui menacées d'extinction (IPBES, 2019), et cet effondrement n'est pas sans conséquence sur nos sociétés (Bradley et al., 2012). Face à ce constat, la préservation de la biodiversité devient un sujet d'importance majeure, et nécessite de connaître de mieux en mieux les espèces vivant sur notre territoire et leurs habitats. L'amélioration de ces connaissances entraîne alors la production d'une importante quantité de données sur la biodiversité, issues d'organismes ou de particuliers aux méthodes de collecte variées (Arpin, 2015). Cette hétérogénéité constitue cependant un frein au partage des données et à leur utilisation dans la production d'analyses à différentes échelles.

En réponse à cette problématique, le Système d'Information sur la Nature et les Paysages (SINP) a vu le jour en 2006 afin d'offrir un cadre méthodologique de référence pour le partage des données et des métadonnées d'observations naturalistes, facilitant ainsi la mise en relation des informations et des acteurs du monde naturaliste. Le Muséum National d'Histoire Naturelle (MNHN), porteur du SINP au niveau national, délègue ses missions de bancarisation des données auprès de plateformes thématiques régionales SINP. En Nouvelle-Aquitaine, cette mission a été confiée à l'Observatoire Aquitain de la Faune Sauvage (OAFS) pour l'ensemble des données sur la faune sauvage. Il assure ainsi, à partir du Système d'Information sur la Faune sauvage (SI Faune), le relais technique des échanges d'informations entre le niveau national et régional pour la faune.

L'hétérogénéité des sources de données et des données elles-mêmes n'est pas que problématique pour leur normalisation et leur diffusion, mais également vis-à-vis de leur qualité. En effet, les informations récoltées dans le SINP ne sont pas exemptes d'erreurs et il est parfois difficile de les identifier manuellement. Ainsi, le MNHN construit et teste depuis 2014 un système de validation dont l'objectif est de détecter automatiquement les données erronées. Ce système de validation est divisé en trois étapes dont la dernière, appelée validation scientifique, consiste à attribuer un niveau de fiabilité à la donnée selon des critères écologiques relatifs au taxon observé.

Les plateformes thématiques régionales du SINP, grâce à leur expertise locale et leur proximité avec les producteurs de données, ont pour objectif d'appliquer à leur échelle cette validation scientifique. Dans ce cadre, l'OAFS a choisi d'opter pour une méthode qui combine une validation automatique, basée sur des résultats d'expertise préalable, et une validation manuelle, à dire d'experts. Le développement de la méthode de validation scientifique automatique au sein de l'OAFS est l'objet de ce stage.

L'objectif principal de mon stage a donc été l'élaboration d'une méthode de validation automatisée d'observations naturalistes. Une analyse des méthodes existantes, ayant pour but de définir les critères écologiques à utiliser et les référentiels sur lesquels se baser, a ainsi été réalisée. Une fois ces points éclaircis, la méthode a été conçue puis mise en œuvre dans une chaîne de traitement. Enfin, une extension au SI Faune a été développée pour stocker et diffuser les résultats de cette validation.

Ce rapport présente tout d'abord le contexte de ce stage et l'état de l'art, puis décrit la méthode et l'extension mises en place. La présentation de quelques résultats d'analyse obtenus grâce à la méthode de validation scientifique automatique permettront enfin d'en évaluer l'efficacité.

1. CONTEXTE DU STAGE

1.1. Le SINP

En réponse aux problématiques soulevées durant la convention de Rio (1992) et aux obligations de mise à disposition des informations environnementales évoquées durant la convention d'Aarhus (1998) et dans la directive Inspire (2007), la Direction de la Nature et des Paysages du Ministère de l'Écologie, du Développement durable, des Transports et du Logement (MEDDTL) a mis en place, en 2007, le Système d'Information sur la Nature et les Paysages (SINP).

Le SINP est un dispositif partenarial entre le ministère en charge de l'écologie et les acteurs de la biodiversité. Il a pour objet la structuration des connaissances sur la biodiversité, la mise à disposition de ces connaissances et la facilitation de leur mobilisation dans l'élaboration de politiques publiques. Il fournit pour cela un socle méthodologique de référence facilitant la mise en relation d'informations (protocoles, standards de métadonnées et de données, référentiels, etc.).

Son organisation est fortement décentralisée et repose sur le niveau régional pour assurer l'animation locale du SINP. Cette tâche est confiée aux Directions Régionales de l'Environnement, de l'Aménagement et du Logement (DREAL). En Nouvelle-Aquitaine, la DREAL délègue ce travail à trois organismes composant la plate-forme régionale thématique SINP :

- l'Observatoire Aquitain de la Faune Sauvage (OAFS) pour la faune ;
- le Conservatoire Botanique National Sud-Atlantique (CBNSA) pour la flore, la fonge et les habitats ;
- la Réserve Naturelle Géologique de Saucats-la-Brède pour la géologie.

Les données produites par les acteurs naturalistes, appelées Données Sources (DS), sont transmises à ces trois pôles pour être analysées, structurées et bancarisées. Elles deviennent alors des Données Sources Régionales (DSR). Elles sont enfin transmises au Muséum National d'Histoire Naturelle (MNHN), gestionnaire des données au niveau national, pour devenir des Données Élémentaires d'Échanges (DEE).

Le SINP s'appuie sur un Système d'Information appelé Inventaire National du Patrimoine Naturel (INPN) afin d'assurer la gestion, la diffusion et la valorisation des DEE. Ce Système d'Information représente ainsi la partie visible du SINP au niveau national. Il joue également le rôle d'interface avec les instances internationales et notamment européennes comme INSPIRE, le GBIF (Global Biodiversity Information Facility) ou Natura 2000.

Ces missions de structuration et de bancarisation des données sur la biodiversité du SINP ont pour objectif final l'analyse des données et la réalisation de synthèses permettant l'orientation des politiques publiques directement auprès du ministère en charge de l'écologie. Dans ce cadre, la connaissance de la validité des données transmises est primordiale afin de ne pas transmettre de données erronées et ainsi de ne pas mettre en péril les démarches engagées en faveur de la biodiversité.

1.2. L’OAFS

La publication en 2006 des Orientations Régionales de Gestion et de conservation de la Faune sauvage et de ses Habitats (ORGFH) en Aquitaine met en avant l’absence d’une structure régionale pour fédérer et coordonner les nombreux acteurs du milieu faunistique sur le territoire (collectivités publiques, associations, bureaux d’études, laboratoire de recherches, etc.). Quelques mois plus tard, en 2007, le SINP voit le jour, créant également le besoin d’une structure régionale ayant pour rôle la structuration des données sur la faune (la flore étant confiée au CBNSA). La Direction Régionale de l’Environnement, de l’Aménagement et du Logement d’Aquitaine (DREAL Aquitaine) décide donc en 2011 de prendre les choses en main et met en place une mission de préfiguration afin d’identifier les besoins et de définir sous quelle forme sera élaborée la structure coordinatrice des connaissances faunistiques en Aquitaine.

C’est ainsi que s’est formé, deux ans plus tard, en 2013, l’Observatoire Aquitain de la Faune Sauvage, afin de coordonner le partage et la structuration des données naturalistes faunistiques, d’animer le réseau d’acteurs naturalistes et d’apporter une valorisation scientifique, fiable et partagée, à la production d’informations sur la faune régionale en Aquitaine puis en Nouvelle-Aquitaine.

Le portage de l'OAFS, d’abord confié à l’Université de Bordeaux, est aujourd’hui confié à l’Unité Mixte de Recherche "Biodiversité, Gènes et Communautés" de l'Université de Bordeaux avec l'appui administratif de l'Adera et l’appui financier de la DREAL, de la Région Nouvelle-Aquitaine et de l’Europe.

Son bon fonctionnement repose sur trois entités :

- **Le Réseau de Contributeurs** (RCOAFS), composé de nombreux organismes ayant adhérés à la charte partenariale, contribue à la connaissance et à l’expertise sur la faune sauvage de Nouvelle-Aquitaine.
    - Au sein de ce réseau, un **collectif d’experts régionaux** composé d’organismes spécialistes d’un ou plusieurs groupes taxonomiques à l’échelle régionale est coordonné par l’OAFS. L’une de ces missions est la participation à la validation des données naturalistes du SI Faune.
- La **cellule de traitement**, chargée de la réalisation des missions portées par l’OAFS, se divise en deux équipes. L’équipe SI est chargée de la gestion des données et du développement d’outils d’analyse et de diffusion. L’équipe Étude est chargée des projets d’analyse et de valorisation des connaissances (états des lieux, hiérarchisation des enjeux, listes rouges, etc.).
- Le **Conseil Scientifique Régional du Patrimoine Naturel** (CSRPN) oriente et valide les productions de la cellule de traitement.

La gestion des données au sein de l’OAFS est effectuée grâce au Système d’Information sur la Faune (SI Faune), nom attribué au système de gestion et de diffusion des données de l’OAFS. Ce système est composé d’une base de données (sous PostgreSQL v.9.3), d’une application web (sous Symfony v.2.8) et de routines informatiques (sous FME 2018) installées sur les serveurs. L’application web permet la mise en relation des acteurs naturalistes avec les données en leur permettant de les insérer, de les modifier ou de les exporter. Elle sert également de vitrine pour les productions internes.

L’OAFS, en tant que plateforme régionale SINP, assure la bonne structuration et la diffusion des données du SI Faune en accord avec le protocole national SINP, et accompagne les structures souhaitant participer à la démarche. Les missions de l’OAFS dépassent cependant ce cadre ; on peut citer par exemple la production de référentiels dressant l’état des connaissances actuelles sur les espèces en Nouvelle-Aquitaine, ou encore sa participation au secrétariat scientifique des Zones d’Intérêt Faunistique et Floristique (ZNIEFF) en Aquitaine.

Au niveau local, les données mutualisées par l'OAFS sont nécessaires à la mise en place de politiques locales telles que les Plans Locaux d'Urbanismes (PLU) établis par les communes ou leur communauté, les Espaces Naturels Sensibles (ENS) gérés par les départements ou encore les plans de gestion de la biodiversité menés par un réseau naturaliste varié. Le contrôle de la validité des données et la diffusion d'une information fiable et consolidée sont donc des objectifs primordiaux.

Par ailleurs, l'OAFS, bien qu'aujourd'hui considéré comme la plateforme thématique régionale du SINP vise actuellement l'obtention d'une habilitation rendant officiel son statut vis-à-vis du SINP (UMS PatriNat, 2018). L'observatoire doit pour cela répondre à un certain nombre de critères. La majorité d'entre eux, comme la mutualisation et la diffusion de l'ensemble des données publiques faunes sur le territoire, sont déjà effectifs, mais certains sont encore à développer. C'est le cas notamment de la mise en place d'un système de validation scientifique des données collectées.

1.3. La validation de données naturalistes

Depuis 2014, le MNHN travaille à la mise en place d'un système de validation permettant d'écarter les données erronées des synthèses sur la biodiversité. En effet, la validité de la donnée est un enjeu majeur de la gestion de l'information sur la biodiversité, qui conditionne à la fois la crédibilité de la démarche et la robustesse des analyses qui en résultent. Il est donc important que les différents utilisateurs soient en mesure de connaître le niveau de validité de chaque donnée partagée.

Dans ce cadre de l'élaboration d'un système de validation, les données étudiées doivent partager le même format. Il serait en effet beaucoup plus difficile, voire impossible, de valider des données hétérogènes, que ce soit manuellement ou automatiquement. Au sein du SINP, ce format standard, qui fait toujours l'objet d'un groupe de travail dont l'OAFS fait partie et qui en est aujourd'hui à la version 2.0, mentionne que l'observation naturaliste, également appelée « occurrence de taxon », doit porter avec elle, afin d'être complète et valorisable (Jomier, 2018) :

- le sujet de l'observation, à savoir le taxon observé (le Quoi) ;
- sa localisation (le Où), qui repose soit sur une localisation précise soit sur une entité géographique connue (commune, département, maille) ;
- sa date de réalisation (le Quand) ;
- les informations sur son auteur (le Qui) ;
- son protocole d'acquisition (le Comment).

Toute donnée ou métadonnée entrant dans le SINP doit donc être **conforme** à ce format, il doit en respecter les règles physiques et conceptuelles. Une occurrence de taxon doit également être **cohérente** vis-à-vis des informations qu'elle porte, lesquelles doivent suivre une logique combinatoire. Par exemple, une donnée dans le SI Faune porte deux dates : une minimale, l'autre maximale ; la première ne peut donc être supérieure à la seconde. Une fois les données conformes et cohérentes, elles sont **validées scientifiquement**. Cette validation scientifique se définit comme un processus d'expertise visant à renseigner sur le degré de confiance qui peut être accorder à la donnée. Ce stage ne porte que sur la validation scientifique des observations naturalistes.

Ces trois étapes de la validation d'une observation naturaliste, sont décrites dans un guide méthodologique qui a été publié courant 2016 par le SINP afin de définir les modalités dans lesquelles elle doit s'établir (Robert, 2016). Il propose ainsi une démarche et des concepts pour la réalisation d'une méthode de validation que chacun est libre d'utiliser. En tant que partenaire SINP, l'OAFS suit ses

prérogatives afin de fournir une méthode parfaitement compatible avec l'INPN, mais va parfois plus loin dans les détails.

La validation scientifique est déterminée à chaque étape de la vie de la donnée : par le producteur de la donnée (sur la DS), par le pôle thématique régional (sur la DSR) et par le MNHN au niveau national (sur la DEE).

Elle peut être réalisée soit :

- **automatiquement**, en testant au sein d'une chaîne de traitement un ensemble de critères s'appuyant sur des résultats d'expertises préalables (référentiels, couches atlas, etc.) ;
- **manuellement,** lorsqu'elle est réalisée par des experts régionaux ;
- de manière **combinée**, si les deux méthodes sont utilisées conjointement.

A minima, la validation scientifique consiste à contrôler la présence du taxon à une date et à un endroit donnés. Si seules ces trois informations sont validées, le processus est décrit comme étant de paramètre « **minimal** ». Si toute autre information jugée pertinente est examinée puis validée également (le stade de développement de l'animal par exemple), la validation scientifique est dite de périmètre « **élargi** ».

Au final, l'application de ce processus de validation scientifique doit aboutir à un niveau de validité permettant de connaître, avant chaque utilisation de la donnée, son degré de fiabilité. Dans le cas d'un processus automatique, le SINP recommande l'utilisation de 6 niveaux de validité :

- **Certain – très probable :** La donnée présente un haut niveau de vraisemblance (très majoritairement cohérente) selon la procédure automatique appliquée. Le résultat de la procédure correspond à la définition optimale de satisfaction de l'ensemble des critères de la procédure automatique, par exemple, lorsque la localité correspond à la distribution déjà connue et que les autres paramètres écologiques sont dans la gamme habituelle de valeur.
- **Probable :** La donnée est cohérente et plausible selon la procédure automatique appliquée mais ne satisfait pas complètement 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.
- **Douteux :** La donnée concorde peu selon la procédure automatique appliquée. 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).
- **Invalide :** La donnée ne concorde pas selon la procédure automatique appliquée. Elle présente au moins une discordance majeure sur un des critères jugés les plus importants ou la majorité des critères déterminants sont discordants. Elle est considérée comme trop improbable (aberrante notamment au regard de l'aire de répartition connue, des paramètres biotiques et abiotiques de la niche écologique du taxon). Elle est considérée comme invalide.
- **Non-réalisable :** La donnée a été soumise à l'ensemble du processus de validation mais l'opérateur (humain ou machine) n'a pas pu statuer sur le niveau de fiabilité (connaissances du taxon ou informations sur l'observation insuffisantes).
- **Non-évalué :** Niveau initial ou temporaire. 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.

2. OBJECTIFS ET ORGANISATION DU STAGE

La mise en place d'un système de validation scientifique des données mutualisées est nécessaire à l'habilitation finale de l'OAFS en tant que plateforme thématique régionale SINP. Afin de garantir un niveau de fiabilité correct, la méthode choisie par l'OAFS, dite « combinée », se compose de deux phases :

- Phase 1 : chaque donnée est soumise à un processus de validation automatique visant à lui attribuer un niveau de fiabilité.
- Phase 2 : un collectif d'experts évalue manuellement la validité scientifique des données pour lesquelles la validation automatique n'est pas suffisante, en se basant entre autres sur le niveau de fiabilité attribué automatiquement.

L'objectif de mon stage est de concevoir puis de mettre en place la phase 1 de cette méthode de validation scientifique des données faune. Dans le cadre du stage, cette méthode est temporairement limitée à l'ex-région d'Aquitaine, mais sera amenée à évoluer par la suite.

Afin d'aboutir à cet objectif, mon stage a été divisé en plusieurs phases :

- La première a été dédiée à la réalisation d'un état de l'art des méthodes de validation scientifique existantes afin d'identifier les principaux critères et bases de connaissances[1] utilisés.
- La deuxième a consisté à choisir les critères et les référentiels qui seront utilisés au sein de notre méthode puis à définir de quelle manière ils allaient être mis en place et regroupés pour aboutir à un niveau de validité automatique.
- Une fois la méthode conceptualisée, la troisième phase a eu pour objectif le développement d'une extension à la base de données pour la validation et la mise en place de la chaîne de traitement de la validation scientifique automatique afin de pouvoir la lancer en routine lors de l'insertion de nouvelles données.
- Enfin, la chaîne de traitement a été testée sur les Mammifères d'Aquitaine afin de s'assurer de son fonctionnement et du fonctionnement de chacun des critères.

Par ailleurs, une note méthodologique, décrivant la méthode développée au cours du stage, a été rédigée à destination du collectif d'expert afin d'être soumise à leur validation. Ce collectif s'est également chargé de la réalisation de certains référentiels qui seront présentés au cours de ce rapport.

L'organisation des tâches effectuées durant ce stage est présenté ci-après (Figure 1).

L'avancée du stage a été régulièrement suivie lors de réunions organisées avec mon maître de stage Paul FROMAGE et la responsable scientifique et technique de l'observatoire Marie BARNEIX ainsi que par mail avec mon tuteur-enseignant David SHEEREN.

1 Une base de connaissances constitue un ensemble structuré de connaissances sur un domaine spécifique. Par exemple, une base de connaissances peut être un référentiel de la difficulté de détermination des espèces d'Aquitaine ou encore une couche géographique de référence des aires de répartitions. Elle constitue donc des données de références pour la mise en place de la validation.

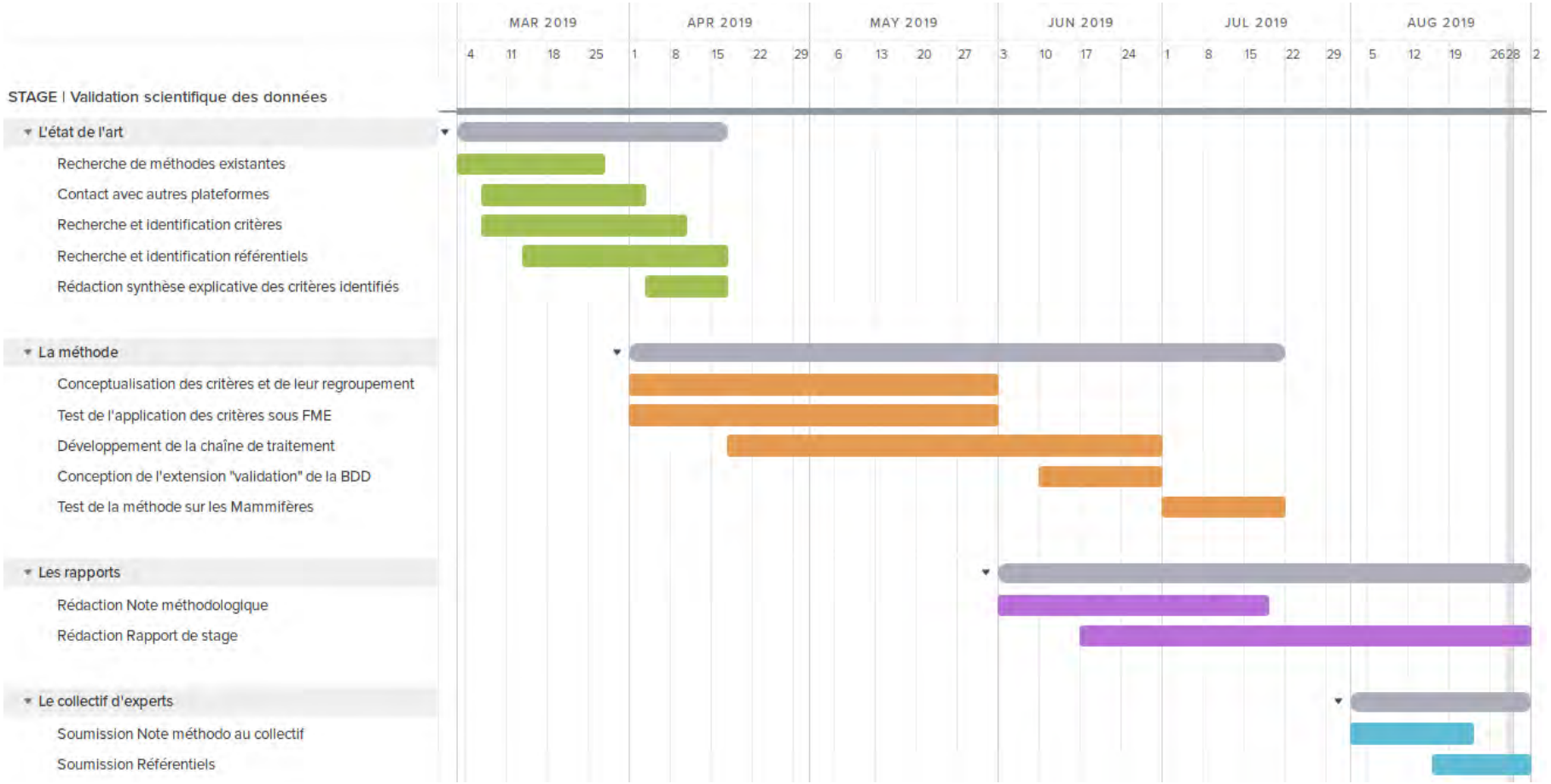

Figure 1 : Diagramme de GANTT récapitulatif de la planification de mon stage.

3. ÉTAT DE L'ART

3.1. Recherche de protocoles existants

Dans la mesure où le guide méthodologique sur la validation au sein du SINP (Robert, 2016) a été publié en 2016, certaines plateformes thématiques régionales SINP avaient déjà intégré au sein de leur Système d'Information une méthode de validation avant le début de mon stage. Par ailleurs, certaines structures externes au SINP, mais manipulant tout de même des données sur la biodiversité utilisent également un système de validation d'observations naturalistes.

Afin d'orienter les choix qui ont été réalisés pour le développement de la méthode de validation automatique, une analyse des méthodes existantes a été réalisée pour identifier, entre autres, les critères et bases de connaissances sur lesquelles elles reposent.

En 2015, Solène ROBERT du MNHN a réalisé un recensement des systèmes de validation existants dans le cadre de la mise en place du système de validation du SINP (Robert, 2015). Mes recherches se sont en partie appuyées sur ces travaux pour l'identification de quelques organismes gestionnaires de données naturalistes. En complément, une recherche approfondie de protocoles de validation a été réalisée parmi d'autres gestionnaires de données naturalistes (plateformes régionales du SINP ou autres institutions gérant des données naturalistes telles que les parcs naturels, les organismes publics et les associations) aussi bien en France qu'à l'étranger. Au total, 19 organisations gérant des systèmes d'information sur la faune et/ou la flore sauvage ont été contactées afin d'obtenir, s'il existe, le protocole de validation qu'ils utilisent.

Certains systèmes n'utilisant pas de méthode automatique ont aussi été étudiés puisque les critères sur lesquels se base la validation manuelle ne sont pas fondamentalement différents de ceux utilisés pour la validation automatique.

Un tableau récapitulatif de ces protocoles a été réalisé et contient les informations suivantes (Annexe 1) :

- nom de l'organisme ;
- date d'édition du protocole de validation ;
- valeurs utilisées comme résultat de la validation ;
- sectorisation, ou non, de la validation en protocoles distincts en fonction du groupe taxonomique étudié ;
- critères utilisés ;
- référentiels sur lesquels les critères sont basés ;
- méthode de regroupement utilisée.

3.2. Bilan des recherches

Sur les 19 organismes contactés, 11 ont répondu à notre demande et possédaient un protocole de validation automatique. Après analyse, voici les conclusions que nous en avons tirées.

3.2.1. Valeurs utilisées comme résultat de la validation

Concernant les valeurs utilisées comme résultat du processus de validation, plus de la moitié des méthodes (7) utilisent un système binaire : une donnée est soit valide soit invalide. Les autres méthodes utilisent entre 5 et 6 valeurs. Ces derniers correspondent davantage aux préconisations faites par le MNHN dans le guide méthodologique et offrent davantage de précision dans la description du niveau de fiabilité de la donnée.

3.2.2. Sectorisation du protocole par groupe taxonomique

Trois des protocoles étudiés sectorisent la méthode par groupe taxonomiques. Par exemple, le SINP de La Réunion utilise des méthodes différentes pour les cétacés, les tortues marines, les mammifères non-volants, etc. Cela permet d'adapter la méthode aux spécificités de certains groupes, mais rend plus compliquée la maintenance et l'évolution de la méthode, et multiplie le temps de la conception par le nombre de groupes différents pris en compte.

3.2.3. Critères fréquemment utilisés

Le niveau de validité d'une donnée se base sur l'évaluation d'un ensemble de critères. L'analyse des différentes méthodes a permis de les rassembler en critères plus généraux afin de pouvoir étudier leur fréquence d'utilisation (Figure 2). Le critère le plus fréquemment utilisé est l'aire de répartition : il s'agit de comparer la localisation de l'observation avec l'aire de répartition du taxon observé. Viennent ensuite à parts égales la phénologie (contrôle de la date d'observation en fonction de la phénologie du taxon), la difficulté de détermination et la rareté du taxon observé. Ces quatre premiers critères correspondent en fait à la validation « minimale » selon le SINP, à savoir le contrôle de la présence du taxon à un endroit et à une date donnés. Nous retrouvons ensuite des critères moins souvent utilisés, mais qui peuvent tout de même être intéressant pour la validation scientifique.

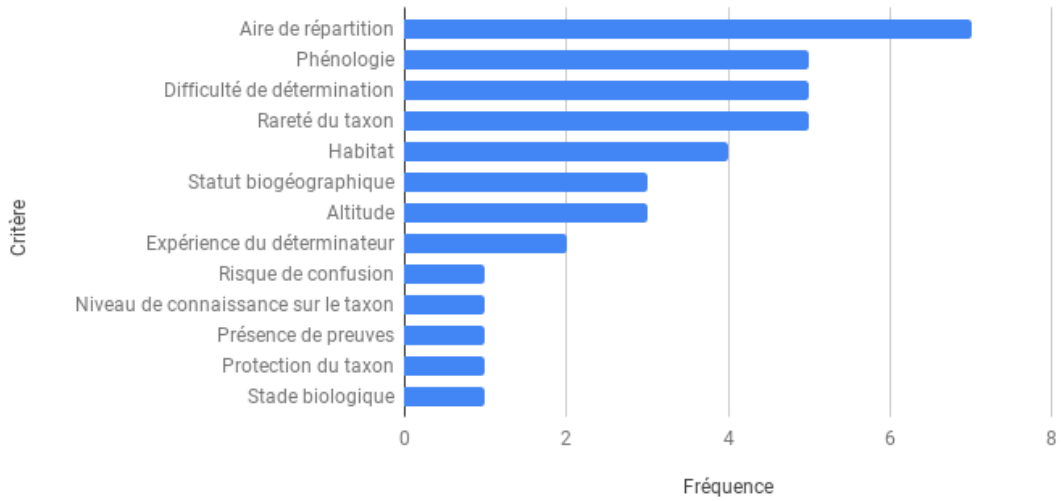

Figure 2 : Fréquence des critères utilisés dans les protocoles étudiés.

3.2.4. Données de références

Les critères précédemment cités s'appuient sur des données de références ou « bases de connaissances ». Parmi les méthodes étudiées, 3 grands types de références ont été identifiés :

- Les référentiels « maison » qui sont des listes d'espèces créées spécialement pour la validation scientifique. Par exemple, le Conservatoire Botanique National Alpin s'appuie sur une liste de taxons difficiles à déterminer, réalisée en interne, pour lesquels une validation manuelle est obligatoire.
- Les référentiels externes qui sont des référentiels déjà existant réutilisés pour la validation. Le plus fréquemment utilisé ici est TAXREF (le référentiel taxonomique national français (Gargominy, 2018) mis en place et géré par le MNHN).
- Des cartes de congruence construites à partir des données déjà validées. Le Conservatoire des Espaces Naturels du Languedoc-Roussillon base par exemple la totalité de la validation scientifique sur ce principe. Ce système nécessite en revanche d'avoir déjà à disposition un ensemble conséquent de données valides. Il peut par ailleurs entraîner des erreurs en cascade dans le cas où des données erronées auraient été validées par le passé.

3.2.5. Méthode d'attribution du niveau de validité final

Si la méthode se base sur plusieurs critères, il est nécessaire de mettre en place une méthode de regroupement afin d'aboutir au niveau de validité final. Parmi les méthodes analysées, seulement trois utilisent une méthode de regroupement (les autres n'utilisant qu'un seul critère ou réalisant le regroupement manuellement). Parmi eux, deux prennent uniquement en compte le critère le plus pénalisant ; les critères ne se compensent donc pas entre eux et chacun d'eux peut entraîner l'invalidité de la donnée. La troisième méthode se base quant à elle sur un arbre décisionnel (ou logigramme) aboutissant à une note de validité finale par la combinaison des différents critères.

3.2.6. Conclusion de l'état de l'art

D'une part, les résultats de cette analyse ont mis en avant que la majorité des méthodes de validation scientifique existantes sont relativement simples et reposent en grande partie sur la validation manuelle. D'autre part, cet état de l'art à permis d'identifier un nombre conséquent de critères qui ont permis d'alimenter les réflexions lors de la construction de la méthode de validation scientifique automatique de l'OAFS.

4. CONCEPTION DE LA MÉTHODE

Suite à l'analyse des travaux de validation de données naturalistes existant, l'objectif a été de construire la méthode dans le respect des recommandations du SINP. Cette étape de conceptualisation a nécessité la prise de nombreuses décisions, pour les choix des critères et des données de référence nécessaires, mais aussi concernant la structuration de la méthode. Ces choix ont été basés sur un compromis entre faisabilité (disponibilité et charge de travail pour les experts, qualité pas forcément suffisante des données disponibles) et gain potentiel vis-à-vis de la qualité de la validation automatique.

Par exemple, il a été décidé de ne pas utiliser de méthodes différentes selon le groupe taxonomique pour limiter les coûts de maintenance et d'intervention des experts, bien que cela améliorerait sûrement la qualité de la validation. Le rapport faisabilité/intérêt pour la méthode n'a pas été jugé suffisant. À l'inverse, il a été choisi d'utiliser les niveaux de validation préconisés par le SINP plutôt qu'un simple niveau binaire (valide/invalide) car il n'implique pas de problématique de faisabilité tout en améliorant la qualité et la précision de la méthode.

4.1. Sélection des critères, choix et limites rencontrées

La méthode de validation automatique de données d'occurrence s'effectue via le contrôle d'un certain nombre de critères pouvant justifier la fiabilité de l'observation naturaliste. Suite à l'état de l'art réalisé en première partie de stage, complété par des propositions internes, dix-neuf critères potentiels ont été identifiés. Afin de juger de leur utilité, chacun d'entre eux a été analysé lors de plusieurs réunions de travail organisées avec l'équipe de l'OAFS tout au long de mon stage. Un tableau synthétisant les critères et les réflexions autour de chacun d'entre eux a été mis en place comme support de discussion (Annexe 2).

Les critères ayant été étudiés sont les suivants :

1. Difficulté de détermination
2. Rareté du taxon
3. Risque de confusion
4. Présence du taxon dans TAXREF
5. Statut de protection du taxon
6. Répartition géographique
7. Habitat
8. Répartition altitudinale
9. Statut biogéographique
10. Répartition temporelle
11. Expérience du déterminateur
12. Existence d'une preuve
13. Existence d'un protocole d'acquisition
14. Niveau de validité fourni par le producteur
15. Dénombrement
16. Processus d'observation
17. Comportement
18. Stade de développement
19. Statut biologique

Cette phase de réflexion et de construction à permis de mettre en avant les éléments suivants :

- La difficulté de détermination d'une espèce (1) ne permet pas, seule, de valider ou d'invalider une donnée. En effet, même une espèce difficile à déterminer peut être observée si le naturaliste possède de bonnes compétences. Ce critère pourrait cependant servir à ajuster un niveau de validité existant.

- La difficulté de détermination (1) d'une espèce pouvant varier selon son sexe ou son stade de développement, il est envisageable de prendre en compte ces informations pour la résolution du critère, tout en gardant à l'esprit que ces informations sont facultatives dans le SI Faune et donc peu renseignées.

- Le risque de confusion avec une autre espèce (3) relevant de la difficulté de détermination, ces deux critères doivent être regroupés.

- La rareté du taxon (2) et son(ses) statut(s) de protection (5) ne reflètent en rien la validité de la donnée. Il n'est pas possible de dire qu'une donnée est invalide sous prétexte que le taxon est rare ou protégé. Ces critères peuvent cependant être utiles pour mettre en avant des données dont la validation serait prioritaire.

- Le contrôle de la présence du taxon dans le référentiel taxonomique national TAXREF (4) est réalisé implicitement durant l'insertion des données dans la base. Ce critère ne sera donc pas pris en compte.

- Le contrôle de la localisation de l'observation selon la répartition géographique du taxon (6) est un critère essentiel. Il permet d'identifier les observations réalisées en dehors de l'aire de répartition connue du taxon. Le critère concernant l'habitat du taxon (7), en lien direct avec la localisation de l'observation, peut être complémentaire à ce dernier. Au contraire, l'altitude de l'observation (8) et les statuts biogéographiques (9) n'ont que peu d'intérêt, car ils traduisent généralement l'habitat et l'aire de répartition des espèces, qui sont déjà pris en compte (6 et 7).

- La comparaison de la date d'observation d'une espèce avec sa période d'observation connue (10) est un critère efficace dans la validation d'une donnée d'occurrence de taxon. Il permet d'invalider des observations faites à une période durant laquelle l'espèce n'est pas observable.

- L'expérience du déterminateur (11) pourrait être prise en compte et permettrait d'ajuster le critère de difficulté de détermination. Cependant, les données naturalistes sont produites par de nombreux acteurs, professionnels ou particuliers, aux compétences très hétérogènes, et la mise en place de ce contrôle nécessiterait de connaître et de classifier l'expérience de chacun d'entre eux, ce qui n'est pas envisageable.

- La mention de l'existence d'une preuve (12) ou l'application d'un protocole d'acquisition (13) ne sont pas réellement des indices de la fiabilité d'une observation. L'utilisation de preuves numériques pourrait être intéressante, mais nécessiterait néanmoins de savoir reconnaître une espèce depuis une photo, ce qui n'est actuellement pas possible automatiquement.

- Avant l'intégration des données dans le SI-Faune, le producteur de la donnée peut avoir déjà validé son observation (14). Cette information peut alors être utilisée afin de prendre en compte de manière prioritaire les observations identifiées comme douteuses par le producteur. Ce critère est intéressant car le producteur se base souvent sur des informations supplémentaires non disponibles dans le SI Faune.

- Le contrôle du dénombrement (15) permet d'invalider des observations où le nombre d'individus observés n'est pas cohérent avec l'espèce. Il n'est par exemple pas possible de voir 1000 ours dans les Pyrénées. Cependant, ce critère consiste davantage à contrôler la cohérence de la donnée (respect de la logique combinatoire des données et des métadonnées) que la validation scientifique.

- Les derniers critères évalués (16 à 19) consiste à vérifier que les informations renseignées pour le processus d'observation, le comportement, le stade de développement et le statut biologique soient cohérentes avec l'espèce observée. Ils relèvent donc eux aussi du contrôle de la cohérence mentionné au point précédent.

Dans un premier temps, l'OAFS préfère mettre en place une validation scientifique de périmètre « minimal » au sens du SINP, qui consiste à ne contrôler que la présence d'un taxon à un endroit et à un moment donné. Ce choix, toujours pris dans une logique de compromis entre la faisabilité de mise en place de la méthode et le gain potentiel généré, a également participé aux décisions prises lors du choix des critères de notre méthode.

En regard des éléments mentionnés ci-dessus, cinq critères ont été retenus :

- « **la validation du producteur** » afin d'identifier les données qui auraient déjà été identifiées douteuses ou invalides par le producteur de données puisqu'il dispose généralement d'informations supplémentaires sur l'observation ;
- « **l 'habitat** » dont l'objectif est d'identifier les données dont la localisation ne correspondant pas avec l'habitat connu du taxon ;
- « **la répartition géographique** » pour identifier les données dont la localisation ne correspond pas avec la répartition connue du taxon ;
- « **la période d'observation** » afin d'identifier les données dont la date ne correspond pas avec la période d'observation connue du taxon ;
- « **la difficulté de détermination du taxon** » qui a pour objectif d'affiner le résultat de la validation selon la difficulté de détermination du taxon observé.

4.2. Choix des données de références et construction des critères

Après avoir identifié les critères pertinents, il reste à définir comment ils seront utilisés et sur quelles données de référence ils se baseront.

Le choix des données de référence s'appuie sur l'état de l'art réalisé au début de mon stage et a été décidé lors de réunions organisées avec Marie BARNEIX et Paul FROMAGE. Les choix d'attribution des niveaux de fiabilité, présentés ci-après, ont été réalisés avec l'appui de l'expertise faunistique de l'OAFS. Ils pourront cependant être sujets à modifications une fois la méthode étudiée par le collectif d'experts régionaux.

4.2.1. La validation du producteur

Le contrôle de la validation du producteur (organisme ayant produit la donnée) se base uniquement sur le champ correspondant de la base de données qui stocke, s'il existe, le résultat de la validation réalisée par le producteur. Aucune donnée de référence n'est donc nécessaire pour ce critère.

Le test consiste donc à regarder le niveau de validation attribué à la donnée par le producteur sachant qu'il en existe cinq différents : « Certain - très probable », « Probable », « Douteux », « Invalide » et « Non réalisable ». L'objectif est d'identifier les données considérées comme douteuses ou invalides par le producteur lui-même. Cette information étant facultative, elle peut ne pas être renseignée : dans ce cas, ce critère n'est pas applicable à la donnée. Les tests effectués sont récapitulés dans la figure 3.

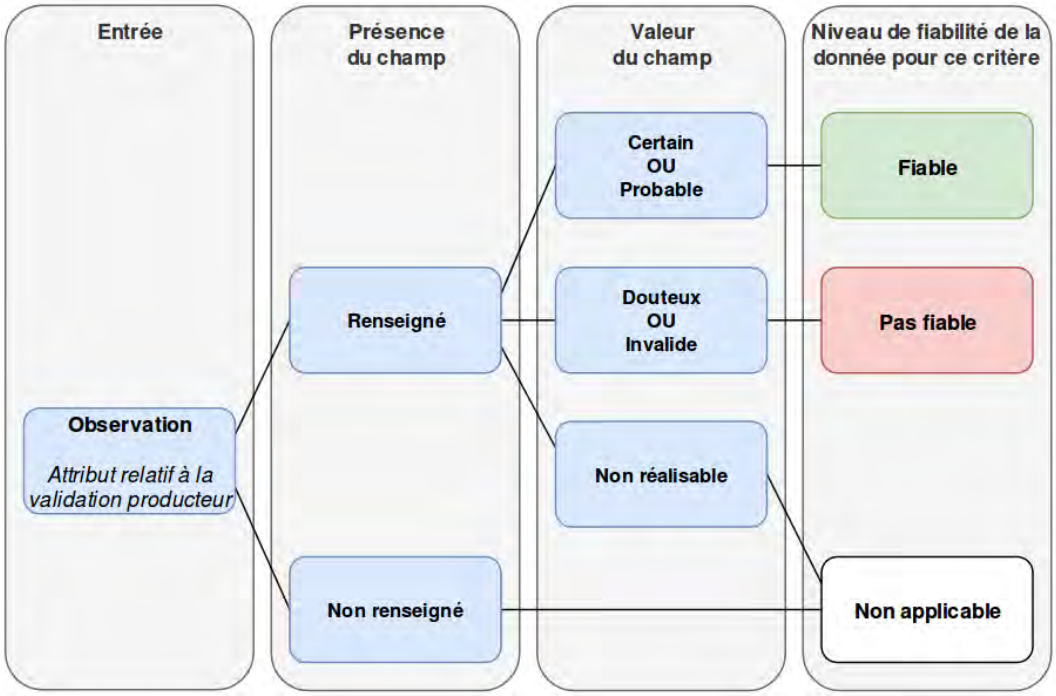

Figure 3 : Schéma de fonctionnement du critère « Validation du producteur »

4.2.2. L'habitat

Afin de mettre en place le critère contrôlant l'habitat du taxon, il est nécessaire de disposer de deux informations : l'habitat correspondant à la localisation de l'observation et l'habitat préférentiel du taxon observé.

La première information nécessiterait l'existence d'une cartographie précise des habitats d'espèces en Nouvelle-Aquitaine, ce qui n'est pas le cas actuellement. Certaines portions du territoire sont cartographiées par le CBNSA, mais cela reste anecdotique. Il n'est donc pas possible de savoir précisément sur quel habitat se situe l'observation à partir de sa localisation. L'utilisation d'une couche d'occupation du sol en remplacement a été envisagée, mais son intérêt au niveau écologique est limité, et il serait par ailleurs nécessaire de connaître l'occupation du sol de prédilection de chaque espèce d'Aquitaine.

En tant que première approche, et afin d'utiliser ce critère, il a été décidé d'utiliser un test simple : le contrôle de l'habitat à partir de la distinction entre le milieu marin et le milieu continental. Pour ce faire, deux couches géographiques ont été construites. La première, définissant le milieu continental, se base sur les limites administratives Françaises issues de la base de données ADMIN-EXPRESS du 14 mars 2019 auxquelles a été ajouté un tampon de 100 mètres pour ne pas exclure les observations d'animaux terrestres (notamment d'oiseaux) s'aventurant proche des côtes. La seconde couche, définissant le domaine marin, a été réalisée par différence avec le milieu continental. Un tampon de 100 mètres est également appliqué à cette couche afin de prendre en compte les observations d'animaux échoués (notamment des tortues et des cétacés). Ces deux couches permettent de connaître l'habitat « général » sur lequel a été réalisé l'observation à partir de sa localisation.

Comme expliqué précédemment, il est également nécessaire de savoir quel est l'habitat de prédilection du taxon observé. Cette information est justement disponible dans le référentiel taxonomique national TAXREF. Pour les besoins de nos tests, les habitats renseignés dans ce référentiel ont été simplifiés (Annexe 3). TAXREF, de par son statut de référentiel national, est la base de connaissances la plus complète et la plus fréquemment mise à jour.

Le contrôle de l'habitat du taxon consiste donc à tester la cohérence entre l'habitat, marin ou continental, sur lequel a été effectué l'observation et l'habitat de prédilection du taxon observé. Ce critère n'a que deux résultats possibles : soit les habitats correspondent et dans ce cas l'observation est fiable vis-à-vis de ce critère, soit ils ne correspondent pas. Toutefois, bien que TAXREF soit l'un des référentiels les plus complets à ce jour, l'habitat de certaines espèces n'est pas renseigné : dans ce cas, ce critère n'est pas applicable. Le fonctionnement de ce critère est synthétisé ci-après (Figure 4).

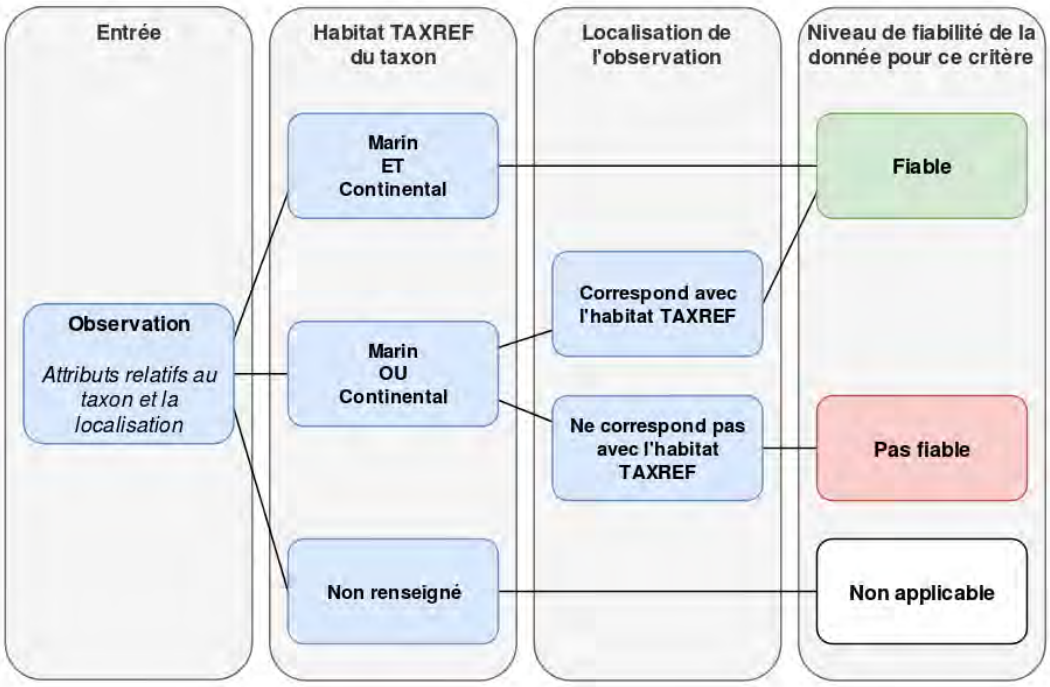

Figure 4 : Schéma de fonctionnement du critère « Habitat »

4.2.3. La répartition géographique

La vérification de la répartition géographique consiste à comparer la localisation de l'observation avec l'aire de répartition connue du taxon observé. Cette aire de répartition peut être renseignée à plusieurs échelles géographiques : nationale, régionale, départementale et précise (maillage 10*10km)[2].

Pour pouvoir utiliser ce critère, il faut donc disposer des aires de répartition connues citées précédemment. Pour chacune des échelles géographiques, voici les bases de connaissances étudiées et les conclusions des réflexions menées :

- à l'échelle **nationale**, la base TAXREF fait aussi office de référence géographique en fournissant par espèce un statut biogéographique traduisant sa présence à l'échelle métropolitaine (Annexe 4) ;
- à l'échelle **régionale**, le référentiel espèces mis en place par l'OAFS fournit pour certains groupes taxonomiques un ensemble d'informations sur les espèces présentes en Aquitaine, dont un statut biogéographique régional (Annexe 5).
- à l'échelle **départementale**, ce même référentiel espèces définit également un statut de présence pour chacun des cinq départements aquitains (Annexe 5). Il est intéressant de noter ici qu'un Atlas de la Biodiversité Départementale et des Secteurs Marins (ABDSM), géré par le MNHN, fournit également des statuts de présence départementaux, mais que ces derniers sont moins à jour que le référentiel de l'OAFS ;
- enfin, à l'échelle de la **maille 10*10km**, il existe en Aquitaine et pour certains groupes taxonomiques des atlas régionaux recensant les mailles de présence de chaque espèce pour une période donnée.

Afin d'être le plus complet et rigoureux possible, il a été décidé d'utiliser chacun des niveaux géographiques cités.

La vérification de la répartition géographique consiste donc à comparer la localisation de l'observation avec l'aire de répartition connue du taxon observé aux quatre échelles géographiques. Le but est de contrôler que le taxon soit présent au niveau métropolitain via TAXREF, puis au niveau régional et départemental via le référentiel espèces de l'OAFS et enfin à la maille via les atlas régionaux. Le niveau de fiabilité résultant de ce critère sera donc variable selon l'échelle géographique la plus fine correspondant à la localisation de l'observation. Le contrôle est considéré non applicable lorsque le taxon est présent en métropole, mais qu'il n'est pas étudié dans le cadre du référentiel espèces (n'importe quel orthoptère par exemple). Le fonctionnement de ce critère est synthétisé ci-après (Figure 5).

2 Dans le cadre des études sur la biodiversité, la grille nationale de carreaux de 10 par 10km est un système de référence. Elle est généralement utilisée comme unité de base pour la réalisation de calculs d'aires de répartition (MNHN, 2018).

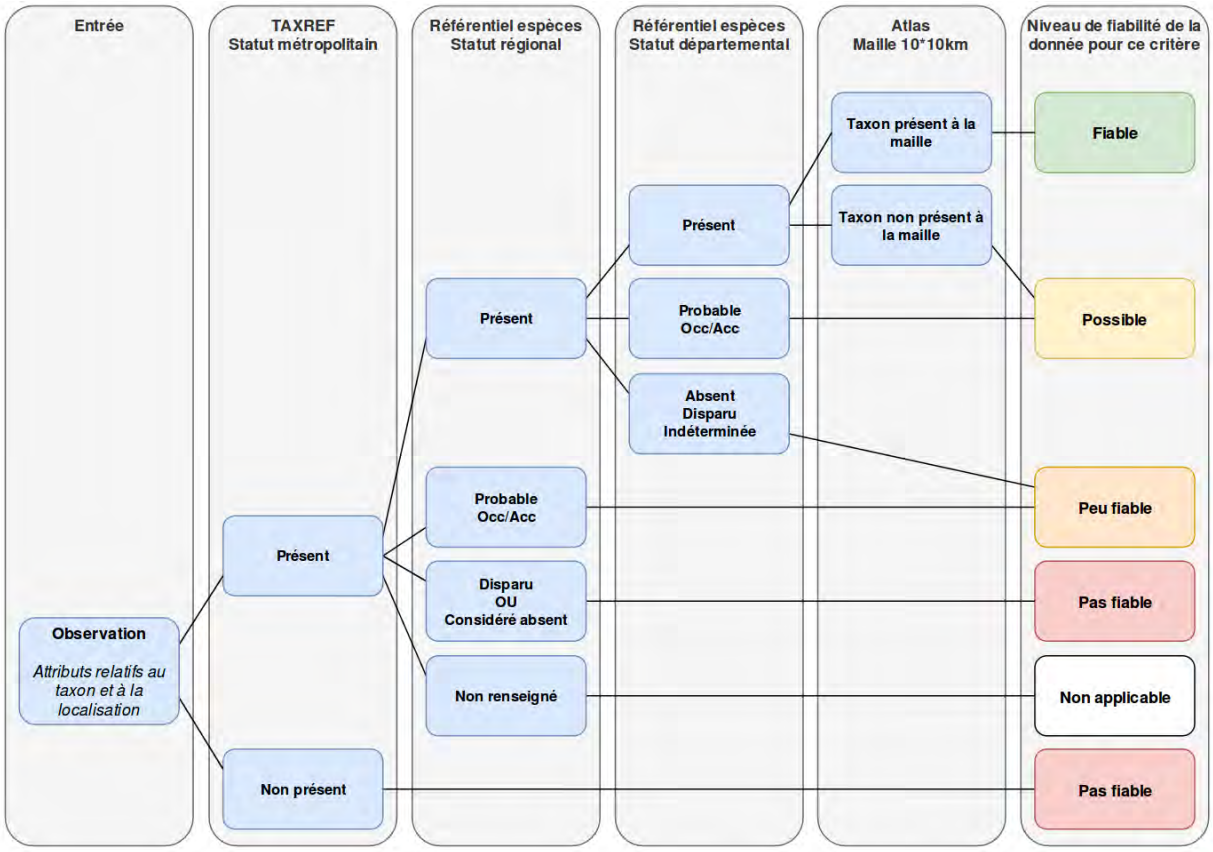

Figure 5 : Schéma de fonctionnement du critère « Répartition géographique »

Par ailleurs, il arrive parfois que l'OAFS reçoive des observations d'espèces non listées dans les référentiels espèces, alors pourtant que les groupes taxonomiques auxquels elles appartiennent ont été étudiés. Puisque les référentiels espèces sont exhaustifs (à une date donné, la biodiversité évoluant sans cesse), ces espèces se voient attribué pour les besoins de la validation scientifique un statut régional « Considéré absent ». Par exemple, les Reptiles formant un groupe étudié de manière exhaustive dans le cadre du référentiel espèces, l'observation d'une Couleuvre astreptophore (non listée dans le référentiel, car non présente en Aquitaine) ne sera « pas fiable ».

Afin de procéder à ces contrôles, le rattachement des observations à un département et à une maille 10x10km est nécessaire. Ce rattachement est précalculé dans le SI Faune (pour améliorer les performances de l'application web). Il arrive cependant qu'une observation intersecte plusieurs départements ou plusieurs mailles (du fait d'une localisation imprécise ou floutée). Dans ce cas, la note du critère est calculée pour chaque entité intersectée et la valeur finale correspond à la valeur maximale attribuée (fiabilité la plus élevée).

Enfin, la validation n'étant pas encore en place au sein du SI Faune, aucune donnée n'est actuellement validée dans la base de donnée (sauf rares exceptions validées manuellement). Il est ainsi actuellement impossible de baser la validation sur des cartes de congruence. Cette possibilité a tout de même été envisagée au sein de ce critère en complément aux données des atlas régionaux qui sont parfois anciennes ou lorsque le référentiel espèces est « non renseigné ». Cela pourra donc éventuellement entraîner la modification de la méthode de validation automatique lorsqu'un nombre suffisant de données seront validées.

4.2.4. La période d'observation

Contrôler la période d'observation d'un taxon nécessite en premier lieu la connaissance de la période de l'année durant laquelle l'observation du taxon est possible, voire fréquente. Pour certaines espèces, cette information est parfois disponible dans la littérature (atlas régionaux par exemple), mais il n'existe pas de référentiel complet pouvant être utilisé. De plus, il convient de bien faire la distinction entre période de présence (information souvent disponible) et période d'observation, une espèce pouvant être présente sans être observable (c'est le cas notamment des espèces hibernantes ou des espèces à stade de vie larvaire). Il a donc été statué que ce référentiel inexistant à l'heure actuelle sera réalisé par le collectif d'experts régionaux qui s'appuiera alors sur un gabarit décrivant les informations à renseigner.

Deux informations sont demandées à ce collectif d'experts : la période d'observation durant laquelle l'observation de l'espèce est courante, et une période plus large durant laquelle l'espèce peut être observée, mais plus rarement. Pour la première version de la méthode de validation automatique, ces périodes sont définies de manière globale pour le taxon, mais elles pourraient être précisées à moyen terme en fonction du sexe ou du stade de développement. Cet aspect n'est pour l'instant pas mis en place, car les attributs nécessaires de la donnée (sexe et stade de développement) sont facultatifs et rarement renseignés, ce qui ne justifie pas l'augmentation du coût que cela entraînerait vis-à-vis du collectif.

Le contrôle de la période d'observation se base donc sur le référentiel en cours de réalisation par le collectif d'experts. L'objectif est évidemment d'attribuer un niveau de fiabilité plus faible aux observations réalisées en dehors des périodes d'observation courantes ou peu fréquente pour l'espèce considérée. Certaines données (historiques notamment) ne possèdent pas de date précise d'observation, mais une période imprécise (souvent l'année). Si cette période excède la période d'observation, ce critère ne peut pas être appliqué. Ces contrôles sont illustrés en figure 6.

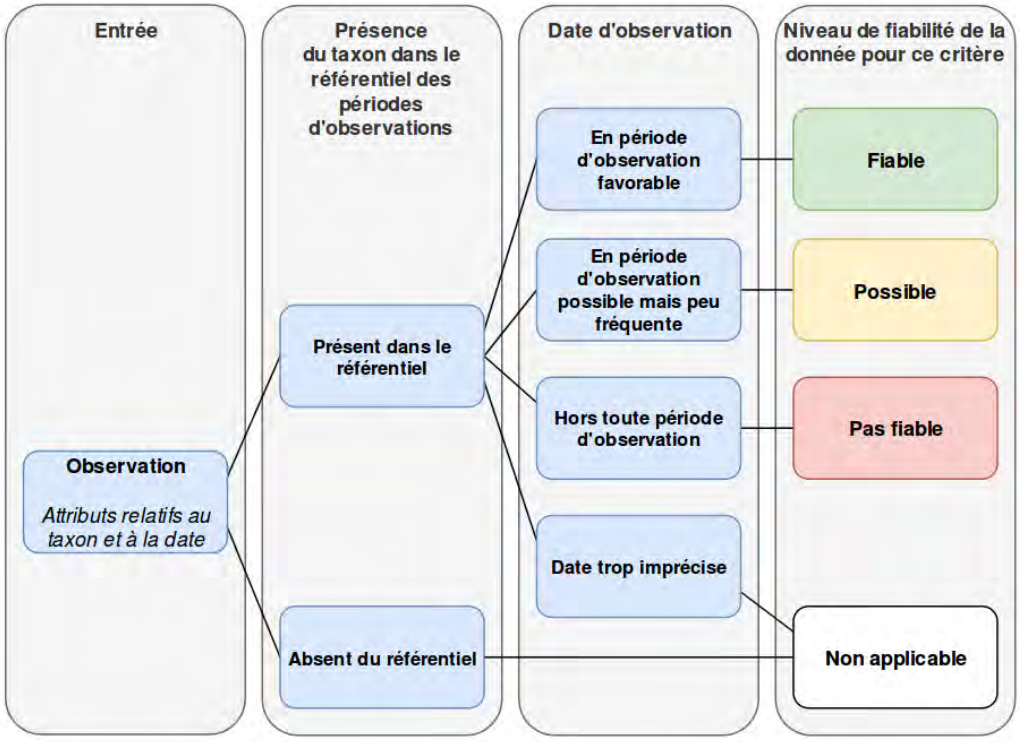

Figure 6 : Schéma de fonctionnement du critère « Période d'observation »

4.2.5. La difficulté de détermination

Certaines espèces sont plus faciles à identifier que d'autres. La mise en place de ce critère permet de prendre en compte cette information dans la méthode de validation scientifique. Il est donc nécessaire de disposer d'un niveau décrivant la difficulté de détermination du taxon observé. Après analyse de la littérature, il a été constaté que ce niveau est uniquement disponible pour les Rhopalocères (Papillons de jour) et les Odonates (Libellules) dans les pré-Atlas aquitains rédigés par le Conservatoire des Espaces Naturels (CEN) Aquitaine. De plus, ces informations ne sont pas disponibles sous forme numérique. Il a donc été décidé que ce référentiel, presque inexistant à l'heure actuelle, sera également réalisé par le collectif d'experts régionaux.

La construction de ce référentiel suit la règle suivante : chaque espèce inscrite doit disposer d'un niveau de difficulté de détermination général. Dans le cas où les spécialistes le jugent nécessaire, une même espèce peut disposer de niveaux supplémentaires pour des sexes et/ou des stades de développement particuliers. Contrairement au critère de « période d'observation », ces informations facultatives (sexe et stade de développement) sont ici prises en compte, car elles peuvent influencer fortement la difficulté de détermination. Les descriptifs sur lesquels se basent les niveaux de difficulté ont été définis à la suite d'une analyse bibliographique et sont présentés en annexe 6.

La difficulté de détermination, contrairement aux critères explicités précédemment, ne permet pas réellement de donner un degré de fiabilité à la donnée. Bien que difficile à déterminer, une espèce peut en effet avoir été identifiée correctement et inversement. Ce critère est donc utilisé en complément de ceux cités précédemment, pour affiner le niveau de validation final (Figure 7).

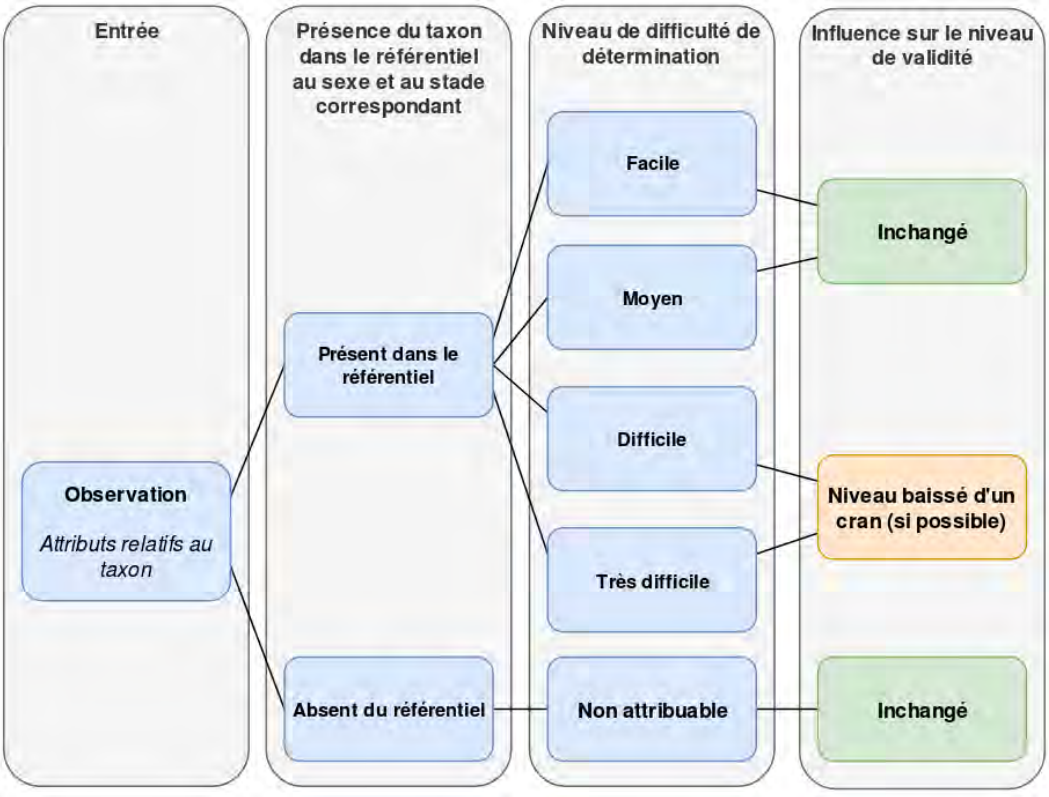

Figure 7 : Schéma de fonctionnement du critère « Difficulté de détermination »

4.2.6. Bilan des données de références utilisées

Comme mentionné plusieurs fois précédemment, la complétude des bases de connaissance utilisées peut varier en fonction des groupes taxonomiques. Le tableau ci-dessous synthétise les données disponibles ou en cours d'établissement en fonction des groupes étudiés.

Référentiel		Atlas aquitains	Référentiel espèces de l'OAFS (présence régionale et départementale)	Référentiel des périodes d'observation	Référentiel des difficultés de détermination
Groupe taxonomique	Oiseaux	Complet			
	Mammifères hors chiroptères	Complet	Complet	À consolider	À consolider
	Chiroptères	Complet	Complet	À consolider	En cours
	Amphibiens	Complet	Complet	À consolider	En cours
	Reptiles	Complet	Complet	À consolider	En cours
	Odonates	Complet	Complet	En cours	En cours
	Rhopalocères	Complet	Complet	À consolider	En cours
	Zygènes	Complet			
	Écrevisses	Complet	En cours	En cours	En cours
	Orthoptères		En cours	En cours	En cours
	Poissons		En projet	En projet	En projet

Tableau 1 : Tableau récapitulatif de la complétude des données de référence utilisées par groupe taxonomique

Chacune de ces bases de connaissances utilisées pour la validation automatique est renseignée à l'espèce. Cependant, afin de pouvoir attribuer un niveau de fiabilité aux observations supra-spécifiques (genre, famille, etc.), ils sont agrégés à partir des informations disponibles sur les taxons inférieurs. Par exemple, pour un genre donné, si au moins une espèce est considérée présente au niveau régional, le genre le sera également.

4.3. Choix et mise en place de la méthode de regroupement des critères

Le regroupement des critères permet l'attribution finale d'un niveau de validité à la donnée analysée. Il peut s'effectuer de plusieurs manières. Trois solutions ont été envisagées :

- retenir, parmi les critères testés, le niveau de fiabilité le plus invalidant comme niveau final. Cette solution, plutôt simpliste, ne permet pas de retranscrire les interactions potentielles entre les critères et n'a donc pas été retenue ;
- utiliser un modèle mathématique liant tous les critères entre-eux, en leur attribuant des notes de pondération. Cette méthode, permettant de hiérarchiser et d'assurer la prise en compte des liens logiques entre les critères, est cependant relativement compliquée à mettre en place dans le temps imparti par mon stage. De plus, l'objectif premier de cette méthode est son appropriation par le collectif d'expert, chargé d'avoir un avis critique, ce qui aurait été compliqué avec ce type de solution ;
- élaborer un arbre décisionnel afin d'établir un lien logique entre les différents critères. Cette méthode, utilisé par le MNHN pour valider les données de l'INPN, permet de prendre en compte aisément les interactions entre critères (du moment qu'elles ne sont pas trop nombreuses) et facilite la lecture de la méthode pour le collectif d'expert.

L'arbre décisionnel construit regroupe les niveaux de fiabilité de chaque critère comme illustré en figure 8. La difficulté de détermination du taxon, de par sa particularité, sert ici à orienter le niveau de validité déterminé en amont.

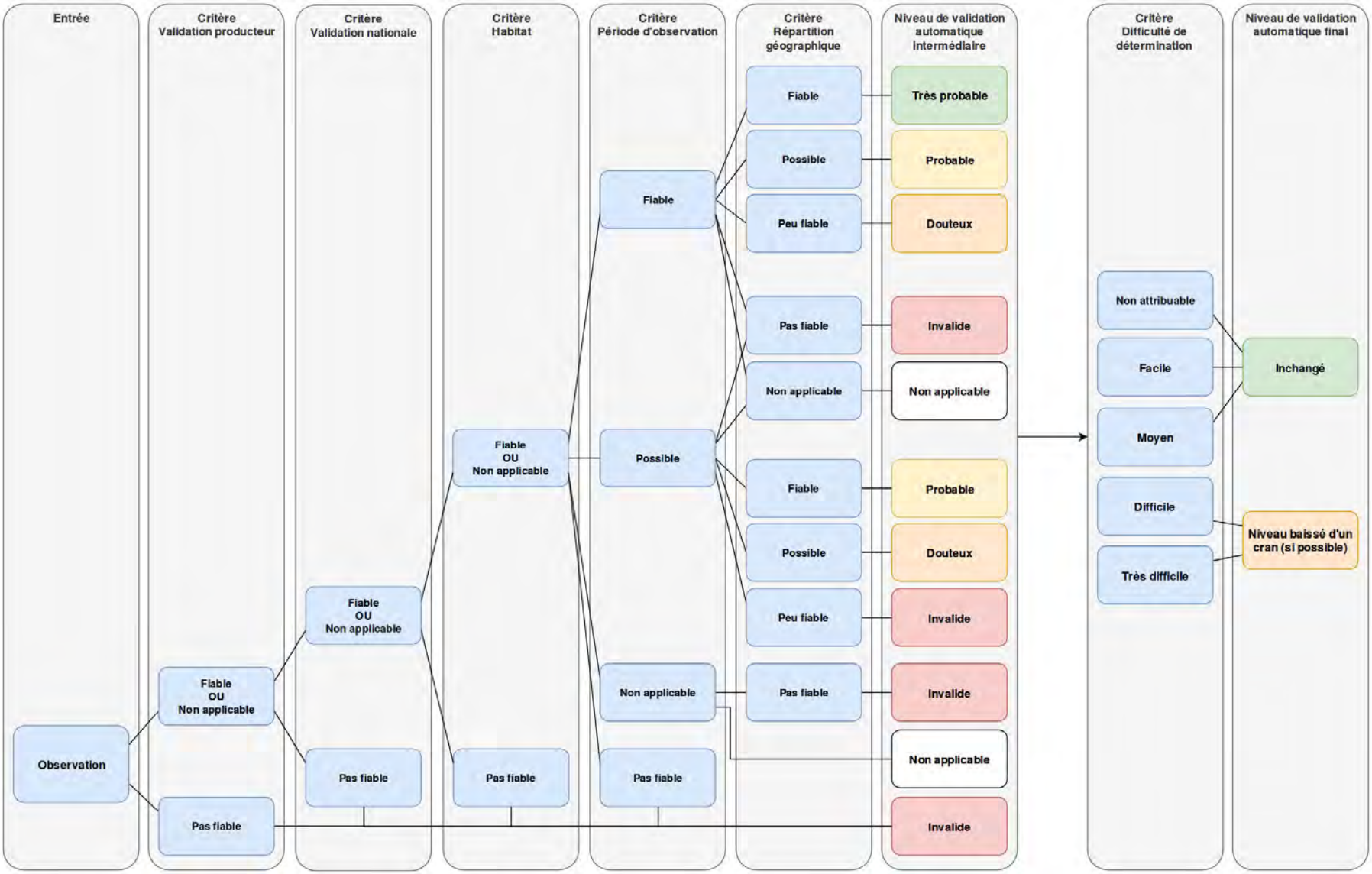

Figure 8 : Schéma du regroupement des critères pour l'attribution du niveau de validité automatique final

5. INTÉGRATION DE LA MÉTHODE AU SYSTÈME D'INFORMATION

Une fois la méthode de validation automatique de données d'occurrences de taxon construite, l'objectif a été de la transcrire sous la forme d'une chaîne de traitement informatique afin de pouvoir l'exécuter automatiquement lors de l'insertion de nouvelles données dans le SI Faune. Cette partie explicite le développement de l'extension à la base de données qui permettra le stockage des résultats de la validation puis les choix techniques réalisés pour la mise en place de la chaîne de traitement.

5.1. Conception de l'extension validation à la base de données

Le système de validation de l'OAFS a comme objectif le calcul et la diffusion d'un niveau de validité (automatique ou combiné) pour chaque donnée d'observation. Il est donc nécessaire de stocker les résultats obtenus dans la base de données, qui pourront ainsi être transmis au SINP et consultés sur le SI Faune. La diffusion sur la plateforme web permet notamment de communiquer aux experts les résultats détaillés de la validation automatique (résultats de chaque critère et note finale) afin de pouvoir les utiliser pour la phase 2 de la validation, soit la validation manuelle.

Une extension à la base de donnée du SI Faune (sous PostgreSQL v9.3) a donc été développée afin d'organiser le stockage de ces informations (aussi bien pour la validation automatique que manuelle). Deux formats ont été envisagés :

- concevoir une unique table dont le but est de stocker les résultats de la validation automatique et manuelle, un champ permet ensuite de savoir de quelle validation il s'agit ;
- stocker les informations dans deux tables distinctes, une pour la validation automatique et une pour la validation manuelle.

Certaines informations sont spécifiques à un type de critère. En effet, la validation automatique fournit un niveau de fiabilité pour chaque critère tandis que la validation manuelle ne défini qu'un niveau de validité final. Par ailleurs, la validation manuelle peut faire l'objet de commentaires (échanges entre les experts) contrairement à la validation automatique. Compte tenu de ces observations, il a été décidé d'utiliser deux tables pour distinguer les informations spécifiques aux deux types de validation dans chacune d'entre-elles.

La première table se nomme « validation_auto », et stocke les résultats obtenus pour chaque donnée. Elle est reliée à la table « observations » (une table existante du SI Faune) dans laquelle sont stockées les données d'observations. Une donnée ne peut avoir qu'un et un seul niveau de validité automatique. Cette table est ensuite reliée à plusieurs tables de références listant les résultats possibles des critères et de la validation finale. Ces tables stockent pour chaque résultat possible le niveau de fiabilité (le libellé) et un descriptif complet expliquant la signification du résultat obtenu. Enfin, la table « validation_auto » stocke également la date à laquelle a été réalisé la validation. En effet, la validation automatique repose sur un état des connaissances à un moment donné, et elle peut être relancée à tout moment si nécessaire. La connaissance de la date de validation est donc primordiale.

Parallèlement à la validation automatique, la table « validation_manuelle » stocke le résultat de la validation manuelle (sous la même forme que la validation automatique), la date, le nom de l'expert validateur et un éventuel commentaire. Dans le cadre de la validation manuelle, les experts peuvent prendre contact avec le producteur de la donnée afin d'obtenir des informations supplémentaires. Le cas échéant, le nom de la personne consultée et la date du contact sont stockés dans la table de manière à éviter les sollicitations multiples du producteur. Contrairement à la validation automatique, une donnée peut avoir plusieurs niveaux de validation manuelle. Il est en effet possible que plusieurs experts valident la même donnée avec des avis différents. Dans ce cas là, chaque validation est conservée et la validation la plus récente est prise en compte lors de la diffusion de la donnée.

Le Modèle Conceptuel de Données (MCD) de cette extension est présenté ci-après (Figure 9).

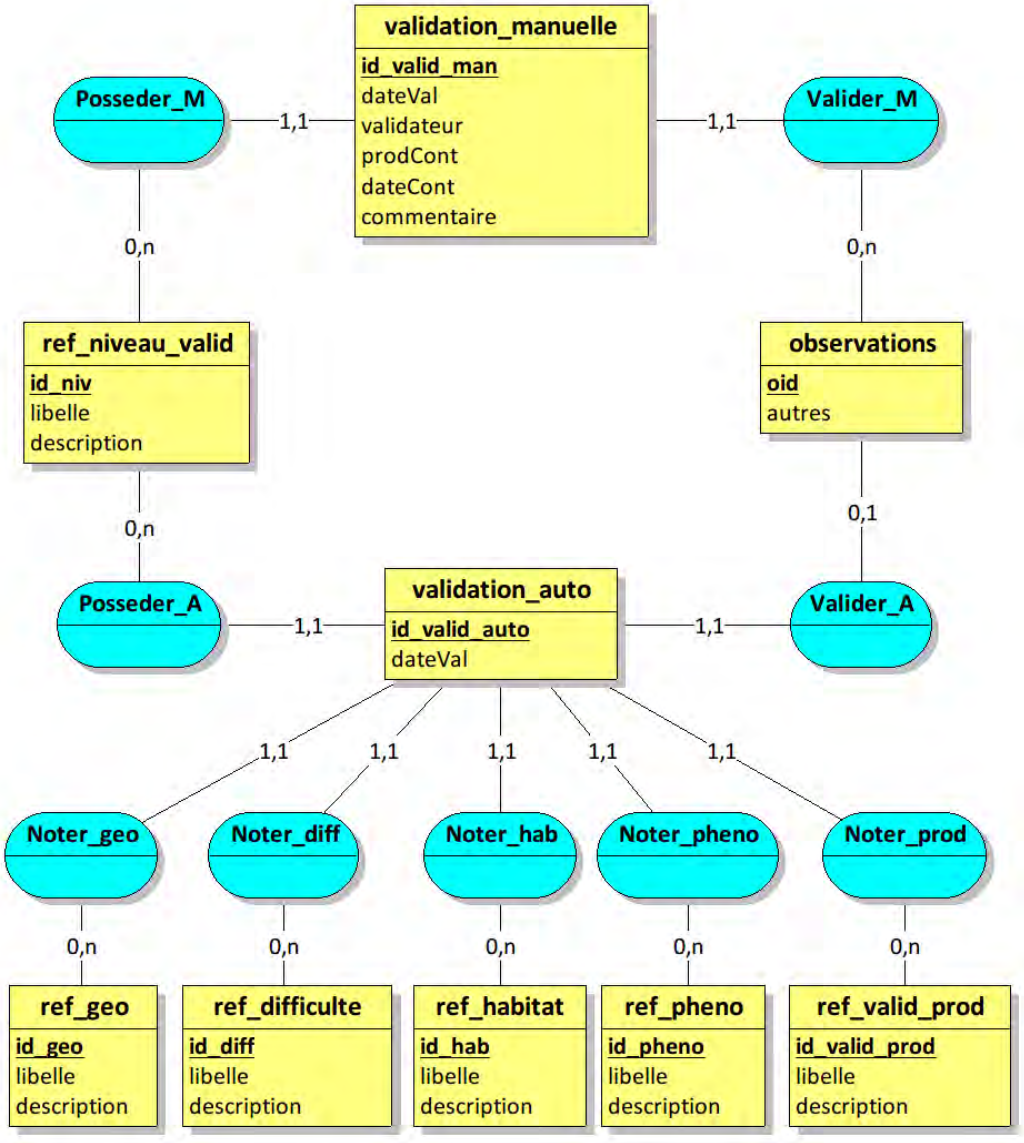

Figure 9 : MCD de l'extension validation à la base de données du SI Faune

5.2. Automatisation de la méthode dans une chaîne de traitement

5.2.1. Choix de la technologie utilisée

La première étape de la mise en place de la chaîne de traitement a été le choix de la technologie à utiliser. Pour cela, trois technologies ont été envisagées :

- FME : « Feature Manipulation Engine » est un ETL (*Extract Transform Load*) propriétaire spécialisé dans le traitement de données (données spatiales comprises). Il supporte une multitude de formats de données et permet grâce à une interface graphique et des *Transformers*[3] de manipuler avec complexité toutes sortes de données.
- SQL et PL/pgSQL : Le « Structured Query Language » est un langage normalisé permettant la gestion d'une base de données relationnelle. Associé au PL/pgSQL, un langage de procédure pour PostgreSQL, le SQL permet d'effectuer, via des fonctions et des triggers, des opérations plus complexes.
- Langage de programmation (Python) : Un langage de programmation permet la création d'algorithmes, qui décrivent un ensemble d'opérations réalisées dans un but précis. Dans notre cas, la procédure de validation pourrait par exemple faire l'objet d'un algorithme sous Python.

Chacune des trois technologies présente des avantages et des inconvénients, lesquels ont été résumés dans le tableau ci-après (Tableau 2).

Technologie	Avantages	Inconvénients
FME	- Développement facilité par une interface graphique et les *Transformers* - Maintenance et évolution facilitée - Logiciel déjà utilisé en interne	- Nécessite une licence payante (l'OAFS en possède une)
SQL / PGplSQL	- Gratuit - Traitement réalisé directement à l'insertion ou la modification d'une donnée (*trigger : ON INSERT / ON UPDATE*)	- Traitement réalisé par la BDD ralentissant ainsi ses performance pour les autres utilisations - Maintenance et évolution complexe
Python	- Gratuit - Gère le multi-processing	- Maintenance et évolution complexe

Tableau 2 : Avantages et inconvénients des technologies envisagées pour la réalisation de la chaîne de traitement

Au final, le choix s'est rapidement tourné vers FME. Il permet en effet de développer aisément la chaîne de traitement tout en facilitant sa maintenance et son évolution. Il est également déjà utilisé en interne par l'OAFS pour différentes missions et d'autres chaînes de traitement, assurant ainsi une reprise facilitée par le reste de l'équipe SI si nécessaire. Au contraire, traduire la procédure uniquement en SQL ou en Python complexifie à la fois le développement, la maintenance et l'évolution du processus de validation.

3 Les *Transformers* sont utilisés dans FME afin de manipuler les données. Ils permettent d'effectuer par simple glisser/déposer les opérations les plus communes.

5.2.2. Structure de la chaîne de traitement

Afin de développer la chaîne de traitement de la procédure de validation automatique, chaque critère a tout d'abord été étudié puis construit de manière indépendante afin de s'assurer de son fonctionnement et de sa correspondance avec les tests imaginés lors de la phase de conception.

Pour chaque critère, le script FME est structuré de la manière suivante :

1. Une requête SQL interroge la base de données et récupère les données sur lesquelles va se baser la validation. Si nécessaire, d'autre requêtes (ou des *Readers*[4]) vont récupérer les données de références stockées dans la base de données.
2. Les tests sur la donnée sont ensuite réalisés conformément à la méthode établie.
3. Le script attribue une valeur correspondant au résultat obtenu pour le critère.

La figure 10 illustre cette procédure pour le critère contrôlant l'habitat du taxon observé.

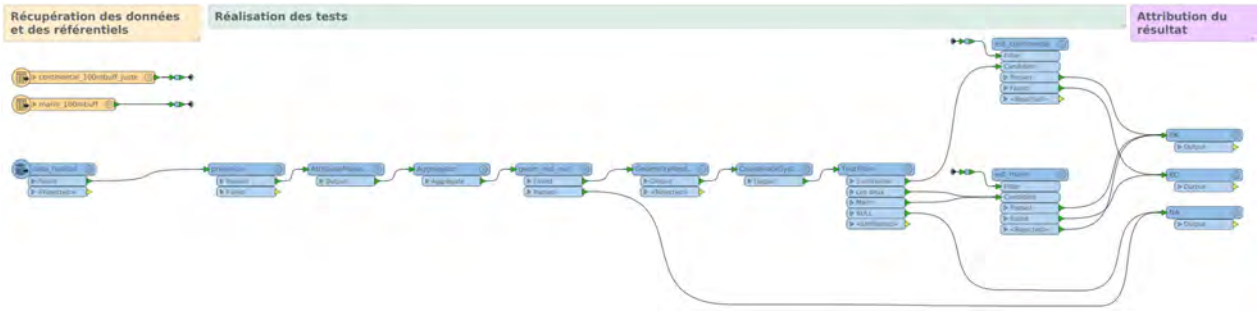

*Figure 10 : Exe*mple de structure d'automatisation avec le critère contrôlant l'habitat

Une fois tous les critères transcris sous FME, les résultats de chaque critères sont regroupés comme décrit durant la phase de conceptualisation, puis envoyés dans la base de données par l'intermédiaire d'un *Writer*[5]. Suite à cette étape, la chaîne de traitement est en capacité de réaliser totalement la validation sur un ensemble de données défini au préalable en tant que paramètre initial du script FME.

La validation doit pouvoir être lancée automatiquement sans avoir à modifier constamment les requêtes réalisées durant la procédure. Pour ce faire, un script « superviseur » a été mis en place. Son objectif est de détecter si de nouvelles données viennent d'être insérées dans la base de données et doivent donc être validées. En effet, il est possible grâce au *Transfomer* « *WorkspaceRunner* » de lancer un script FME à partir d'un autre script FME. Ainsi, le script « superviseur » teste la présence de données à valider, et s'il en trouve, lance le script de validation en lui précisant quelles données sont à valider (Figure 11).
Ce système a été mis en place sur les serveurs de l'OAFS (Linux) et le « superviseur », qui est un script demandant très peu de ressources, est lancé tous les jours (la fréquence d'exécution pourra être amenée à changer). Le lancement du script « superviseur » peut être fait à partir de FME Server[6] ou grâce au programme cron[7]. La licence FME Server coûtant environ 12 500 €, les autres option lui sont préférées. C'est donc cron qui lance le « superviseur » à heure fixe.

---

4 Dans FME, un *Reader* est une entité permettant la lecture fichier au format supporté, dont PostgreSQL.

5 Dans FME, un *Writer* est une entité permettant l'écriture de données dans format supporté, dont PostgreSQL.

6 FME Server est la version orientée serveur de FME. Il permet entre autre la planification de tâches.

7 *cron* est un programme permettant aux utilisateurs des systèmes Unix d'exécuter automatiquement des commandes à une date et une heure spécifiées à l'avance, ou selon un cycle défini à l'avance.

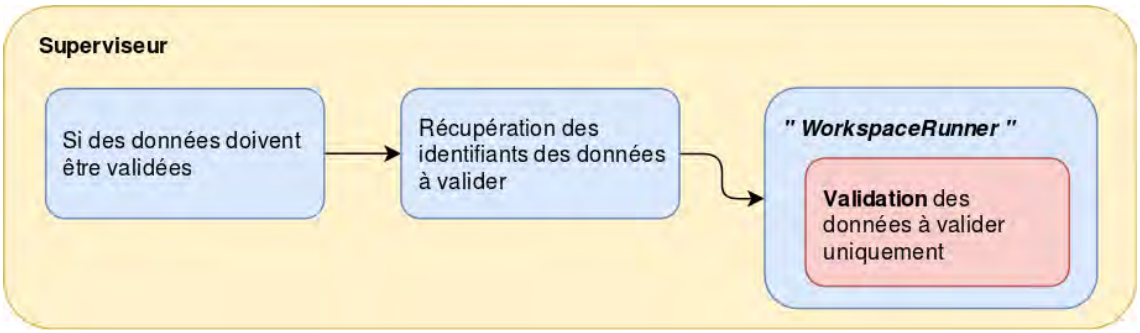

Figure 11 : Schéma du fonctionnement du script « superviseur ».
En jaune le script « superviseur », en bleu ses Transformers et en rouge le script de validation.

6. RÉSULTATS DES TESTS ET DISCUSSIONS

Afin de mettre à l'épreuve la chaîne de traitement développée au cours de ce stage, des tests ont été réalisés sur l'ensemble des observations de Mammifères continentaux non volants du SI Faune. C'est en effet l'unique groupe taxonomique pour lequel les référentiels des périodes d'observation et des difficultés de détermination ont pu être complétés en interne grâce à l'expertise de Thomas RUYS (Chargé de projets connaissances & expertises). De plus, la liste rouge des Mammifères d'Aquitaine étant en cours de réalisation par l'OAFS, l'ensemble des données disponibles en Aquitaine ont été mutualisées dans le SI Faune, constituant de fait un jeu de données complet et représentatif de la diversité du réseau d'acteurs, idéal pour l'analyse.

Au total, 181 117 données d'observations de Mammifères non volants en Aquitaine ont donc été soumises au processus de validation automatique. Parmi ces observations, 94,4 % obtiennent les niveaux « Très Probable » ou « Probable » ce qui est encourageant vis-à-vis de la qualité des données présentes dans le SI faune (Figure 12). Ces résultats confortent en partie les choix effectués lors de la phase de conception de la méthode : rejeter uniquement les valeurs les plus aberrantes.

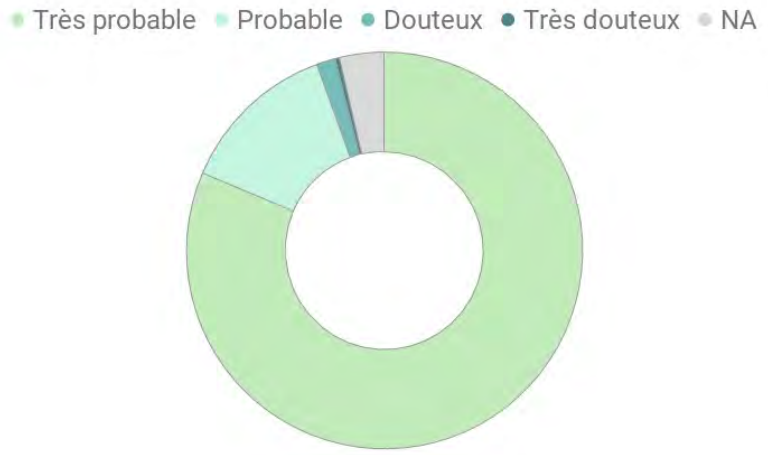

Figure 12 : Niveaux de validation automatiques obtenus sur les mammifères non volants
Très probable : 81,4 % ; Probable : 13 % ; Douteux : 1,6 % ; Très douteux : 0,4 % ; NA : 3,6 %

Afin de vérifier que la méthode identifie de manière efficace les données invalides, certaines données préalablement invalidées par Thomas RUYS (Chargé de projets connaissances & expertises) ont été comparées aux résultats obtenus. Les résultats pour chaque critère sont présentés ci-après.

Vérification du critère « répartition géographique » et « difficulté de détermination »

Une observation de Chat forestier (*Felis sylvestris*) avait été identifiée manuellement comme invalide au sein du SI Faune. Cette observation a en effet été réalisée en Gironde alors que cette espèce n'est présente que dans les Pyrénées Atlantique et en Dordogne. Suite la validation automatique, la donnée a obtenu le niveau de validation « Très Douteux », confirmant ainsi la capacité de la chaîne de traitement à identifier les erreurs liées à l'absence de l'espèce au département. En effet, comme décrit précédemment, l'absence d'une espèce au niveau départemental entraîne, si tous les autres critères sont bons, un niveau final « Douteux ». Étant donné que le Chat forestier est considéré comme difficile à déterminer, le niveau final est alors « Très Douteux ».

Autre exemple : celui d'une observation de Muscardin (*Muscardinus avellanarius*) dont la présence au niveau régional n'est pas certaine ou seulement accidentelle. Dans ce cas, le résultat de la validation automatique est censé être « Peu fiable » vis-à-vis du critère contrôlant la répartition géographique, et donc « Douteux » après regroupement si les autres critères contrôlés sont corrects. Suite à la validation, cette observation a en effet obtenu ce statut, puisque le Muscardin est facilement identifiable, justifiant donc le bon fonctionnement du processus automatique avec ces critères.

Globalement, les données de Mammifères sont majoritairement fiables au regard du critère contrôlant la répartition géographique : moins de 1 % des observations sont peu ou pas fiables (Figure 13). Concernant, la difficulté de détermination, on remarque qu'environ 11 % des données sont des observations d'espèces difficiles à identifier (Figure 14). Cette valeur explique en partie le nombre de données « Probable » au niveau global puisqu'une détermination difficile baisse le niveau de validité d'un cran.

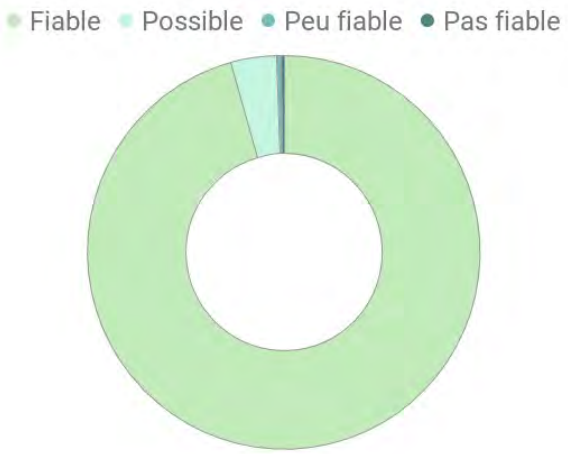

Figure 13 : Niveaux de fiabilité obtenus pour le critère « répartition géographique » sur les mammifères non volants
Fiable: 95,6 % ; Possible: 3,8 % ; Peu fiable: 0,4 % ; Pas fiable : 0,2 % ; NA : 0 %

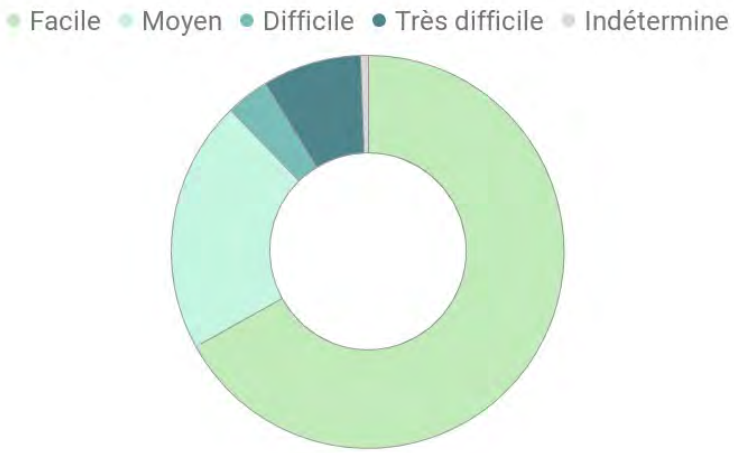

Figure 14 : Niveaux de difficulté de détermination des mammifères non volants
Facile: 67 % ; Moyen: 20,8 % ; Difficile: 3,5 % ; Très difficile : 8,1 % ; Indéterminé : 0,6 %

Vérification du critère « Habitat »

Une observation de Blaireau européen (*Meles meles*) a été identifiée comme « Très Douteuse » vis-à-vis du critère contrôlant l'habitat du taxon. Cette observation a en effet été effectuée au-delà (à environ 50m) du tampon de 100m effectué autour du domaine continental. Cette espèce effectuant évidemment la totalité de son cycle de vie sur le domaine terrestre, cette observation est donc légitimement invalide et le critère semble fonctionnel pour cette donnée.

De manière générale, les résultats obtenus pour le critère habitat (Figure 15) montrent une très grande majorité de données fiables : seulement 3 observations de mammifères sont effectuées en dehors de leur habitat de prédilection. Ces résultats sont bons mais soulèvent néanmoins l'intérêt que pourrait avoir l'amélioration de la méthode en distinguant des habitats plus précis afin d'identifier davantage de données invalides.

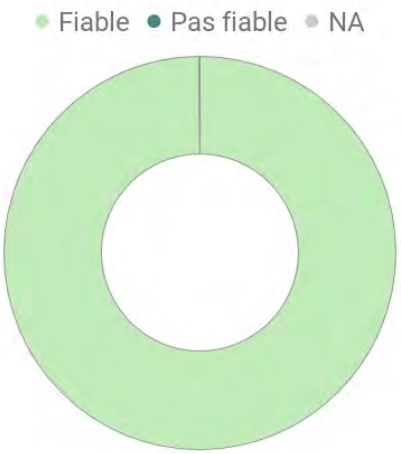

Figure 15 : Niveaux de fiabilité obtenus pour le critère « habitat » sur les mammifères non volants
Fiable: 99,9 % ; Pas fiable : 0 % ; NA : 0,1 %

Vérification du critère « Période d'observation »

Une donnée de Marmotte des Alpes (*Marmota marmota*) a été observée en février alors qu'elle hiberne durant cinq mois l'hiver, la rendant ainsi quasi-impossible à observer à cette période. Les résultats obtenus confirment ce constat, l'observation ayant obtenu le niveau « Pas Fiable » au critère contrôlant la période d'observation et donc « Très Douteuse » après le regroupement final.

Les autres données validées sont principalement « Fiables » et les dates correspondent avec le référentiel (Figure 16). Plus de 350 observations ont tout de même été identifiées comme « Possibles » (abaissant ainsi le niveau de validité final d'un cran) ou « Pas fiable » (en dehors de toutes les périodes d'observations plausibles).

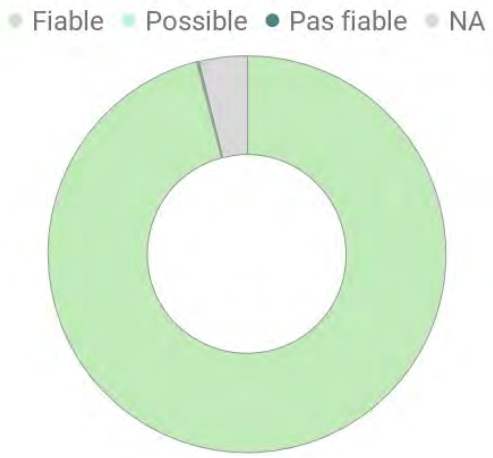

Figure 16 : Niveaux de fiabilité obtenus pour le critère «période d'observation» sur les mammifères non volants
Fiable: 95,9 % ; Possible: 0,1 % ; Pas fiable : 0,1 % ; NA : 3,9 %

Vérification du critère « Validation du producteur »

L'attribut renseignant la validation du producteur dans le SI Faune étant très peu renseigné, moins de 3 % des données disposent d'un résultat pour ce critère (Figure 17). Sur l'ensemble des données de Mammifères validées, une seule a été soulevée comme douteuse par le producteur. Dans ce cas, le processus de validation automatique a bien effectué le regroupement et le résultat de la validation est donc « Très Douteux ».

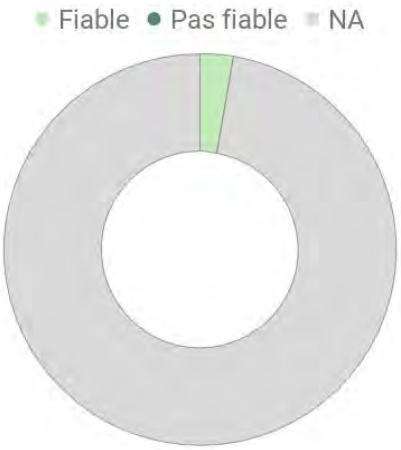

Figure 17 : Niveaux de fiabilité obtenus pour le critère « validation du producteur » sur les mammifères non volants
Fiable: 2,8 % ; Pas fiable : 0 % ; NA : 97,2 %

En conclusion, ces résultats montrent que la chaîne de traitement développée tout au long de ce stage fonctionne et permet l'identification des données invalides. Le processus n'est néanmoins pas sans faille et c'est pourquoi intervient ensuite la phase 2 de la validation combinée : assurer la validation des données pour lesquelles la validation automatique n'est pas suffisante.

BILAN PERSONNEL

Fort d'un premier diplôme dans la gestion de la biodiversité et le suivi environnemental (Master 2 Biodiversité et Suivis Environnementaux), j'ai toujours désiré associer à ces compétences en écologie mon affection pour la gestion des données, notamment géographiques, et l'informatique. Au cours de ce stage, j'ai dû mobiliser conjointement les compétences acquises au cours de la formation SIGMA et mes compétences en écologie. J'ai également pu mettre un pied dans le milieu naturaliste Aquitain. Ainsi, participer à la mise en place de cette méthode de validation des données naturalistes au sein de l'OAFS aura été pour moi un énorme bond en avant dans la réalisation de mon projet professionnel. Ce stage représentait donc une opportunité que je suis ravi d'avoir saisi.

Mon travail au cours de ce stage a englobé l'ensemble des phases de réalisation d'un projet, caractéristiques d'un véritable travail d'ingénieur. En effet, à mon arrivée, aucune démarche n'avait encore été engagée pour la mise en place du système de validation. J'ai donc dû commencer à la base avec la lecture des guides du SINP et la réalisation de l'état de l'art. Ont suivies ensuite les phases de conception et d'intégration, avec tous les choix que cela implique et toutes les difficultés qui peuvent être rencontrées (auxquelles il est alors nécessaire de trouver des solutions). Enfin, les tests effectués sur la méthode ont permis de rendre compte de son efficacité.

Ces six mois ont également été l'occasion d'améliorer beaucoup de mes compétences. Côté base de données par exemple, j'ai eu l'occasion de réaliser de multiples requêtes SQL, parfois très complexes, pour interroger et analyser les données. Côté traitement de données, lors de la production de la chaîne de traitement informatique, je me suis formé de manière approfondie à l'utilisation du logiciel FME. Par ailleurs, j'ai pu prendre en main l'environnement Linux, sur lequel j'ai travaillé durant la totalité du stage.

Ce stage aura aussi été très enrichissant humainement. Bien qu'ayant réalisé une grande partie de mon stage en autonomie, mes encadrants et toute l'équipe de l'OAFS ont toujours su se montrer à l'écoute de mes besoins lorsque cela était nécessaire. Ils ont été à la base d'un lien de confiance et de respect où chaque remarque était prise en compte. Il a donc été très agréable de travailler dans une structure à taille humaine (7 salariés actuellement), où la confiance et la communication sont les clés d'un travail d'équipe de qualité.

Pour conclure, ce stage, très complet et formateur, m'a permis :

- d'adopter une véritable démarche d'ingénierie, depuis la conception d'un projet jusqu'à sa réalisation technique, en passant par la recherche de solutions aux difficultés rencontrées ;
- d'associer dans un même projet mes affinités pour la gestion de données et la gestion de la biodiversité ;
- de prendre plaisir à travailler en autonomie au sein d'une équipe réduite qui sait être présente et à l'écoute quand cela est nécessaire ;
- d'améliorer grandement mes compétences techniques en SQL et avec FME.

Cette première expérience en tant que géomaticien se poursuivra en septembre avec un premier CDD de quatre mois afin de poursuivre la mise en place du système de validation et de participer aux nombreuses autres missions menées par l'OAFS. J'aurais notamment l'occasion de me former au développement Web sous Symfony afin de mettre en place l'outil de validation manuelle sur la plateforme web du SI Faune.

CONCLUSION ET PERSPECTIVES

La validité des données est aujourd'hui un enjeu majeur de la gestion et de la valorisation des données naturalistes, l'absence de prise en compte de cette dernière pouvant avoir de graves conséquences sur la crédibilité des démarches engagées en faveur de la biodiversité. Fort de ce constat, le SINP a rédigé un guide méthodologique contenant les principes et les recommandations d'usage pour la mise en place d'une méthode de validation de données. En tant que plateforme thématique régionale SINP, il était temps pour l'OAFS de se doter d'une telle méthode, et la création de ce stage avait donc pour objectif la conception et la mise en route d'un système de validation permettant l'attribution automatique d'un niveau de fiabilité aux données mutualisées par l'observatoire.

L'état de l'art des méthodes de validation existantes a permis, après de nombreuses concertations internes, l'identification de cinq critères retenus pour le calcul du niveau de validation d'une donnée d'observation naturaliste : l'habitat du taxon observé, sa répartition géographique, ses périodes d'observation, sa difficulté de détermination et le niveau de validité fourni par le producteur de la donnée. Ils reposent presque tous sur des données de références auxquelles sont comparées les données d'observations. Le choix de ces référentiels ont également fait l'objet de concertations qui ont abouti à l'utilisation du référentiel taxonomique national TAXREF, du référentiel espèces de l'OAFS, de certains atlas régionaux et à la construction de nouveaux référentiels : les référentiels des périodes d'observation et de difficulté de détermination. Les résultats obtenus lors du contrôle de chaque critère sont ensuite regroupés dans un arbre décisionnel dont l'objectif est d'attribuer le niveau de validité automatique final à la donnée naturaliste.

Ce processus de validation, encore à l'état de concept, a ensuite été mis en place avec le logiciel FME, permettant d'effectuer les calculs et d'enregistrer les résultats obtenus dans une extension de la base de données du SI Faune.

Ce stage à donc permis le développement d'un système de validation automatique fonctionnel qui est intégré dans le SI Faune afin de valider automatiquement les nouvelles données insérées dans la base de données via le script « Superviseur ».

La fin de ce stage est marquée par la communication d'une note méthodologique à destination du groupe d'experts pour relecture et validation. Le processus de validation automatique pourra donc potentiellement être amené à évoluer en fonction des remarques reçues. Ce même collectif a également la charge de réaliser les référentiels nécessaires aux contrôles de la période d'observation et de la difficulté de détermination.

Certains points pourront néanmoins faire l'objet de futures améliorations qui permettront à la méthode d'identifier avec davantage de précision les données invalides. C'est par exemple le cas du critère contrôlant l'habitat, qui ne permet actuellement que d'identifier les observations d'espèces marines effectuées sur le domaine continental et inversement. Une connaissance plus fine des habitats naturels (c'est notamment l'objectif du programme CarHAB initié par le ministère de l'écologie) permettrait d'identifier des données invalides sur d'autres types d'habitat plus précis. Par exemple, certaines espèces inféodées aux zones humides ne pourraient pas êtres observés en dehors de ces dites zones.

À moyen terme, une fois un nombre conséquent de données validées, la méthode pourra également prendre en compte des cartes de congruence générées à partir des données préalablement validées, qui permettront de compléter les référentiels utilisés pour le contrôle de l'aire de répartition des espèces.

Cette méthode, limitée à l'Aquitaine dans le cadre d'une première approche, sera normalement étendue en début d'année prochaine à la Nouvelle-Aquitaine avec la prise en compte des atlas régionaux du Limousin et du Poitou-Charentes et l'ajustement des référentiels des périodes d'observation et des difficultés de détermination.

Enfin, ce stage se poursuivra par la mise en place de la phase 2 du système de validation de l'OAFS, c'est à dire la validation manuelle des observations naturalistes dont les niveaux de validation automatique sont les plus douteux. Ce travail aura comme objectif de confirmer ou de modifier le niveau de validation attribué par le processus automatique. Pour cela, un outil web sera développé en interne pour permettre aux experts compétents de consulter les résultats obtenus dans le cadre de ce stage, de dialoguer entre eux, et de prioriser les données à valider.

BIBLIOGRAPHIE

ARPIN I., CHARVOLIN F., FORTIER A., 2015, Les inventaires naturalistes : des pratiques aux modes de gouvernement », Études rurales volume 195, p.11-26.

BRADLEY J. CARDINALE, J. EMMETT DUFFY, ANDREW GONZALEZ, DAVID U. HOOPER, CHARLES PERRINGS, PATRICK VENAIL, ANITA NARWANI, GEORGINA M. MACE, DAVID TILMAN, DAVID A. WARDLE, ANN P. KINZIG, GRETCHEN C. DAILY, MICHEL LOREAU, JAMES B. GRACE, ANNE LARIGAUDERIE, DIANE S. SRIVASTAVA, SHAHID NAEEM, 2012, Biodiversity loss and its impact on humanity, Nature volume 486, p.59–67

GARGOMINY, O., TERCERIE, S., RÉGNIER, C., RAMAGE, T., DUPONT, P., DASZKIEWICZ, P. & PONCET, L., 2018, TAXREF v12, référentiel taxonomique pour la France : méthodologie, mise en œuvre et diffusion. Muséum national d'Histoire naturelle, Paris. Rapport Patrinat 2018-117, 156 p.

IPBES, 2019, Communiqué de presse : Le dangereux déclin de la nature : Un taux d'extinction des espèces « sans précédent » et qui s'accélère. Consulté sur https://www.ipbes.net/news/Media-Release-Global-Assessment-Fr

JOMIER R., PONCET L., ROBERT S., MILON T., ARCHAMBEAU A-S., PAMERLON S., BOURGOIN T., DE MONICAULT L., BARNEIX M., FROMAGE P., CARPY B., CALLOU C., CANDELIER S., HUGUET A., MEUNIER D., WÓDKA-GOSSE A., GROSSIORD F., GUICHARD B., LANDRIEU G., LAFAGE B., 2018, Standard de données SINP Occurrences de taxons, version 2.0, UMS 2006 "Patrimoine naturel", Paris, 100 p.

MNHN, 2018, La grille Lambert 93, maille de 10 km par 10km, 1p.

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 MNHN-SPN 2015-44, 54 p.

ROBERT S., BARNEIX M., BODY G., CASTANET J., CAZE G., CELLIER P., DESSE A., DE MAZIÈRES J., FROMAGE P., GOURVIL J., JOMIER R., JUSTE A., LANDRY P., LEBEAU Y., LECOQ M.E., LESCURE J., MARAGE D., MEYER D., PAMERLON S., PAPACOTSIA A., PONCET L., QUAINTENNE G., SALTRÉ A. & TOUROULT J., 2016, Guide méthodologique pour la conformité, la cohérence et la validation scientifique des données et des métadonnées du SINP – Volet 1 : occurrences de taxons, Version 1. Rapport pour le SINP, rapport MNHN-SPN 2016-77, 63 p.

UMS PATRINAT, 2018, Habilitation des plateformes régionales du SINP - Thématique « Occurrences de taxon » - Principes de l'habilitation, Muséum national d'histoire naturelle, Paris. 6 p.