UNIVERSITÉ DE NANTES

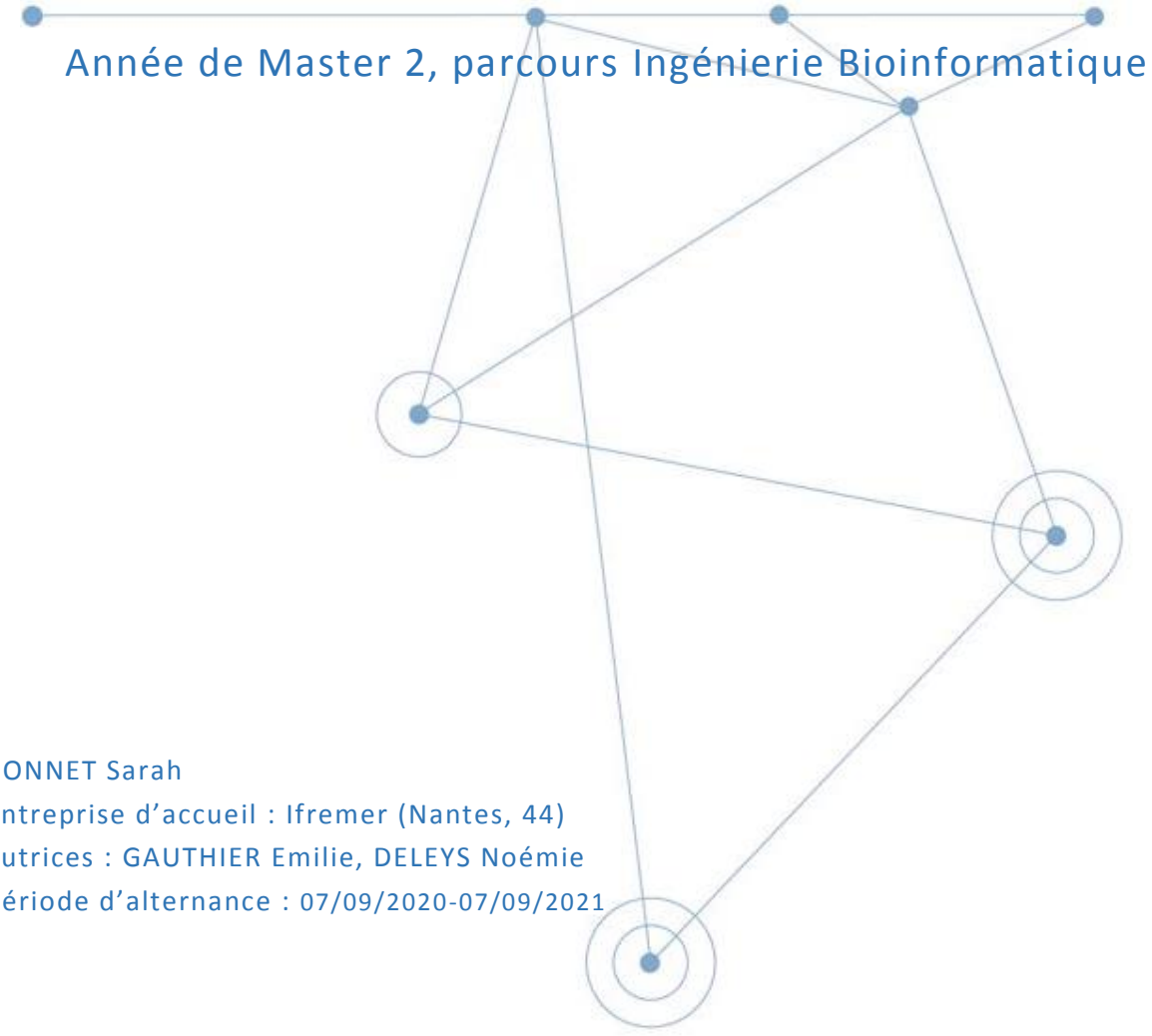
Ifremer

Master de Bioinformatique, Université de Nantes (44)
Promotion 2019-2021

Mémoire d'alternance

Qualification Biogéographique de données taxinomiques

Année de Master 2, parcours Ingénierie Bioinformatique



BONNET Sarah
Entreprise d'accueil : Ifremer (Nantes, 44)
Tutrices : GAUTHIER Emilie, DELEYS Noémie
Période d'alternance : 07/09/2020-07/09/2021

Ref. Ifremer/ODE/VIGIES/21-10
BONNET Sarah
27 août 2021 – Diffusion libre

Résumé

Le Système d'Information Quadrige, développé par l'Ifremer, représente aujourd'hui le Système d'Information national de référence pour la gestion des données de surveillance littorale effectuée conformément à la Directive Cadre sur l'Eau de l'Union Européenne. Ce Système d'Information, associé à une Base de données du même nom, gère donc ces données de surveillance du littoral, géolocalisées, parmi lesquelles comptent des données taxinomiques, dont chaque occurrence correspond à l'observation d'un taxon à un endroit donné, à un instant *t* et sur un **paramètre** donné.

Tout processus d'acquisition de données et de bancarisation pouvant présenter des erreurs, le Système d'Information Quadrige intègre différents outils et procédures de qualification des données. Il s'agit de détecter les anomalies pour les corriger et leur attribuer un niveau de qualité (« Bon », « Douteux », « Faux ») informant les usagers de la donnée de ses limites d'utilisation. Or aucune procédure de ce type n'existait pour le contrôle de la qualité des données taxinomiques, notamment pour les informations géographique et taxinomique. Pour pallier ce manque, les travaux que j'ai menés ont permis de mettre en place une procédure de contrôle biogéographique des données taxinomiques. Les tests pertinents ont été identifiés, puis des scripts ont été développés pour effectuer ces tests sur un jeu de données ciblé (phytoplancton du réseau REPHY des années 2010 à 2020). Les résultats ont permis d'élaborer les fiches descriptives permettant de rejouer les tests à la demande, certains tests ont été intégrés dans des outils déjà en exploitation, et l'ensemble a été ordonné et documenté, constituant la procédure de qualification attendue.

Mots-clés : Qualité des données, procédure, données taxinomiques, base de données, Quadrige

Abstract

The Quadrige Information System, developed by Ifremer, is currently the national reference information system for the management of coastal monitoring data carried out in accordance with the European Union's Water Framework Directive.

This Information System, is associated with a Database of the same name, and manages these geolocalized coastal monitoring data, including taxonomic data. Each occurrence corresponds to the observation of a taxon at a given location, at a given time and on a given parameter.

Since any data acquisition and banking process may contain errors, the Quadrige Information System integrates various tools and procedures for data qualification. The aim is to detect anomalies in order to correct them and assign a quality level ("Good", "Doubtful", "False") providing users with a confidence level thus a relevance in using those data.

However, no such procedure existed for the quality control of taxonomic data, especially for geographic and taxonomic information.

In order to overcome this lack, the work I carried out led to a procedure for biogeographical control of taxonomic data.

Relevant tests have been identified, and scripts were developed to perform these tests on a selected dataset (phytoplankton from the REPHY network from 2010 to 2020).

Results have been used to develop descriptive sheets allowing to run the tests on demand. Some tests have been integrated into existing tools and the result has been ordered and documented to provide the expected qualification procedure.

Keywords : data quality, procedure, taxonomic data, database, Quadrige

Remerciements

Dans un premier temps, je tiens à remercier tout particulièrement mes deux cotutrices, Emilie GAUTHIER et Noémie DELEYS, qui ont tenu une place centrale dans le bon déroulement de mon projet, pour leur appui thématique et technique précieux lors du processus de réflexion de la procédure et lors de la rédaction de ce rapport, pour tout le temps qu'elles m'ont consacré, mais aussi pour l'inspiration et la motivation qu'elles m'ont apportées lors de nos échanges.

Merci à toutes les personnes du service VIGIES pour m'avoir accueillie aussi chaleureusement, malgré le contexte particulier, et de m'avoir à l'occasion donné un aperçu de leur domaine respectif.

Merci à la coordination REPHY, assurée par Maud LEMOINE et Nadine MASSON, pour avoir pris le temps de me renseigner et d'échanger sur ce réseau surveillance, dont sont issues les données sur lesquelles j'ai travaillées pour l'établissement des tests de la procédure.

Merci aussi à Nicolas CHOMERAT et Kenneth MERTENS de l'Ifremer, qui ont accepté mes sollicitations pour une réunion avec la coordination REPHY et mes cotutrices, afin de me donner leur point de vue thématique sur l'avancement de mes travaux, ce qui m'a permis de réajuster mes recherches, et a inspiré certains tests de la procédure.

Merci à Steven PIEL de l'OFB, pour ses renseignements sur les différents référentiels géographiques nationaux et internationaux intéressants dans le cadre de mes travaux, et qui ont permis la mise en place des tests cartographiques retenus dans la procédure.

Je remercie mes responsables de formation pour l'accompagnement et les conseils qu'ils m'ont dispensés tout au long de cette année.

Merci enfin l'Ifremer de m'avoir accueillie dans le cadre de mon alternance, pour cette dernière année d'études enrichissante m'ayant permis de mettre un premier pas dans le monde professionnel.

Table des matières

Table des figures

Table des tableaux

Abréviations

Abréviation	Définition
BD	Base de Données
DCE	Directive Cadre sur l'Eau
DCSMM	Directive Cadre Stratégie pour le Milieu Marin
Ifremer	Institut Français de Recherche pour l'Exploitation de la Mer
INPN	Inventaire National du Patrimoine Naturel
MNHN	Muséum National d'Histoire Naturelle
Département ODE	Département Océanographie et Dynamique des Ecosystèmes
OFB	Office Français de la Biodiversité
ON	Open Nomenclature. Nomenclature taxonomique recommandée pour les identifications de taxons sur images.
REPHY	REseau de surveillance de PHYtoplancton et de phycotoxines
SANDRE	Service d'Administration Nationale des Données et Référentiels sur l'Eau
SAR	Service d'Administration des Référentiels marins
SI	Système d'information
SIE	Système d'Information sur l'Eau
SINP	Système d'Information sur la Nature et les Paysages
Projet SIVS	Projet Système d'Informations et de Valorisation de la Surveillance
VIGIES	Valorisation de l'Information pour la Gestion Intégrée Et la Surveillance
WoRMS	World Register of Marine Species

1. Contexte du projet d'alternance

1.1 Situation en entreprise

Mon contrat d'alternance s'est déroulé au sein de l'Institut Français de Recherche pour l'Exploitation de la Mer (Ifremer). Fondé en 1984, cet institut est considéré comme une référence nationale concernant la connaissance du milieu marin et de ses ressources. Il participe notamment à la surveillance du milieu marin et du littoral et au développement durable des activités maritimes (Ifremer, 2021). J'ai travaillé au sein du service Valorisation de l'Information pour la Gestion Intégrée Et la Surveillance (VIGIES) (Figure 1), lui-même situé dans le département Océanographie et Dynamique des Ecosystèmes (ODE). Ce service assure un soutien opérationnel et méthodologique aux unités en charge de l'observation et de la surveillance du littoral. Il s'y décline plusieurs métiers travaillant en synergie : la maîtrise d'ouvrage du Système d'information (SI) Quadrige et la valorisation des données, ainsi qu'une part de coordination de la surveillance. Cette coordination intervient autour des thématiques du phytoplancton et des phycotoxines, de la Directive Cadre Eau (DCE) et de la Directive Cadre Stratégie pour le Milieu Marin (DCSMM).

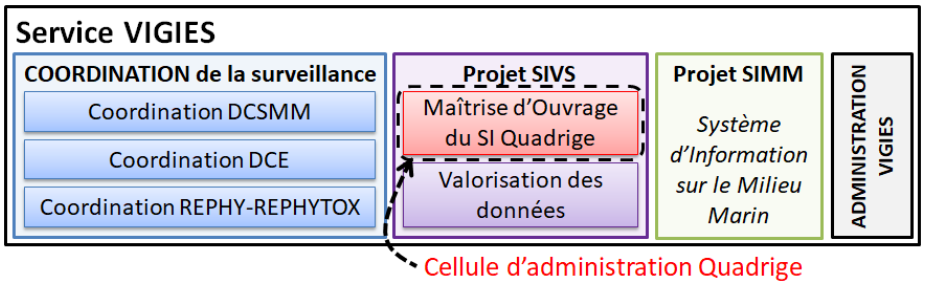


Figure 1 : Organisation du service VIGIES (adapté de Ifremer, 2017a).

Pour mener à bien ces missions, face à la nécessité de gérer de façon pérenne les données acquises et garantir leur accessibilité, le SI Quadrige a été créé autour de la Base de Données (BD) du même nom. Le SI Quadrige propose « des interfaces de saisie, plusieurs bases de données et une panoplie d'outils d'interprétation et d'élaboration de produits de valorisation » (Ifremer, 2020). Il représente aujourd'hui le « système d'information de référence pour la gestion des données de surveillance littorale effectuée dans le cadre de la DCE » (Ifremer, 2020). Au-delà du cadre réglementaire des politiques de surveillance, le SI Quadrige gère aussi d'autres données scientifiques, toujours en lien avec le milieu marin.

Les travaux que j'ai menés sur la qualification biogéographique des données taxinomiques (*cf.* § 1.3.2) du SI Quadrige s'intègrent plus précisément aux missions de la maîtrise d'ouvrage du SI Quadrige, assurées par la Cellule d'administration Quadrige (désignée par **Cellule Quadrige** dans la suite du présent document) (Figure 1). La Cellule Quadrige a été créée en 2008, afin de maintenir en conditions opérationnelles le SI Quadrige, de le développer et d'assurer la bonne bancarisation des données. Cette équipe est en charge de l'assistance aux utilisateurs du SI, de la gestion des formations, du support à l'intégration de données, du recueil et du suivi des évolutions. Elle assure également la gestion des référentiels du SI, et le support à la *qualification* (*cf.* § 1.3.1.1) et au contrôle de la qualité des données (Ifremer, 2017b).

1.2 Organisation des données dans le SI Quadrige

Les travaux que j'ai menés portent sur les données taxinomiques de la BD Quadrige. Dans le SI Quadrige, une **donnée taxinomique** se définit comme un résultat d'observation d'un taxon à un endroit donné, à un instant *t* et sur un **paramètre**[1] donné. Un **taxon** désigne quant à lui une « unité quelconque (famille, genre, espèce, *etc.*) de la classification zoologique ou botanique » (INPN ,2021). Afin de permettre la compréhension de ce document, il convient d'expliquer les éléments structurant les données de cette BD. Sont présentées ci-après des notions clés réinvesties dans la suite du rapport (en gras et en italique).

[1] Un paramètre est une propriété du milieu ou d'un élément du milieu qui contribue à en apprécier les caractéristiques et/ou la qualité et/ou l'aptitude à des usages (Cellule Quadrige, 2009)

Dans le SI Quadrige, les données sont organisées selon (Figure 2) :

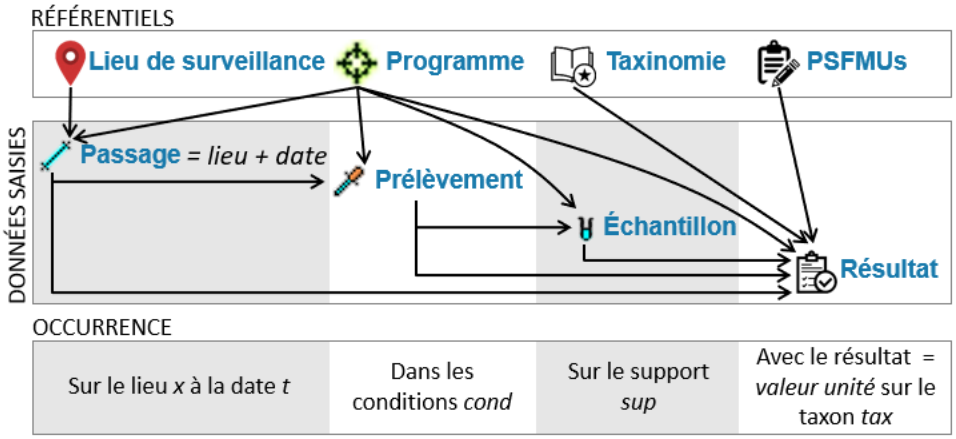


Figure 2 : Organisation des données dans le SI Quadrige.

- <u>des référentiels :</u> listes communes à tous les utilisateurs, définies et gérées au niveau national et sur lesquelles s'appuient les données d'observation et de mesure. Parmi ces référentiels, citons :
  - les ***lieux de surveillance*** : entités géographiques de référence sur lesquelles sont effectués les suivis. Différentes informations, telles que des coordonnées de centroïde, un libellé et une bathymétrie sont associées à un lieu de surveillance. Un lieu de surveillance est localisé de façon unique par son emprise géographiques qui peut être ponctuelle, linéaire ou surfacique ;
  - les ***programmes*** : activités à l'origine de la collecte d'un ensemble cohérent de données, que ce soit pour les réseaux de surveillance ou pour des études limitées dans le temps. Ces activités sont toujours mises en œuvre selon une procédure déterminée à l'avance (Cellule Quadrige, 2009) ;
  - les *taxons* : le référentiel taxinomique Quadrige est propre au SI Quadrige. Il possède ses propres règles d'administration et un périmètre limité aux données bancarisées dans le SI. Chaque taxon du référentiel est décrit notamment par un libellé, un auteur, un niveau taxinomique, un taxon parent, et possède un identifiant unique nommé *TAXON_NAME_ID*. Il contient des équivalences avec des référentiels externes qui font foi pour ses mises à jour : le WoRMS (*cf.* § 2.3.2.1), TAXREF (*cf.* § 2.3.2.2) et le SANDRE[2] ;
  - les ***PSFMUs*** : combinaison de 5 éléments définissant tout résultat de la BD Quadrige : *Paramètre (P) = caractéristique mesurée ; Support (S) = support/élément analysé e.g. de l'eau ou un bivalve ; Fraction (F) = partie du support analysée ; Méthode (M) = méthode utilisée pour mesurer le paramètre ; Unité (U) = unité de mesure du résultat.*
- <u>des données d'observation et de mesure</u> sont saisies par les producteurs de données et s'appuient sur les référentiels. Les données sont organisées selon différents niveaux (Figure 2) liés par une relation d'héritage :
  - le ***passage*** : il se définit par un lieu auquel il est rattaché, une date et un ou plusieurs programmes dans lesquels il s'inscrit. Il peut comporter d'autres informations précisant les conditions d'observation : *e.g.* sonde (hauteur d'eau sous le bateau), coordonnées reprécisées ;
  - le ***prélèvement*** : il est rattaché à un passage et réunit les informations concernant la mise en œuvre d'une opération de prélèvement → l'engin utilisé, la taille du prélèvement l'organisme préleveur impliqué, le ou les programmes auxquels il est rattaché. Des informations d'intérêt sur les conditions du prélèvement (coordonnées réelles du prélèvement ou encore profondeur à laquelle est effectué le prélèvement) peuvent être ajoutées ;
  - l'***échantillon*** : subdivision du prélèvement, il porte les informations propres à celle-ci → le support de l'échantillon (partie qui est recueillie pour analyse ou dénombrement), le ou les programmes concernés, et des informations complémentaires, telles que le taxon support de l'échantillon s'il y en a un à préciser ou encore la taille de l'échantillon ;
  - le *résultat* : rattaché à un (et un seul) des niveaux d'information précédents, il est caractérisé par la valeur (qualitative ou quantitative) ou le fichier résultat, le PSFMU, le taxon ou groupe de taxons mesuré s'il y en a un, l'organisme analyste ayant établi le résultat et le ou les programmes

[2] SANDRE : Service d'Administration Nationale des Données et Référentiels sur l'Eau, à l'origine du référentiel des données sur l'eau du Système d'Information sur l'Eau (SIE), visant à établir une uniformité des données sur l'eau du secteur public à l'échelle nationale afin d'en faciliter les échanges et la diffusion.

dans lesquels s'inscrivent le résultat.

Chaque niveau hérite des informations de l'élément auquel il se rattache. Dans le cas des coordonnées, le passage a par défaut la même emprise que le lieu de surveillance auquel il est rattaché : il aura des **coordonnées** dites **héritées** du lieu de surveillance. Il arrive que l'observation soit effectuée à un endroit légèrement différent du lieu théorique (dérive du bateau, conditions de marée, …) : dans ce cas des coordonnées relevées sur le terrain peuvent être saisies au niveau du passage. Le passage a alors des **coordonnées** dites **réelles**.

Toutes ces données sont saisies par les utilisateurs de différentes manières. L'intégration de données peut être effectuée à partir d'applications développées par Ifremer, telles que Quadrige^2 (Q^2), BD Récif et DALI. Ces applications contiennent des contrôles à la saisie limitant le risque d'erreur. Toutefois, il est aussi possible d'importer des données *via* de dépôt en ligne de fichiers respectant des formats standardisés. L'import de ces fichiers est également soumis à des contrôles de cohérence, notamment vis-à-vis des référentiels, mais certains de ces contrôles sont différents des applications.

Durant l'acquisition, la saisie ou encore la gestion des données, des erreurs et des incohérences peuvent survenir. Néanmoins les données bancarisées dans la BD Quadrige ont pour vocation d'être utilisées pour définir la qualité des eaux dans le cadre des réseaux de surveillance ou encore pour répondre aux directives européennes. Les données bancarisées doivent donc être de qualité. Pour cela, elles doivent être contrôlées et qualifiées.

1.3 La qualification des données dans le SI Quadrige

1.3.1 Présentation de la qualification des données dans le SI Quadrige

1.3.1.1 Qu'est-ce que la qualification des données ?

Les erreurs ou incohérences des données peuvent nuire à leur interprétation lors de leur traitement et valorisation. Ainsi quand une donnée est considérée comme douteuse ou fausse, il est important de justifier et d'argumenter cette analyse afin d'informer tout utilisateur des limites d'utilisation de la donnée. Dans le cas contraire, si la donnée est estimée de bonne qualité, il convient aussi de le signaler, afin de prévenir les utilisateurs que cette donnée est analysable et éviter sa ré-expertise.

Dans le SI Quadrige, la **qualification** des données et des métadonnées, pour laquelle la Cellule Quadrige offre un appui technique, consiste en l'attribution d'un niveau de qualité aux données et métadonnées. Cette qualification est datée. Les différents niveaux de qualification possibles dans la BD Quadrige sont : « *Non qualifié* » pour les éléments n'ayant pas encore été qualifiés, « *Bon* » pour les éléments estimés comme cohérents et dont la considération pour d'éventuelles analyses est pertinente, « *Douteux* » pour les éléments suspectés d'une erreur pouvant biaiser les analyses, et « *Faux* » pour les éléments aberrants ou comportant une erreur identifiée, à ne pas inclure dans les analyses (Cellule Quadrige, 2019). Pour les niveaux « douteux » et « faux », le niveau de qualification est accompagné d'un commentaire justifiant le niveau attribué.

1.3.1.2 La qualification, partie intégrante du cycle de vie standard de la donnée

Les données bancarisées dans la BD Quadrige suivent un cycle de vie standard (Figure 3). Elles sont issues de mesures ou d'observations effectuées *in situ* ou d'analyses effectuées en laboratoire par un organisme analyste. Elles sont ensuite intégrées dans la BD Quadrige par saisie manuelle ou par le biais d'un import de fichiers de données. Une fois bancarisées, l'organisme saisisseur de la donnée vérifie la cohérence des informations bancarisées : c'est l'étape du contrôle. S'il estime la saisie correcte, il valide les données. Dès lors, les données sont rendues accessibles par tous. La validation implique que le producteur de données ne peut plus modifier ses données sans passer par un agent habilité pour dévalider les données en question. Dans le SI Quadrige, la qualification intervient après le contrôle et la validation des données, c'est-à-dire une fois que les données sont définitives et non modifiables par un tiers. Une dévalidation de données entraîne la suppression immédiate de la qualification de celles-ci.

1.3.1.3 Les acteurs de la qualification

La qualification des données est réalisée par différents acteurs (Figure 3) (Cellule Quadrige, 2019) :

- les producteurs de données (préleveur, analyste) et saisisseurs peuvent être à l'origine de la qualification en signalant une anomalie dans processus d'acquisition de la donnée ou être sollicités *a posteriori* pour corriger et qualifier leurs données (notamment en vérifiant les informations d'après les cahiers de paillasse et de terrain) ;
- des experts thématiques qui vont définir des règles de qualification en lien avec leur domaine d'expertise, analyser les données qui ne respectent pas ces règles et attribuer les niveaux de qualité aux données en les commentant si besoin ;
- la Cellule Quadrige, qui va aider à mettre en place et à utiliser les outils de qualification (applications, programmes informatiques).

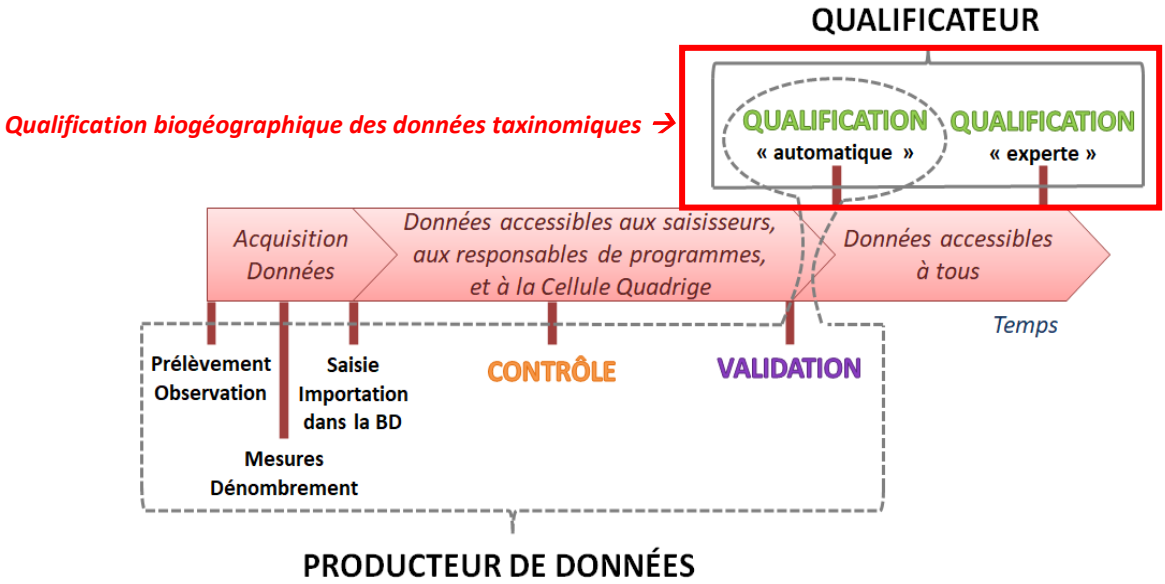


Figure 3 : Cycle de vie standard des données dans le SI Quadrige, et acteurs de la qualification.

1.3.1.4 Différents modes de qualification

Il existe différents modes de qualification dans le SI Quadrige. Les deux processus présentés ci-dessous sont liés à la thématique d'acquisition de la donnée (Figure 3) et s'inscrivent dans la procédure de qualification biogéographique des données taxinomiques élaborée à l'occasion de ce projet d'alternance (Cellule Quadrige, 2019) :

- la qualification dite « automatique » permet de vérifier la cohérence entre les informations composant une donnée d'après des règles définies par un expert de la thématique concernée. Ces contrôles sont effectués à l'aide de programmes élaborés par la Cellule Quadrige, et permettent de déceler automatiquement des anomalies qu'il convient d'analyser (*e.g.* des résultats d'analyse hydrologique[3] renseignant une température d'eau de mer à 31°C est suspecte et doit être vérifiée) ;
- la qualification dite « experte » permet de vérifier la cohérence d'un résultat par rapport aux autres données constituant le jeu de données et replacé dans son contexte spatio-temporel. En effet, des données extrêmes peuvent être détectées grâce à des analyses statistiques effectuées à l'aide de programmes informatiques puis présentées à un expert pour qualification. *E.g.* : une température de l'eau de mer = 12°C est cohérente prise isolément, mais elle peut être exceptionnellement basse ou élevée par rapport aux valeurs habituellement rencontrées sur ce lieu à cette saison.

Par ailleurs, un outil de qualification, « AlerteAno », a été mis en place par la Cellule Quadrige pour déceler des anomalies techniques dans les données de l'ensemble de la base (quelle que soit la thématique), notamment des incohérences structurelles provenant de bugs applicatifs en attente de correction (*e.g.* donnée de résultat sans aucune valeur, ni numérique ni qualitative, qui aurait dû être

[3] Hydrologie : analyse physico-chimique de l'eau, notamment de la température, la salinité, l'oxygène dissous ou encore les sels nutritifs.

supprimé à l'enregistrement par l'application Q[2]). Cet outil est un programme Talend exécutant quotidiennement des requêtes SQL identifiant des anomalies et produisant des rapports d'anomalie et des listings de données concernées. Si la correction est univoque, des requêtes SQL de correction sont exécutées automatiquement. Le cas échéant, la Cellule Quadrige analyse les anomalies et contacte, si nécessaire, des producteurs de données pour les corriger.

1.3.2 Emergence du besoin de qualification biogéographique des données taxinomiques

A l'occasion d'analyses et de contrôles qualité menés sur des jeux de données bancarisés dans la BD Quadrige, des erreurs ont été détectées et ont été signalées à la Cellule Quadrige. Ces erreurs étaient de plusieurs ordres, *e.g.* le taxon renseigné dans la BD Quadrige est incohérent avec la thématique (*e.g. requin dans des données de phytoplancton*), la localisation de l'observation du taxon renseignée en base est incohérente avec son aire de répartition (*e.g. taxon indo-pacifique dans l'océan Atlantique*), ou encore les coordonnées renseignées sont incohérentes avec la thématique (*données marines se situant en milieu terrestre, e.g. Figure 4)*.

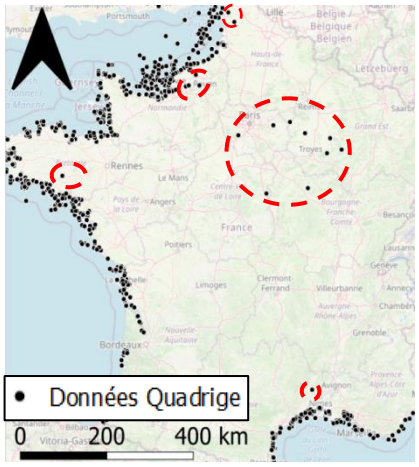


Figure 4 : Localisation de données marines du SI Quadrige en France métropolitaine : *des données à terre, considérées comme douteuses, sont entourées.*

Jusqu'à présent la détection de telles erreurs se faisait de manière ponctuelle et opportuniste. Ainsi, de telles anomalies n'étaient pas identifiables de façon systématique et globale à l'échelle de toute la BD, alors que cela serait intéressant à mettre en place pour améliorer la qualité des données de la BD Quadrige.

La Cellule Quadrige a déjà mis en place des procédures de qualification sur certains domaines thématiques comme l'hydrologie. En revanche aucune procédure n'a été mise en place pour les données taxinomiques. Les travaux que j'ai menés ont eu pour objectif de préparer la mise en place d'une telle procédure.

1.3.3 Caractérisation de la procédure de qualification

Les travaux que j'ai menés ont consisté à élaborer une procédure de **qualification biogéographique des données taxinomiques**, intégrable au système de qualification du SI Quadrige. Afin d'assurer le maintien, l'application régulière et le déploiement à grande échelle, cette procédure doit se présenter sous la forme de tests de qualité :

- reproductibles ;
- respectant une architecture et des règles communes aux tests de qualité déjà en place ;
- reposant sur des outils informatiques qui permettent l'analyse de jeux de données complets ;
- produisant des rapports d'anomalie adaptés aux processus de qualification automatique et experte de Quadrige.

Ainsi, mon travail a consisté à :

- dresser un état de l'art des connaissances et pratiques actuelles sur la qualification, en particulier biogéographique (publications, outils existants, démarches de qualification mises en place dans le SI Quadrige et d'autres structures) ;
- identifier les éléments de la donnée pouvant impacter la qualité de la composante biogéographique, sur lesquels baser les tests ;
- déterminer les référentiels à utiliser pour les tests ;
- mettre en place et ordonner des tests, ainsi que présenter des préconisations de tests supplémentaires non implémentés pour l'instant;
- établir des fiches techniques standards pour l'application de chacun des tests ;
- intégrer les outils, scripts et la documentation de la procédure dans l'architecture réseau mise en place dans le service, afin d'en faciliter la prise en main par les agents du service.

Certains des tests de cette procédure ont pu être éprouvés et les résultats sont présentés dans le présent rapport, tandis que d'autres sont à l'état de préconisation, et présentés dans le § 5 Perspectives.

2. Pré-requis à l'établissement de la procédure de qualification

2.1 Des données interprétables, selon des formats standards

La nécessité de standards de données revient dans la plupart des sources abordant la démarche qualité sur des données (*e.g.* Chapman et al., 2020). Avant toute chose, pour effectuer des tests, il est nécessaire de s'assurer que toutes les données respectent un format standardisé. Cela permet d'avoir des données homogènes, de les rendre interprétables informatiquement, et de limiter la manipulation des jeux de données avant traitement.

Le SI Quadrige offre déjà une partie de ces garanties *via* :

- l'architecture de la BD (*e.g.* respect du type de variable),
- des contrôles applicatifs[4] : saisie limitée à des listes *(formulaires)* basées sur les référentiels communs, règles de contrôle communes (*e.g.* respect de l'intervalle [-180 ; 180] pour les valeurs de longitudes) et thématiques (*e.g.* l'heure de prélèvement des données hydrologiques est obligatoire).

Ainsi, assurer le respect de standards de données permet d'étudier à grande échelle les différentes données et métadonnées présentes dans la BD Quadrige, et sur lesquelles les tests devront se baser.

2.2 Identification des métadonnées d'intérêt

Chapman et al. (2020) ont mené une étude sur des données relatives à la biodiversité et intégrées dans un Système d'Information Géographique (SIG). Pour ces données, les métadonnées (*i.e.* « éléments d'information ») ont été articulées selon le vocabulaire du Darwin Core (TDWG, *s.d.*) et classées par cas d'utilisation (usages courants de ce type de données, *e.g.* études de la densité de population, ou de la répartition). Les éléments d'information les plus utilisés ont alors été considérés comme ceux dont l'amélioration de la qualité impacterait le plus la qualité générale de la donnée. Les 5 éléments ayant le plus grand niveau d'implication sont dans l'ordre d'importance : les coordonnées ; le nom scientifique du taxon identifié dans la donnée ; les valeurs décimales de longitudes puis de latitudes ; la date ; le pays.

Ainsi, les catégories de métadonnées impactant le plus la qualité de la donnée sont : la localisation, la taxinomie, et ensuite la temporalité.

Aussi, une erreur dans l'une de ces catégories de métadonnées pourrait invalider les résultats des tests sur la qualité biogéographique qui exploitent ces informations. Il semble ainsi important de mettre en place des tests sur ces métadonnées en amont des tests sur la biogéographie, selon une hiérarchie d'exécution cohérente et optimisant la recherche et la résolution des erreurs.

Pour vérifier la cohérence des métadonnées et données relatives à la taxinomie, la localisation, puis la biogéographie, il est nécessaire de s'appuyer sur des référentiels externes qui vont permettre de valider ou invalider les observations effectuées.

2.3 Identification des référentiels pertinents

2.3.1 Référentiels géographiques

Les données bancarisées dans le SI Quadrige sont acquises dans le milieu marin, du large aux eaux de transition, ce qui inclut une partie des estuaires et les lagunes. Afin de vérifier que la localisation des données est cohérente avec le périmètre d'action du SI Quadrige, trois référentiels complémentaires ont été sélectionnés :

- l'espace maritime français : référentiel mis à jour en mai 2020 par l'Office Français de la Biodiversité (OFB) s'étendant du trait de côte histolitt® (mars 2020) à la limite de la Zone Economique Exclusive (ZEE) (12 milles marins au large) ;
  ↳ Considéré comme pertinent car les données stockées dans la BD Quadrige proviennent en grande partie d'analyses effectuées dans les aux marines françaises, et devraient donc se situer dans l'espace maritime français. Cependant, ce référentiel n'inclut pas les eaux de transition. Or la BD

[4] Contrôle applicatif : règle imposée par une application, ici une application du SI Quadrige permettant la saisie de la donnée.

Quadrige contient des données y étant acquises. Pour que ces données ne soient pas identifiées comme anomalies potentielles à tort, il faut compléter les contrôles avec un autre référentiel qui inclut les eaux de transition ;

- les masses d'eau DCE françaises 2019 (février 2021), administrées par le SANDRE, servant de zonage d'évaluation de l'état des eaux dans le cadre de la DCE ;
  ↳ Considéré comme pertinent car couvre à la fois les eaux littorales et de transition françaises, ce qui correspond aux zones suivies dans une grande partie des réseaux de surveillance bancarisés dans la BD Quadrige. Cependant il ne couvre pas toutes les zones étudiées, *e.g. celles* situées au large, suivies dans le cadre de la DCSMM. Certains suivis étant situés en dehors des deux précédents zonages (*e.g.* à proximité de Jersey), il est nécessaire de rajouter un autre référentiel ;
- les Zones Marines Quadrige ont été élaborées spécifiquement pour le SI Quadrige en 2004. Ce zonage couvre l'ensemble des milieux suivis dans le cadre du SI Quadrige. Ses entités ont été découpées de sorte que chaque lieu soit associé à une et une seule Zone Marine ;
  ↳ Ce référentiel couvre des zones complémentaires aux deux zonages précédents. Cependant, certains lieux ajoutés récemment se situent hors des Zones Marines historiques.

La couverture spatiale proposée par la combinaison de ces trois référentiels est censée englober la totalité des milieux suivis dont les données sont bancarisées dans la BQ Quadrige. Des données situées hors de ces trois référentiels seraient donc à vérifier. La complémentarité de ces référentiels est illustrée sur la Figure 5.

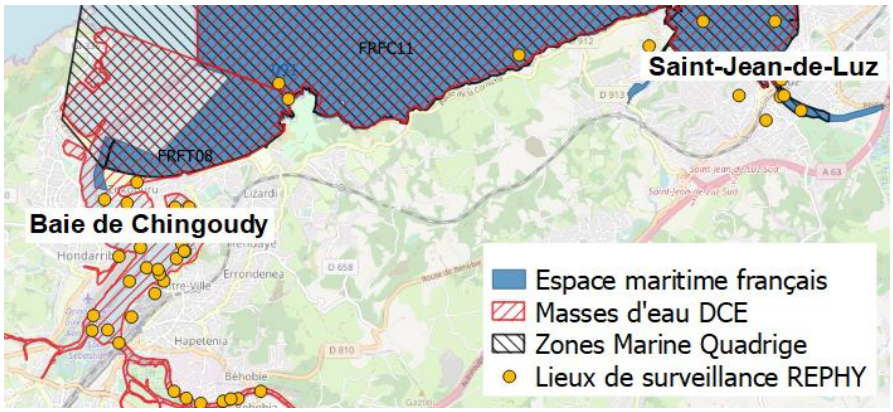


Figure 5 : Illustration de la complémentarité des référentiels.

2.3.2 Référentiels taxinomiques

La qualification biogéographique des données taxinomiques comporte deux niveaux de contrôles exploitant les métadonnées taxinomiques :

- la nomenclature taxinomique utilisée est-elle correcte ? (*i.e.* le taxon identifié dans la donnée correspond-il a une entité scientifiquement correcte, connue et identifiable de façon unique ?) ;
- l'observation de ce taxon à cet endroit-là est-elle probable ? (*i.e.* biogéographie).

Les référentiels considérés comme pertinents pour mener à bien ces contrôles sont décrits ci-après.

2.3.2.1 World Register of Marine Species (WoRMS)

Le WoRMS est un référentiel mondial des espèces marines, dont le but est de fournir de façon la plus exhaustive possible la liste des taxons marins sous tous leurs noms connus. Son contenu est défini selon la libre contribution d'experts en taxinomie. Le WoRMS est un des référentiels les plus consensuels et universels dans le domaine scientifique marin. Il se base sur la BD Aphia développée par le VLIZ[5]. Il assigne à chacun des taxons référencés un identifiant unique, l'***AphiaID***, qui est repris dans le SI Quadrige pour 99,8% des taxons de son référentiel taxinomique. Il comprend des espèces aquatiques, non seulement du milieu marin, mais aussi des eaux de transition : il couvre donc le périmètre thématique des données Quadrige.

Par ailleurs ce référentiel est facilement accessible par le biais de web services ce qui facilite la mise en place d'outils informatiques d'interopérabilité entre référentiels. Le WoRMS est donc le référentiel taxinomique qui a été retenu pour effectuer les mises à jour régulières du référentiel taxinomique Quadrige. Ces mises à jour sont effectuées soit manuellement, taxon par taxon, par la Cellule Quadrige au

[5] VLIZ : Vlaams Instituut voor de Zee, soit l'institut Marin de Flandre en Néerlandais

gré des demandes des utilisateurs, soit de façon globale pour l'ensemble du référentiel *via* une procédure semi-automatisée basée sur des scripts Talend.

Chaque *AphiaID* est notamment associé à un libellé, un auteur, une classification taxinomique *via* des liens père-fils permettant d'établir une généalogie sous forme d'arbre taxinomique et un statut couplé à un item « *validAphiaID* » permettant de déterminer si le taxon est référent ou non.

Par ailleurs, le WoRMS consigne des informations concernant les types de milieu de vie possibles de certains taxons. Ces informations sont utiles à la réalisation des tests de biogéographie. L'information relative aux types de milieu de vie est un booléen qui prend la valeur « NULL » quand le statut n'est pas défini, 0 si le taxon est absent du milieu et 1 si le taxon est présent dans le milieu :

- *isMarine* indique si le taxon peut être retrouvé en milieu aquatique marin;
- *isBrackish* indique si le taxon peut être retrouvé en eau saumâtre ;
- *isFreshwater* indique si le taxon peut être retrouvé en eau douce ;
- *isTerrestrial* indique si le taxon peut être retrouvé en milieu terrestre.

2.3.2.2 Référentiel taxinomique TAXREF

TAXREF est le référentiel taxinomique national pour la faune, la flore et la fonge française. La taxinomie étant en perpétuelle évolution, une nouvelle version de ce référentiel est publiée chaque année par l'Inventaire National du Patrimoine Naturel (INPN). Il est géré par le Muséum National d'Histoire Naturelle (MNHN) dans le cadre de la mise en œuvre du Système d'Information sur la Nature et les Paysages (SINP). Ses objectifs sont *(Gargominy et al., 2020)* :

- de donner un nom scientifique unique, non ambigu, et consensuel aux niveaux national et international, pour chacun des taxons de France ;
- de permettre une interopérabilité entre les différents jeux et bases de données à l'échelle nationale ;
- de gérer les évolutions taxonomiques et nomenclaturales dans les données concernant ces espèces (suivi et gestion de la synonymie et de la hiérarchie taxonomique).

Pour avoir un seul et même nom pour chaque taxon, TAXREF s'appuie en partie sur les *Global Species Database* (BD de référence mondiale, telles que le WoRMS) revues par des groupes d'experts internationaux, mais aussi sur les compétences en interne au MNHN. Aussi, certaines données sont issues de référentiels locaux, et intégrées dans le cadre de la consolidation nationale par un réseau national d'experts. Chacun des taxons du référentiel possède un identifiant unique : le ***CD_NOM***.

TAXREF constitue le support principal de mise à jour du référentiel taxonomique du SANDRE, dont les identifiants uniques (*taxon_ID*) doivent être mentionnés dans le SI Quadrige. En effet, certains éléments du référentiel du SANDRE sont réinvestis dans des protocoles d'observation, ou dans le calcul d'indicateurs d'évaluation de l'état des eaux, dont les données sont stockées dans la BD Quadrige.

Le référentiel Quadrige contient des correspondances entre les identifiants des taxons Quadrige (*TAXON_NAME_ID*) et ceux de TAXREF (les *CD_NOM*). Cependant ce transcodage entre TAXREF et le référentiel taxinomique de Quadrige n'est pas exhaustif : 93% des taxons Quadrige ont une correspondance avec un *CD_NOM*. Ainsi, les 7% restants ne pourront pas faire l'objet des tests biogéographiques décrits par la suite puisqu'ils se basent sur le référentiel TAXREF.

TAXREF est ainsi un référentiel reconnu à l'échelle nationale, qui à la différence du WoRMS intègre des spécificités françaises (parfois non considérées au niveau mondial), mais peut aussi être moins exhaustif, référençant moins de taxons que le WoRMS. Il est également pris en compte lors des mises à jour du référentiel taxinomique Quadrige et fait foi en cas de désaccord avec le WoRMS. N'étant pas encore exhaustif sur le milieu marin, il est moins intéressant que le WoRMS pour les tests de nomenclature.

De plus, TAXREF consigne aussi les **statuts biogéographiques** (Tableau 1) des taxons dans chaque zone géographique française, notion de présence (présence/absence), d'origine (indigénat ou introduction) et de surface d'aire d'occupation (endémisme) (Gargominy et al., 2020).

Les statuts biogéographiques sont élaborés par des experts en taxinomie et selon des publications. Il convient cependant de considérer que ces statuts sont sujets au changement, suivant les efforts d'observation, les connaissances des spécialistes impliqués, ou l'évolution des aires de répartition des taxons. Ce référentiel ne peut donc pas être utilisé pour toute qualification automatique : une intervention

d'expert sera nécessaire.

Tableau 1 : Codes et statuts biogéographiques TAXREF, présentés dans *(Gargominy et al., 2020).*

STATUT	DESCRIPTION	STATUT	DESCRIPTION	STATUT	DESCRIPTION
P	Présent (indigène ou indéterminé)	J	Introduit envahissant	W	Disparu
E	Endémique	M	Introduit non établi (dont domestique)	X	Éteint
S	Subendémique	B	Occasionnel	Y	Introduit éteint / disparu
C	Cryptogène	D	Douteux	Z	Endémique éteint
I	Introduit	A	Absent	Q	Mentionné par erreur

Par ailleurs les taxons possèdent un statut « Habitat » pouvant prendre différentes valeurs numériques selon le(s) type(s) de milieu colonisé(s) (Tableau 2).

Tableau 2 : Valeurs de la table TAXREF_HABITATS (Gargominy et al., 2020), présentant les valeurs que peuvent prendre les statuts indiquant les habitats colonisés par les taxons. Les statuts soulignés feront l'objet des tests biogéographiques.

HABITAT	DESCRIPTION	REMARQUES
1	Marin	Espèces effectuant l'intégralité de leur cycle de vie en milieu marin. Les espèces vivant en mer mais pouvant occasionnellement supporter les eaux douces entrent dans cette catégorie (exemple de la sardine).
2	Eau douce	Espèces effectuant l'intégralité de leur cycle de vie en eau douce. Les espèces vivant en eau douce mais pouvant occasionnellement supporter les eaux saumâtres entrent dans cette catégorie.
3	Terrestre	Espèces vivant uniquement en milieu terrestre.
4	Marin & Eau douce	Espèces pouvant être présentes en eau douce et en mer de par leur cycle de vie diadrome pour les organismes amphihalins, ou par tolérance aux fortes variations de salinités pour les organismes euryhalins.
5	Marin & Terrestre	Espèces effectuant une partie de leur cycle de vie en mer et l'autre partie à terre (cas des pinnipèdes, des tortues et des oiseaux marins par exemple).
6	Eau saumâtre	Espèces vivant exclusivement en eau saumâtre.
7	Continental (terrestre et/ou eau douce)	Espèces continentales (non marines) dont on ne sait pas si elles sont terrestres et/ou d'eau douce (taxons provenant de Fauna Europaea).
8	Continental (terrestre et eau douce)	Espèces terrestres effectuant une partie de leur cycle en eau douce (odonates par exemple), ou fortement liées au milieu aquatique (loutre par exemple).

Si le statut habitat a une valeur unique par taxon, le statut biogéographique peut différer selon les territoires pour un même taxon, comme présenté dans l'exemple du Tableau 3.

Tableau 3 : Extrait du référentiel TAXREF, présentant le statut Habitat et les statuts biogéographiques TAXREF du taxon *Thalassiosira eccentrica*. Les abréviations des noms de colonne correspondent aux territoires français (*e.g.* FR = France métropolitaine, GF = Guyane Française, TA = Terre Adélie, PF = Polynésie Française).

Nom taxon valide	Habitat	FR	GF	TA	PF
Thalassiosira eccentrica (Ehrenb.) Cleve, 1904	1	D	*NULL*	P	P

2.4 Sélection d'un jeu de données test pour éprouver les tests de la procédure

Le SI Quadrige compile plus de 5 millions d'occurrences taxinomiques. Afin de faciliter les investigations et d'effectuer plus rapidement certains tests à mettre en œuvre, un jeu de données a été ciblé : le REseau de surveillance de PHYtoplancton et de phycotoxines (**REPHY**), dans le cadre duquel des données taxinomiques sont acquises. En effet, ce jeu de données s'étend sur une large période de temps (depuis 1984) et présente également un volume important d'occurrences taxinomiques (près d'un million, soit 18% des occurrences taxinomiques de la BD Quadrige), ainsi qu'une diversité élevée de taxons (~1300, soit 15% des taxons impliqués dans les résultats de la BD Quadrige) (Figure 6). Toutefois dans le cadre des travaux que j'ai mené, seules les données validées de 2010 à 2020 ont été considérées. En effet, ces données étant assez récentes, il est fort probable que les intervenants soient encore disponibles ce qui facilite les éventuelles corrections.

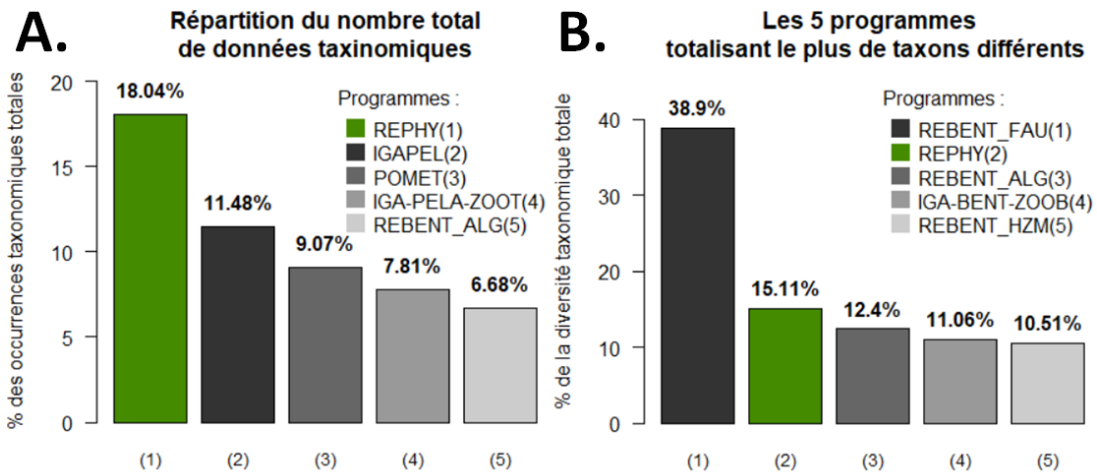


Figure 6 : Présentation des cinq programmes totalisant le plus d'occurrences taxinomiques[6] (A.) et la plus grande diversité taxinomiques (B.) dans la BD Quadrige *(totaux parmi toutes les données taxinomiques de Quadrige).*

Cette quantité et diversité de données apporteront ainsi de la pertinence aux études statistiques les concernant. De plus, ces données possèdent un protocole d'acquisition et de contrôle cadré, ce qui permet d'obtenir en amont des données cohérentes et analysables.

2.5 Organisation de la procédure

Afin de rendre la procédure opérationnelle, elle doit être documentée. Les tests doivent notamment être décrits avec précision pour que les acteurs de la qualification sachent comment les exécuter. J'ai donc proposé un modèle de fiche standard de description des tests. La procédure de qualification biogéographique des données taxinomiques devant s'intégrer aux autres procédures déjà en place, je me suis inspirée des méthodes et du vocabulaire déjà mis en place par la Cellule Quadrige, pour respecter une certaine homogénéité dans les documents et en favoriser l'appropriation par la Cellule Quadrige. J'ai également repris les concepts de certains éléments présentés dans la description technique des tests du projet Kurator (Morris et al., 2018). La liste des champs retenus dans ces fiches est présentée dans le Tableau 4.

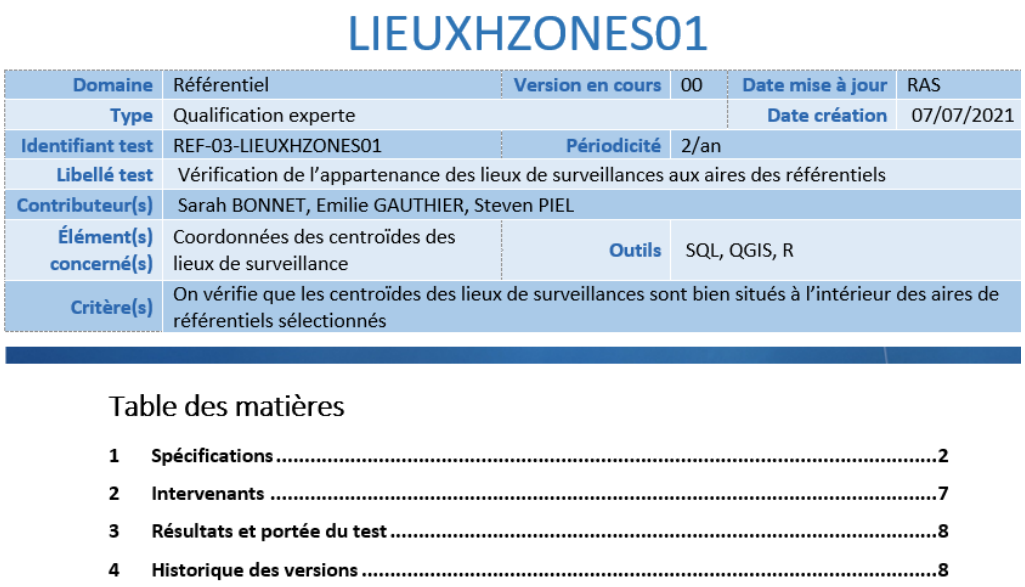
Tableau 4 : Description du format standard des fiches de tests : liste des champs et définition.

Champ de la fiche	Définition
Libellé court du test	Code alphanumérique bref unique pour chaque test permettant d'identifier rapidement le test concerné. Il permet de nommer le test par ailleurs (*e.g.* dans d'autres documents).
Domaine	Périmètre dans lequel le test intervient, ce qu'il implique, ce qui est impacté.
Version	Numérotation des différentes versions du test (permet d'en suivre les évolutions).
Date de mise à jour	Date de la dernière mise à jour du document.
Date de création	Date de création du document.
Historique des versions	Tableau permettant de dater les différentes versions, mais aussi d'identifier les modifications apportées et les personnes à l'origine de ces modifications.
Type de test	Indique s'il s'agit d'une opération effectuée automatiquement ou manuellement.
Identifiant	Code structure permettant d'identifier et d'ordonner les tests selon le séquençage de la procédure. Il est repris pour le nom du fichier de la fiche procédure. Pour cela, il respecte une forme standard : • *Le **domaine** du test, désigné par une courte séquence alphabétique* ; • *Le numéro indiquant l'ordre **d'exécution** des tests au sein de chaque domaine ;* • *Le libellé court **du test*** ; • *Le **numéro** de la **version** désignée (facultatif, selon si une version en particulier est désignée).* *e.g.* **REF-01-TAXOWORMS01-v00**
Périodicité	Indique à quelle fréquence le test est exécuté.

[6] Une occurrence taxinomique : une identification individuelle d'un taxon sur un lieu donné à une date donnée.

Champ de la fiche	Définition
Libellé long	Nom littéral complet du test, assez explicite pour comprendre le sujet et la problématique du test.
Contributeurs	Noms des agents ayant participé à l'élaboration du test ou à sa mise en œuvre (auteurs, relecteurs, experts sollicités…).
Eléments concernés	Indication du périmètre de données et des métadonnées concernées par le test.
Outils	Langages de scripts, logiciels et scripts mis en jeu dans les tests. Permet d'identifier rapidement les compétences nécessaires pour exécuter, modifier ou corriger le test.
Critère(s)	Critère utilisé pour discriminer les éléments sujets au test, la règle suivie.
Spécifications	Description pas à pas de la démarche pour réaliser le test : nom et emplacement des fichiers impliqués, méthode d'utilisation des scripts, méthodologie de traitement des résultats, intégration des corrections / qualification en base.
Intervenants	Définition des acteurs prenant part à chaque étape de l'exécution du test.
Résultats et interprétation	Description des résultats produits par le test, et de l'interprétation à adopter pour déterminer les différents niveaux de qualité des données selon ces résultats. Cela permet un traitement objectif et de garantir l'homogénéité dans l'affectation des niveaux de qualité.

Un extrait des fiches techniques des tests de la procédure mis en place est présenté en Figure 7.



LIEUXHZONES01

Domaine	Référentiel	**Version en cours**	00	**Date mise à jour**	RAS
Type	Qualification experte			**Date création**	07/07/2021
Identifiant test	REF-03-LIEUXHZONES01	**Périodicité**	2/an		
Libellé test	Vérification de l'appartenance des lieux de surveillances aux aires des référentiels				
Contributeur(s)	Sarah BONNET, Emilie GAUTHIER, Steven PIEL				
Élément(s) concerné(s)	Coordonnées des centroïdes des lieux de surveillance	**Outils**	SQL, QGIS, R		
Critère(s)	On vérifie que les centroïdes des lieux de surveillances sont bien situés à l'intérieur des aires de référentiels sélectionnés				

Table des matières

Figure 7 : Exemple de fiche technique d'un test de la procédure.

Grâce à ce cadre d'élaboration de la procédure, il a alors été possible de mettre en place les tests de la procédure et les ajuster plus facilement au fur et à mesure, en fonction des cas pratiques rencontrés.

3. Vers une procédure de qualification biogéographique des données taxinomiques

L'ensemble de mes travaux a permis d'identifier des contrôles qualité concernant l'aspect biogéographique des données taxinomiques. Ces contrôles se présentent sous la forme d'un ensemble de tests et de démarches de veille de référentiels. Ces tests sont présentés selon un ordre défini qui permet :

- de respecter la hiérarchisation des informations contrôlées (*e.g.* vérifier les référentiels avant de vérifier les données saisies sur ces référentiels) ;
- d'éviter le doublonnage des erreurs détectées, en effectuant les tests globaux avant de vérifier les règles plus fines et précises ;
- de garantir la bonne exécution des tests de la procédure.

Les résultats des tests qui ont pu être effectués sur le jeu de données REPHY depuis 2010 bancarisé dans la BD Quadrige sont consignés et discutés dans le présent rapport. D'autres tests sont envisagés pour compléter la procédure mise en place mais n'ont, à ce jour, pas pu être développés ; ils sont présentés à l'état de préconisations.

3.1 Vérification des référentiels Quadrige

Avant le lancement de tout test, il est primordial de disposer de référentiels complets et à jour. En effet, cela est essentiel pour distinguer les erreurs dans les informations saisies par les utilisateurs de celles provenant des référentiels. Par exemple, si un lieu de surveillance est mal géolocalisé, les passages héritant de sa géométrie seront, par relation d'héritage, mal géolocalisés eux aussi. Il faut donc d'abord vérifier la cohérence des référentiels avant de vérifier la cohérence des données qui s'appuient dessus, afin de ne pas identifier de façon multiple une même source d'erreur.

3.1.1 Référentiels géographiques : les lieux de surveillance

Pour rappel, dans le SI Quadrige, les lieux peuvent être ponctuels, linéaires ou surfaciques. Afin d'homogénéiser leur traitement, les coordonnées retenues pour les tests sont celles de leur centroïde.

Selon le principe d'héritage entre les différents niveaux des données, expliqué partie 1.2, une erreur de localisation au niveau du lieu de surveillance va impacter toutes les données qui lui sont rattachées et qui héritent de ses coordonnées. Afin d'éviter cet enchaînement d'erreurs, il est donc essentiel de contrôler les coordonnées des lieux de surveillance en premier. De plus, la validité de ces coordonnées permettra d'obtenir des résultats corrects dans les tests qui les exploitent.

Ce contrôle consiste à vérifier que les lieux de surveillances se situent dans l'emprise d'au moins un des 3 référentiels externes retenus (*cf.* § 2.3.1) ; pour rappel il s'agit de l'espace maritime français, des masses d'eau DCE françaises et des zones marines Quadrige. Si c'est le cas, les coordonnées du lieu de surveillance sont considérées comme cohérentes. Dans le cas contraire, il convient de déterminer la raison pour laquelle le centroïde de ce lieu se situe hors de l'emprise de chacun des référentiels : *e.g.* une erreur dans les coordonnées, une erreur due à l'approximation faite par l'utilisation des coordonnées des centroïdes, ou une imprécision des référentiels. S'il réside une possibilité d'erreur dans les coordonnées, il faut alors retourner vers la personne à l'origine de la demande de création du lieu de surveillance pour une expertise manuelle, afin d'en corriger les coordonnées si nécessaire, ainsi que celles des passages en ayant hérité. Afin d'optimiser la répétition de ce test, une liste des identifiants des lieux hors zonages de référentiel déjà expertisés et dont la localisation a été estimée comme correcte a été établie : ils seront exclus du test lors des prochaines exécutions.

Tableau 5 : Résultats et suites à donner des tests sur la localisation des lieux de surveillance selon l'emprise des référentiels géographiques, effectués sur les lieux de surveillance REPHY.

Résultats	Actions mises en œuvre	Suites actions
Comme présenté dans la Figure 8, dans le cadre du REPHY, une partie des lieux se situe dans un ou plusieurs référentiels. Cependant, sur les 1 226 lieux de surveillance REPHY, **5%** se situent hors référentiels (**66** lieux ponctuels, aucun lieu surfacique ou linéaire).	Des lieux se situant dans certains référentiels mais pas dans d'autres prouvent la nécessité de cumuler les trois référentiels sélectionnés. Depuis 2010, ces lieux hors zones ont vu 931 passages hériter de leurs coordonnées (soit 3% des passages REPHY), ce qui démontre l'importance de la validité de la localisation des lieux de surveillance dans la pratique.	Des échanges sont en cours concernant des lieux de surveillance REPHY hors zone, afin de déterminer les corrections et qualifications à apporter à ces lieux et aux données rattachées. Ce test a également mis en évidence un besoin de mise à jour du référentiel des zones marines Quadrige détaillé au § 5.3.

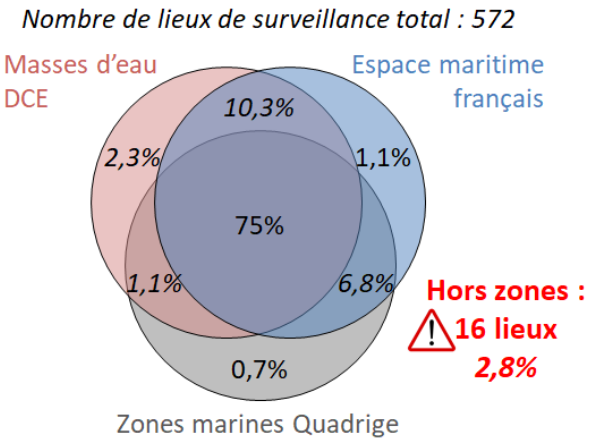


Figure 8 : Diagramme de Wenn présentant la proportion de lieux de surveillance REPHY (centroïdes) situés dans chaque référentiel et en dehors.

3.1.2 Référentiels taxinomiques et biogéographiques

Les données taxinomiques et biogéographiques de Quadrige s'appuient sur des référentiels internes, qu'il convient de maintenir à jour selon différents référentiels externes. Ces référentiels externes offrent aussi la possibilité d'opérer des vérifications sur des jeux de données précis.

3.1.2.1 Cohérence de la nomenclature selon le WoRMS

A ce stade de la procédure de qualité biogéographique, il convient de s'assurer de la concordance du référentiel taxinomique Quadrige avec celui du WoRMS, en effectuant une mise à jour sur la base de celui-ci. Cette opération étant longue, ayant un grand impact sur les données, et nécessitant donc un niveau d'expertise que je n'ai pas eu le temps d'acquérir durant cette année, elle a été laissée à la charge de la Cellule Quadrige. Cependant, elle constitue une étape importante de la procédure, qu'il convient tout de même d'expliquer ici.

Pour mener à bien l'opération, un manuel de procédures semi-automatiques a été rédigé antérieurement à mon alternance, à destination de la Cellule Quadrige. Ce manuel détaille les outils et étapes à suivre. Il est à noter que ce manuel décrit une mise à jour s'effectuant sur la totalité du référentiel taxinomique de Quadrige, la rendant coûteuse en terme de temps (au moins 10 jours), ce qui ne permet pas de l'effectuer plus d'une fois par an. Cependant, la présente procédure de qualification biogéographique des données taxinomiques nécessite la meilleure correspondance possible entre les référentiels taxinomiques en interne et en externe. Il a donc été décidé d'effectuer la mise à jour du référentiel taxinomique Quadrige selon la méthodologie du manuel, mais sur un ensemble délimité de taxons, estimés d'intérêt car mis en jeu dans les données à qualifier. De cette manière, le nombre de taxons à actualiser et donc la durée de traitement sont réduits, ce qui rend la tâche plus aisément applicable. Ainsi, il est envisageable de l'exécuter plus régulièrement.

Lors de cette mise-à-jour, il y a deux étapes principales :

- identifier les taxons du référentiel interne de Quadrige ne possédant pas d'*AphiaID*, afin de leur en fournir un dès que c'est possible : les taxons sans *AphiaID* ne pouvant pas être traités par les opérations suivantes, ceci doit être effectué en premier ;
- pour les taxons avec un *AphiaID*, se référer à leur statut dans le WoRMS afin d'effectuer en interne les modifications nécessaires. Il faut notamment prêter attention :
    - aux statuts « *deleted* », ou « *quarantined* », nécessitant de changer en interne l'*AphiaID* du taxon concerné ;
    - aux changements de synonymie (taxon référent devenu synonyme et inversement) ;
    - pour les taxons synonymes, aux changements du taxon référent indiqué ;
    - aux modifications du nom et/ou de l'auteur.

En cas d'informations ne respectant pas la structuration du référentiel interne Quadrige, il est à la charge de l'agent effectuant la mise à jour de déterminer les rectifications à apporter pour rétablir la cohérence par rapport à la nomenclature WoRMS.

3.1.2.2 Taxons aquatiques, selon les référentiels externes retenus

Le périmètre thématique des données taxinomiques du SI Quadrige couvrant le domaine marin, les estuaires et les lagunes, les taxons renseignés dans la BD Quadrige doivent en théorie pouvoir se retrouver *a minima* dans un de ces milieux. Pour vérifier cette affirmation, il est possible d'utiliser les informations d'habitat présentes dans les référentiels WoRMS et TAXREF. Ils possèdent chacun des informations sur les milieux de vie potentiels occupés par les taxons. A noter que cette vérification n'est possible que pour les taxons possédant en base un *AphiaID* (correspondance avec le WoRMS), ou un *CD_NOM* (correspondance avec TAXREF). Autre limite de ce test : le remplissage des informations d'habitat dans TAXREF et WoRMS n'est pas exhaustif : *i.e.* 23% des taxons Quadrige possédant un *CD_NOM* n'ont pas de statut Habitat dans TAXREF.

Pour les taxons impliquant des données dans la BD Quadrige et ayant une correspondance avec un AphiaID et/ou un CD_NOM, le test consiste à extraire de la BD des informations relatives aux taxons : nom, auteurs ayant décrit ce taxon, *TAXON_NAME_ID*, *AphiaID*, *CD_NOM*. Ensuite, les informations relatives aux milieux de vie colonisés par ces mêmes taxons sont récupérées dans le WoRMS et TAXREF. Les taxons indiqués absents du milieu aquatique selon au moins un des deux référentiels sont extraits pour expertise :

- pour WoRMS : le test doit relever les taxons dont les statuts *isMarine*, *isBrackish* ou *isFreshwater* (*cf.* § 2.3.2.1) ont tous trois la valeur de 0, ce qui signifie qu'ils sont absents du milieu aquatique marin, des eaux douces et des eaux de transition ;
- pour TAXREF : le test doit relever les taxons dont le statut Habitat TAXREF indique qu'ils se retrouvent uniquement en milieu terrestre (statut = 3) ou en milieu continental (statut = 7 ou 8) (Tableau 2). Autrement dit seront vérifiés les taxons étant indiqués absents des milieux aquatiques marins et des eaux de transition.

A l'issue de ce test, une liste de taxons *a priori* non marins, estuariens et/ou lagunaires est produite. Cette liste doit être transmise à des experts en taxinomie pour analyse afin de déterminer des suites à donner. Selon le retour de ces experts, la Cellule Quadrige devra, au cas par cas, décider de l'action à entreprendre, *e.g.* d'éventuelles corrections du référentiel taxinomique Quadrige, un signalement auprès des producteurs de données concernés par d'éventuelles corrections afin d'éviter la répétition de l'erreur *a posteriori*, une qualification à « douteux » ou « faux » des données dont le taxon n'est pas corrigeable, *etc*.

Pour optimiser la répétition de ce test et éviter les expertises redondantes, il conviendrait de renseigner une liste de taxons déjà expertisés en interne et estimés présents en milieu aquatique, afin de les exclure de l'analyse dès son début.

Tableau 6 : Résultats et suites à donner des tests sur la cohérence des milieux de vie connus des taxons selon le périmètre d'actions du SI Quadrige, effectués sur les taxons du référentiel Quadrige impliqués dans des données.

Résultats	Actions mises en œuvre	Suites actions
Parmi les données taxinomiques Quadrige : - aucun taxon non aquatique selon le WoRMS ; - 25 taxons ont un statut habitat TAXREF non marin ou continental dont : • 15 marins selon le WoRMS ; • 5 taxons indiqués non marins dans le WoRMS parmi lesquels un seul uniquement terrestre ; • 5 taxons absents du WoRMS, donc suspectés non marins.	RAS	Les 10 taxons suspectés « non marins » doivent être expertisés par des experts thématiques avec sollicitation éventuelle des producteurs de données. Selon leurs retours, la marche à suivre reste à définir.

La cohérence des référentiels géographique, taxinomique et biogéographique ainsi vérifiée, la suite des tests peut alors traiter des données et métadonnées saisies par l'utilisateur, se basant pour certaines sur les référentiels.

3.2 Vérification de la temporalité

Les métadonnées temporelles d'une donnée sont cruciales. En effet, selon la période de l'année voire l'heure d'observation, les tendances d'observation des espèces aquatiques peuvent différer (*e.g.* variations

saisonnières de présence des espèces, variations quotidiennes selon l'alternance jour/nuit, changements selon les conditions du milieu comme la température). Enfin, lors des études ou tests effectués sur un intervalle de temps donné, une erreur de date peut entraîner l'inclusion ou l'exclusion à tort de certaines données. Il paraît alors essentiel de vérifier la cohérence des métadonnées ayant trait à la temporalité.

3.2.1 Cohérence des dates

La cohérence de la date renseignée en base peut être vérifiée selon son contexte, *e.g.* en vérifiant si la date se situe bien dans un intervalle de dates possibles. Autrement dit, en premier lieu, il est possible de borner les années saisies entre celle la plus reculées qu'il est possible de trouver, et la plus récente. Ce type de contrôle est inspiré d'un test du projet Kurator (#84) (Morris et al., 2018). Il est assuré dans le SI Quadrige *via* les stratégies d'acquisition des données qui limitent la saisie de la date d'une observation à des intervalles de temps sur lesquels les suivis sont censés être effectués. Par le biais d'une contrainte applicative, il est ainsi impossible de saisir des données sous une stratégie si la date d'observation de la donnée ne se situe pas dans l'intervalle de temps durant laquelle la stratégie est active.

3.2.2 Cohérence entre les différentes dates d'une même donnée

Il convient d'effectuer des tests sur les dates d'observation, de validation et du jour qui consiste à vérifier si leur chronologie est cohérente (Figure 9). Il est important d'exécuter ces tests régulièrement pour signaler rapidement les anomalies après leur validation, de sorte que les producteurs de données disposent encore des fiches terrain et laboratoire et effectuent une vérification. Pour cela, ces tests doivent être réalisés dans le cadre des jobs de l'outil « AlerteAno » de la Cellule Quadrige qui détectent quotidiennement l'apparition de nouvelles anomalies.

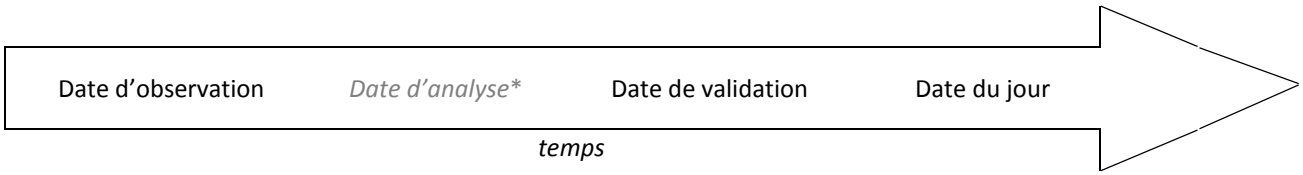


Figure 9 : Chronologie de différentes dates présentes, ou à venir, des données du SI Quadrige. ** La date d'analyse n'est pas bancarisée dans Quadrige* (*cf.* § 5).

Ce test s'effectue avec une requête SQL, intégrable à l'outil « AlerteAno » de la Cellule Quadrige.

Tableau 7 : Résultats et suites à donner des tests effectués sur la chronologie des dates.

Résultats	Actions mises en œuvre	Suites actions
Test effectué : la date observation d'une donnée est antérieure à sa date de validation. Une donnée a été détectée dans l'ensemble de la BD Quadrige.	Cette donnée a été renvoyée à son producteur qui a corrigé l'erreur sur la date d'observation.	Ce test a été mis en production *via* l'outil « AlerteAno » quotidien de la Cellule Quadrige depuis le 16/06/2021. Aucun nouveau cas n'a été détecté depuis.

3.3 Vérification des métadonnées de localisation

Avant de vérifier la cohérence du couple taxon – localisation (biogéographie) (*cf.* § 3.5), il convient de s'assurer de la validité des métadonnées de localisation séparément. Pour rappel, les lieux, passages et prélèvements pouvant avoir une géométrie ponctuelle, linéaire ou surfacique, les tests de localisation concerneront le centroïde des différentes géométries, car il permet de résumer une emprise complexe de type ligne ou polygone à un couple de coordonnées X, Y, format retrouvé chez les lieux ponctuels.

3.3.1 Coordonnées nulles

Le projet Kurator (Morris et al., 2018) mentionne un test (#87) visant à alerter en cas de valeurs de longitude et/ou de latitude égale(s) à 0. Ce cas de figure est certes possible mais il est plus probable qu'il provienne d'une erreur de saisie ou de manipulation.

Dans le SI Quadrige, ce test peut s'appliquer aux champs des coordonnées des centroïdes des passages et des prélèvements. Un prélèvement pouvant hériter des coordonnées du passage auquel il est rattaché,

ce test doit donc être réalisé en deux temps pour éviter les doublons inutiles d'anomalies : tout d'abord sur les passages, puis après traitement des résultats des passages sur les prélèvements.

Ce test peut être effectué par une requête SQL, et est donc tout à fait intégrable aux autres jobs « AlerteAno » exécutés par un programme Talend quotidiennement à la Cellule Quadrige. La requête créée dans le cadre de ce projet permet ainsi de fournir un fichier récapitulant tous les cas concernés et les informations nécessaires pour permettre aux experts de vérifier l'exactitude des coordonnées au cas par cas, et corriger ou qualifier les coordonnées concernées.

Tableau 8 : Résultats et suites à donner des tests effectués sur les coordonnées nulles pour les centroïdes des passages.

Résultats	Actions mises en œuvre	Suites actions
Tests effectués : - Latitude et/ou longitude de centroïde de passage = 0 : aucun cas ; - Latitude et/ou longitude de centroïde de prélèvement = 0 : aucun cas.	RAS	Ce test a été mis en production *via* l'outil « AlerteAno » quotidien de la Cellule Quadrige depuis le 28/07/2021. Aucun nouveau cas n'a été détecté depuis.

3.3.2 Signe(s) de la valeur de latitude et/ou longitude

Des erreurs de signes de longitudes ont été signalées à la Cellule Quadrige à l'occasion d'une expertise de données Quadrige effectuée par le MNHN. Par exemple, la Figure 4 (*cf.* § 1.3.2) met en évidence des points correspondant au littoral breton mais disposés en symétrique par rapport au méridien de Greenwich (0° longitude) : il s'agit d'erreur de signe de la longitude.

Dans le projet Kurator (Morris et al., 2018) est mentionné un test (#54) de modification automatique du signe de la latitude/longitude des données, dans le cas où les coordonnées ne seraient pas cohérentes avec la zone indiquée : la modification est effectuée automatiquement si la cohérence est rétablie en inversant le signe de la latitude et/ou de la longitude des coordonnées.

Dans le SI Quadrige, un contrôle sur les coordonnées des passages fait déjà l'objet d'une recherche quotidienne *via* l'outil « AlerteAno ». Les inversions de signes entre les coordonnées des passages et celle des lieux de surveillance associés sont recherchées. Toutefois, ce test se limitait en l'état aux lieux et passages à géométrie ponctuelle. Les entités linéaires et surfaciques n'étaient pas prises en compte. J'ai ainsi testé les inversions de signe de latitude et longitude des centroïdes des passages par rapport au centroïde de leur lieu de rattachement. Cela a pu mettre en évidence des anomalies (Figure 10).

Il est à noter que les lieux surfaciques et linéaires chevauchant le méridien de Greenwich (0° longitude) ont été exclus du test. En effet, leurs centroïdes et ceux des passages associés peuvent se situer de part et d'autre du méridien. Donc pour ces lieux, une différence de signe sur la valeur de la longitude n'est pas nécessairement une erreur.

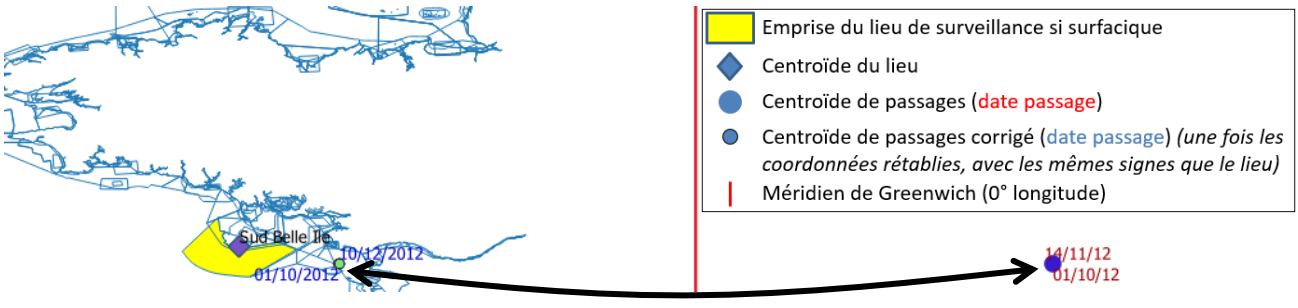


Figure 10 : Exemple d'inversions du signe de la longitude dans des données Quadrige (Méridien de Greenwich : trait rouge).

Ce test est effectué *via* une requête SQL, intégrable aux jobs « AlerteAno » exécutés quotidiennement par la Cellule Quadrige. Il produit un fichier récapitulatif avec les cas concernés et toutes les informations utiles associées. Les producteurs de données concernés sont alors sollicités pour contrôler les passages signalés. Ensuite la Cellule Quadrige effectue les corrections nécessaires au cas par cas.

Tableau 9 : Résultats et suites à donner des tests effectués sur les inversions de signes de latitude et de longitude.

Résultats	Actions mises en œuvre	Suites actions
Test préexistant effectué : Lieux et passages ponctuels : longitude ou latitude du lieu avec signe inverse à la longitude ou latitude du passage : 3 cas ont été détectés au cours de mon contrat. *Test ajouté dans le cadre de mes travaux :* Pour tous les lieux et passages : longitude ou latitude du centroïde du lieu avec signe inverse à la longitude ou latitude du centroïde du passage : 74 cas (après exclusion des lieux linéaires et surfaciques chevauchant le méridien de Greenwich).	Données en cours de correction : - 1 passage supprimé (erreur de saisie) - 7 passages avec coordonnées corrigées - 15 passages ont changé de lieu de surveillance dont 6 passages dont les longitudes ont été aussi inversées - 33 passages dont la longitude a été inversée permettant de réintégrer le lieu de surveillance - 18 passages dont les coordonnées redéfinies avaient été dupliquées à tort depuis un précédent passage et dont les coordonnées ont été ré-héritées du lieu	Test mis en production *via* l'outil « AlerteAno » quotidien de la Cellule Quadrige depuis le 13/09/2019. Les cas détectés depuis ont été corrigés au fil de l'eau. Quand les corrections évoquées précédemment seront effectives en BD, une requête pourra être ajoutée à l'outil « AlerteAno » quotidien pour tenir compte des centroïdes.

3.3.3 Cohérence de la localisation des passages selon l'emprise des référentiels

La correction des signes de la latitude et de la longitude ayant permis de corriger des erreurs assez évidentes de coordonnées, il est maintenant possible d'affiner le contrôle des localisations en recherchant les données de surveillance du littoral localisées hors des référentiels marins externe (*cf*. § 2.3.1). J'ai ainsi mis en place un test qui identifie les cas des données hors zonages maritimes de référence, selon un raisonnement similaire à celui du test effectué sur les lieux de surveillance (*cf*. § 3.1.1).

L'analyse ne concerne que les passages avec des coordonnées réelles. En effet, étant donné que les coordonnées des lieux de surveillance ont été vérifiées précédemment, effectuer le test sur les coordonnées de passages héritées serait redondant. Le test consiste à récupérer les informations des passages dont les centroïdes ne sont situés dans aucune des zones des référentiels choisis. À l'aide d'une requête SQL, les données à traiter et les coordonnées de leur centroïdes sont extraites. Ensuite, sur QGIS, des jointures spatiales sont réalisées entre les couches cartographiques des référentiels et les données. Un script R est ensuite lancé sur le fichier résultant des jointures afin d'en extraire les données n'ayant été jointes avec aucun des référentiels, donc hors référentiels. Le résultat de ce test se présente sous la forme d'un CSV des cas de données hors référentiels et de leurs informations associées, afin de pouvoir les soumettre aux producteurs de données concernés pour vérification. Ensuite, selon leur expertise, un agent de la Cellule Quadrige va effectuer dans la BD les éventuelles corrections et la qualification de la localisation de ces passages. Afin d'optimiser la répétition de ce test, il est nécessaire de tenir une liste des passages hors référentiels déjà expertisés dont la localisation a été estimée juste, afin de les exclure des résultats du test avant expertise.

Tableau 10 : Résultats et suites à donner des tests effectués sur l'emplacement des centroïdes de passages selon l'emprise des référentiels géographiques.

Résultats	Suites données	Suites
Parmi les 33 434 passages REPHY depuis 2010 : - 2,9% étaient hors zone, soient 971 passages - 0,012% étaient hors zone et possédaient des coordonnées réelles, soient 40 passages	Après une comparaison des coordonnées des centroïdes des passages hors référentiels avec celles des centroïdes de leurs lieux de surveillance, il s'est avéré que 3 d'entre eux provenaient probablement d'une erreur de saisie (*e.g.* oubli de saisie d'une décimale de la valeur de latitude comme le cas de la Figure 11)	RAS

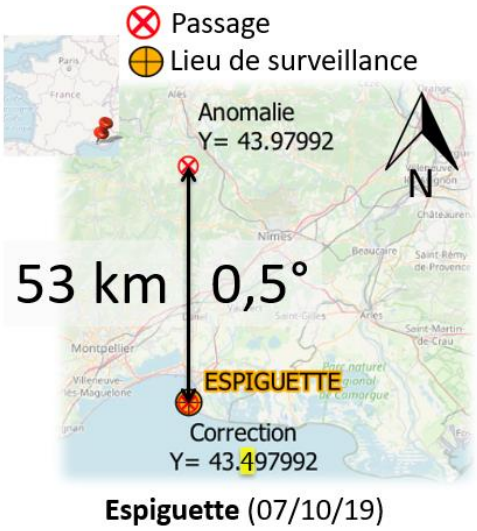


Figure 11 : Exemple d'un cas de coordonnées douteuses sans inversion de signe.

3.3.4 Distances entre les passages et leur lieu de surveillance, et entre les passages d'un même lieu de surveillance

3.3.4.1 Principe des tests

Jusqu'ici, les tests ont permis de corriger les inversions de signe des coordonnées et les coordonnées hors zonages de référence. Cependant, certaines coordonnées de centroïdes de passages restent douteuses. En effet, certains passages semblent se situer anormalement loin soit de leur lieu de surveillance, soit des autres passages rattachés au même lieu de surveillance (Figure 12).

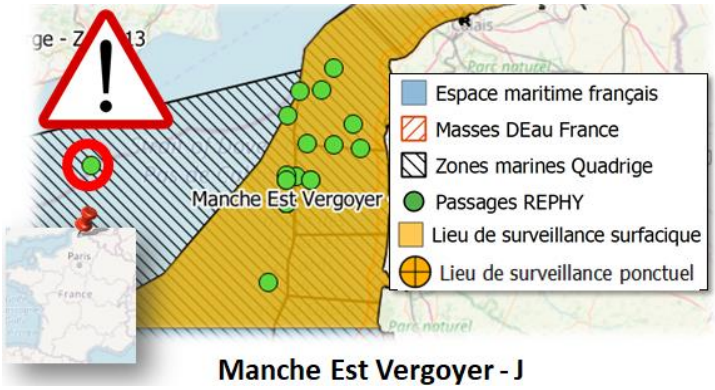


Figure 12 : Passages REPHY rattachés au lieu de surveillance Manche-Est Vergoyer-J. Un passage exceptionnellement éloigné des autres passages est entouré en rouge.

Les deux tests qui suivent visent à mettre en évidence les passages exceptionnellement excentrés par rapport aux autres passages rattachés au même lieu de surveillance. Il est cependant compliqué de définir un seuil fixe de distance au-delà duquel l'éloignement serait considéré comme exceptionnel, car la variabilité de la localisation des observations diffère selon le protocole de prélèvement (à pied, en bateau ou en plongée, observation humaine ou appareil de mesure *in situ*, *etc.*) et la thématique. Ces deux éléments impactant la variabilité sont cependant fixes pour un même lieu. Ainsi, l'analyse des distances s'effectuera séparément lieu par lieu, afin de comparer des distances d'une variabilité théoriquement similaire permettant d'établir une valeur seuil cohérente. Cette valeur seuil correspondra, selon le principe d'un box plot, au 3^{e} quartile + 1,5 l'écart interquartile (Figure 13).

L'intérêt de ces tests est de déceler des erreurs de saisie, peu importe le nombre de données reprenant l'information d'une même saisie. En effet, les passages ayant hérité des coordonnées de leur lieu de surveillance seront représentés une seule fois par le point du lieu de surveillance dans le calcul de l'ensemble des distances car leurs coordonnées résultent d'une même saisie (Figure 14). De plus, une erreur de localisation d'un passage peut se répercuter sur un autre passage lorsque la saisie est effectuée par duplication d'un passage existant (sorte de copier-coller qui reprend les informations d'un passage et les dupliquent sur un autre passage). Ceci peut donc occasionner plusieurs passages superposés dont les coordonnées résultent d'une même saisie originelle : ces passages ne seront représentés que par un point dans le calcul de l'ensemble des distances (Figure 15).

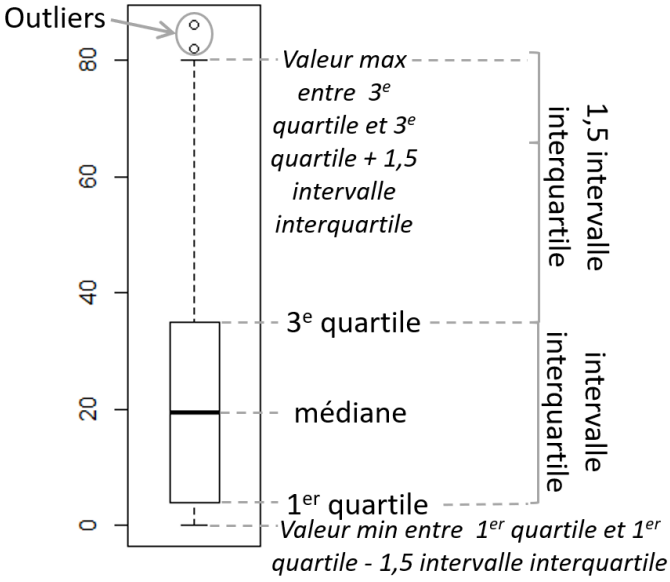


Figure 13 : Principes d'un boxplot.

Figure 14 : Méthode de prise en compte des coordonnées héritées des lieux de surveillance.

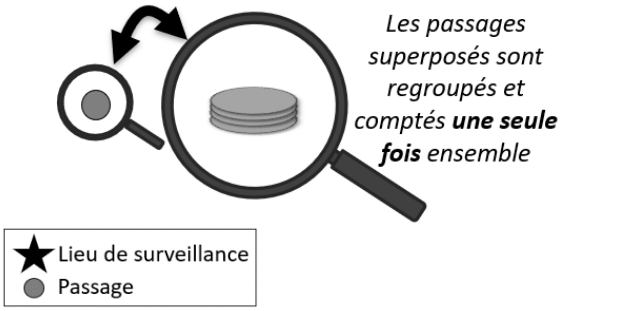


Figure 15 : Méthode de prise en compte des passages superposés hors lieu de surveillance.

3.3.4.2 Distances entre les passages et leur lieu de surveillance

A ce stade de la procédure, suite à l'analyse des cartes de distribution des passages une tendance semblait émerger : les passages aux coordonnées douteuses étaient souvent plus éloignés du centroïde du lieu de surveillance que les autres passages du même lieu.

Il a donc été décidé dans un premier temps d'établir un test pour étudier les distances entre les centroïdes des passages ayant des coordonnées réelles (*cf*. §1.2) et le centroïde de leur lieu de surveillance, afin d'extraire les passages qui en sont exceptionnellement éloignés (Figure 12). Ce test s'exécute lieu de surveillance par lieu de surveillance.

Ce test nécessite d'extraire de la BD les coordonnées et autres informations utiles des lieux de surveillance ainsi que des passages à étudier. Ensuite, un script R calcule la distance entre chaque point unique de passage(s) aux coordonnées réelles (Figure 15) et le lieu de surveillance associé, puis en détermine les valeurs *outliers* (établies selon la fonction boxplot, Figure 13) afin de relever les passages concernés. Un fichier récapitulatif (.csv) contenant toutes les informations de ces passages est produit. Ce fichier est utilisé comme support pour procéder à une qualification experte (*cf.* § 1.3.1.4). Pour optimiser la répétition de ce test, il convient de tenir une liste des passages relevés, expertisés et dont les coordonnées ont été estimées valides, afin de les exclure des résultats par la suite, et ne pas répéter l'expertise.

Tableau 11 : Résultats et suites à donner des tests effectués sur la distance entre les passages et les lieux de surveillance.

Résultats	Actions mises en œuvre	Suites actions
A ce jour, ce test n'a pas été exécuté dans sa version finale : les résultats suivants ont été calculés avec une seule valeur seuil calculée sur la base de tous les passages de l'ensemble des lieux de surveillance du programme REPHY de ces 10 dernières années. Le test a relevé 75 passages excentrés de leur lieu de surveillance, avec des distances allant de 3,8km à 201km.	Les distances relevées étant douteuses dans le cadre de ce programme, j'ai cartographié les passages excentrés et leur lieu de surveillance : le passage le plus éloigné (201km) et d'autres sont effectivement à vérifier. Pour les autres passages, il conviendra de finaliser le test avant toute investigation complémentaire.	Le test est à effectuer dans sa version finale afin de s'assurer de sa pertinence. Il pourra être exécuté ponctuellement et manuellement par un agent, sur un jeu de données délimité (selon les problématiques rencontrées ou d'éventuelles demandes).

3.3.4.3 Distance entre passages eux-mêmes

Une fois les passages anormalement éloignés de leur lieu identifiés et corrigés, il est encore possible d'affiner la recherche de coordonnées exceptionnelles en comparant les coordonnées des passages entre eux. En effet, tout en étant proches de leur lieu de rattachement, il a été constaté que les passages aux coordonnées erronées se trouvaient souvent plus éloignés des autres passages du même lieu de surveillance (Figure 16). Il a donc été décidé d'étudier cette fois les distances entre les passages d'un même lieu de surveillance présentant suffisamment de passages aux coordonnées variable pour permettre au test d'être pertinent.

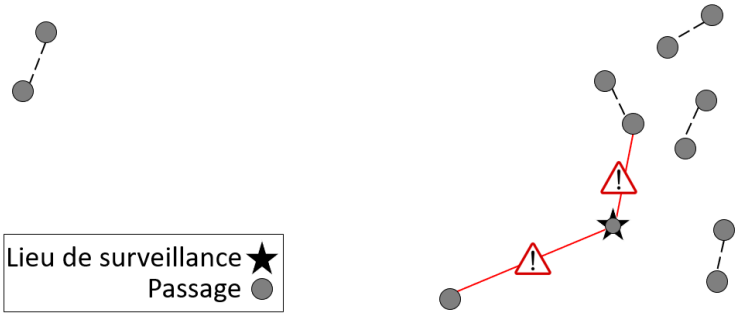


Figure 16 : Schéma explicatif du test des distances exceptionnelles entre passages du même lieu de surveillance, pour un seul lieu de surveillance.

Ainsi, selon un raisonnement similaire au précédent test : il est nécessaire d'extraire de la BD par SQL les coordonnées et informations utiles sur les passages à étudier. Ensuite, grâce à un script R, tous les passages d'un même lieu de surveillance superposés en termes de coordonnées sont convertis en une seule entité (Figure 14 et Figure 15), puis les distances entre passages de même lieu sont calculées. Suite à cette étape, pour chaque passage, la distance minimale au passage le plus proche du même lieu de surveillance est récupérée. Ces distances minimales sont ensuite réunies afin d'extraire les valeurs outliers (établies selon la fonction boxplot) et d'en déduire les passages exceptionnellement excentrés des autres passages de même lieu de surveillance. Comme précédemment, un fichier csv est établi avec les passages retenus pour servir de support à une qualification experte.

Tableau 12 : Résultats et suites à donner des tests effectués sur les distances entre passages d'un même lieu de surveillance.

Résultats	Actions mises en œuvre	Suites actions
Ce test étant dépendant du précédent test (*cf.* § 3.3.4.2) qui n'est pas finalisé, ce test n'a pas pu être exécuté. Cependant, par observation de la cartographie des passages REPHY des 10 dernières années, des cas ont été relevés. Ces passages devront être inclus dans les résultats du présent test afin de vérifier la pertinence de cette analyse.	RAS	Comme pour le test précédent (*cf.* § 3.3.4.2), cette analyse doit être finalisée et pourra être effectuée par un agent sur des jeux de données ciblés, selon les besoins.

3.4 Vérification des informations taxinomiques

3.4.1 Cohérence des taxons avec leur thématique

Les suivis dont les données sont bancarisées dans la BD Quadrige ciblent généralement un compartiment précis de l'environnement avec un groupe taxinomique précis. Dans le cas du programme REPHY, les taxons observés appartiennent au phytoplancton : c'est la thématique étudiée. Grâce à un test effectué précédemment, seuls les taxons considérés comme potentiellement présents dans les milieux des données bancarisées dans le SI Quadrige sont retenus. Cependant, selon la thématique dans laquelle entrent les données, une analyse plus fine est possible. Ainsi, il est intéressant de rechercher les taxons non phytoplanctoniques présents dans les données REPHY.

Or aucun des référentiels Quadrige, WoRMS ou TAXREF ne comportent d'information « thématique » en tant que telle. En revanche, la position du taxon dans l'arbre taxinomique peut fournir des indications sur la thématique. Par exemple la thématique « Poissons » regroupe les taxons fils des branches des vertébrés nommées « *Chondrichthyes* » et « *Osteichthyes* ». Dans le cas du phytoplancton, l'identification de taxons parents est plus compliquée car les branches taxinomiques dont font partie les taxons phytoplanctoniques regroupent d'autres micro-organismes qui ne sont pas phytoplanctoniques. Toutefois, il existe un embranchement ne contenant aucun taxon phytoplanctonique : les « *Animalia* » (animaux).

Un test a donc été mis en place pour rechercher les taxons des données REPHY ayant le taxon *Animalia* comme taxon parent. Ce test consiste en une requête SQL et concerne les données validées non qualifiées. A terme, il pourra être effectué automatiquement *via* l'outil « AlerteAno » de la Cellule Quadrige.

Tableau 13 : Résultats et suites à donner des tests effectués sur la cohérence des taxons avec la thématique des données dans lesquelles ils apparaissent.

Résultats	Actions mises en œuvre	Suites actions
Un taxon animal a été détecté dans 2 résultats REPHY. Il s'agit du genre *Psammodiscus* : le taxon identifié dans les données est *Psammodiscus* Günter, 1862 qui est un poisson, alors qu'il aurait dû s'agir de *Psammodiscus* F.E. Round & D.G. Mann, 1980 qui correspond à une diatomée phytoplanctonique.	Ces observations ont été transmises aux producteurs de données qui ont confirmé l'erreur. Elle reste à corriger par la Cellule Quadrige à la fois dans les données et dans le référentiel. Le taxon Poisson sera « gelé » afin d'empêcher de nouvelles saisies sur ce taxon).	Des groupes thématiques de programmes ont été créés. Il serait intéressant de créer des « groupes de taxons thématiques », à alimenter à partir de chaque taxon nouvellement apparu dans la thématique et validé et/ou de chaque taxon signalé comme hors thématique afin d'identifier automatiquement dans les résultats les taxons de ces listes d'inclusion et d'exclusion thématiques. Cependant, la compétence d'expertise nécessaire pour générer et gérer ces listes est difficilement accessible.

3.5 Vérification de la cohérence biogéographie : statuts biogéographiques TAXREF

Il est possible de vérifier que la **présence** d'un taxon aux coordonnées indiquées en base est cohérente avec les informations biogéographiques mentionnées dans le référentiel sélectionné. Les coordonnées des données de présence d'observations sont comparées, lorsque l'information est disponible, avec les statuts biogéographiques TAXREF des taxons concernés. Les incohérences entre l'emplacement des taxons dans la BD Quadrige et les statuts biogéographiques renseignées dans TAXREF sont relevées. Ces incohérences sont listées dans un fichier récapitulatif afin de les soumettre à une expertise. Sur la base du retour d'expertise, la Cellule Quadrige corrige ou qualifie en conséquence les données expertisées.

Il est à noter que ce test ne peut s'effectuer que sur les données de taxons ayant un CD_NOM (*cf.* § 2.3.2.2) et des statuts biogéographiques renseignés dans TAXREF. Cela représente 82% des 20 429 taxons du référentiel taxinomique Quadrige ayant un *CD_NOM*.

Ce test est exécutable avec une requête SQL et donc intégrable à l'outil « AlerteAno » exécuté quotidiennement par la Cellule Quadrige. Un fichier récapitulatif listant les données concernées et les informations utiles associées serait produit et utilisé pour analyser ces incohérences.

Tableau 14 : Résultats et suites à donner des tests effectués sur la cohérence biogéographique selon les statuts biogéographiques TAXREF.

Résultats	Actions mises en œuvre	Suites actions
Test effectué sur les données REPHY métropole 2010-2020 : 794 résultats concernant 15 taxons dont le statut biogéographique TAXREF est « D » (douteux) pour la zone indiquée (soit 0,2% des données analysées).	RAS	Ce test est à finaliser selon les résultats et les retours d'experts, mais il est probable qu'il faille tenir une liste d'exclusion de taxons pour certaines zones de ce test (*e.g.* insectes terrestres de plage pour des eaux littorales).

3.6 Vérification par rapport au paramètre : cas du REPHY, avec les paramètres flore totale et flore indicatrice

Les tests précédents ont permis d'écarter les taxons n'étant pas cohérents à l'échelle du SI Quadrige dans sa globalité puis à l'échelle d'une thématique. Il est maintenant possible d'effectuer un contrôle d'opportunité plus ciblé, *e.g.* en se basant sur les consignes de la stratégie d'acquisition des données si celles-ci conditionnent l'identification des taxons dans les données. Dans le cas du REPHY, les taxons

identifiés sont soumis à des règles précises (Neaud-Masson, 2020). Des erreurs d'identification de taxons sur des paramètres donnés ont été signalées par la coordination du programme REPHY. Cela a motivé l'établissement du présent test qui consiste à relever les incohérences concernant l'identification de taxons par rapport aux stratégies appliquées.

Ce test prend en compte les données sur les paramètres « Flore indicatrice » (FLORIND) et « Flore totale » (FLORTOT) pour lesquelles des suspicions de confusions de saisies sur les paramètres ont été communiquées à la Cellule Quadrige. Un relevé de **FLORTOT** consiste à identifier et à dénombrer tous les taxons présents dans l'échantillon, alors qu'un relevé de **FLORIND** correspond à l'identification et au dénombrement d'une liste de taxons d'intérêt ou dont l'abondance dans l'échantillon dépasse 100 000 cell/L. De plus, le lieu de surveillance donne une indication sur le paramètre censé être mesuré d'après la stratégie de surveillance (tous les paramètres ne sont pas suivis sur tous les lieux). Ce test va donc relever les données sur le paramètre FLORIND ne respectant pas les conditions de la stratégie indiquée lors du relevé ou dont le lieu n'est pas sujet au relevé de ce paramètre (Tableau 15). Un fichier récapitulant tous les résultats sur échantillon associés aux flores relevées par le test est ensuite communiqué à des experts (coordinateurs du réseau de surveillance REPHY). La qualification et la correction de données sont effectuées selon l'expertise par la Cellule Quadrige.

Ce test consiste en une requête SQL et concerne les données validées et non qualifiées. Ensuite un traitement des données relevées en SQL avec un programme R est effectué. Ce test est à effectuer régulièrement afin de corriger des erreurs de pratiques avant de multiplier les anomalies.

Tableau 15 : Méthodologie appliquée pour le test de cohérence entre taxons et résultats sur le paramètre FLORIND.

1. Récupération des données sur le paramètre FLORIND validées et non-qualifiées
2. ***Absence inattendue de taxons*** *:*
   Sélection des flores ne présentant pas de résultat sur la totalité les taxons attendus systématiquement dans une flore indicatrice. → genres *Alexandrium*, *Dinophysis* et *Pseudo-Nitzschia.*
3. ***Présence inattendue de taxons*** *:*
   Sélection des flores possédant un résultat sur taxon inattendu :
   Ceci exclut les résultats sur taxons étant attendus dans une flore indicatrice, c'est-à-dire :
   a) Ceux dont le taxon est attendu selon la stratégie d'acquisition des données :
      ↳ A toute date : genres *Alexandrium*, *Dinophysis* ou *Pseudo-Nitzschia ;*
      ↳ Depuis 2012 **:** *Ostreopsis*, *Gonyaulax spinifera*, *Lingulodinium polyedra*, *Protoceratium reticulatum*, *Vulcanodinium rugosum*, *Karenia mikimotoi*, *Prorocentrum lima.*
   b) Ceux dont la valeur résultat est > 100 000 cell/L.
4. ***Lieux inadéquats :***
   Sélection des flores dont le lieu associé ne fait pas partie des lieux où une stratégie d'acquisition FLORIND s'applique.
5. Regroupement des précédentes sélections en une liste de flores à anomalie sans doublon.
6. Etablissement de la liste de tous les résultats d'échantillons de ces flores pour expertise, en spécifiant le type d'anomalie rencontré pour aider à l'expertise.

Tableau 16 : Résultats et suites à donner des tests effectués sur la cohérence entre taxons et paramètres de flores du REPHY.

Résultats	Actions mises en œuvre	Suites actions
Plusieurs milliers de résultats ont été relevés par le test (~8000 résultats concernant ~1000 flores soit 7% des flores)	La liste des données taxinomiques relevées par le test a été communiquée à la coordination du réseau REPHY pour expertise. Selon son retour, la présence de taxons ne rentrant pas à proprement parler dans le cadre d'acquisition des données FLORIND était parfois due à un approfondissement volontaire d'identification des taxons, ceci pour diverses raisons *(e.g.* à l'occasion d'une opération d'habilitation d'un agent, nécessitant une comparaison de deux flores d'agents différents).	La saisie des données dans l'application Quadrige² est conditionnée par les stratégies : le paramètre pré-rempli dans la grille est défini dans ces stratégies. Les résultats du présent test ont révélé la nécessité de mettre à jour ces paramètres par défaut pour certains lieux selon les mesures réellement effectuées par les laboratoires. Dans d'autres cas, le paramètre associé aux données doit être corrigé.

4. Conclusion

Pour l'établissement de cette procédure, il a tout d'abord été nécessaire de se familiariser avec l'environnement du SI Quadrige. Ceci a permis dans un premier temps de m'approprier les données à qualifier, mais aussi de cibler des métadonnées d'intérêt. Pour cela une analyse bibliographique sur les démarches qualité mises en place pour ce type de données, et sur des pratiques ayant cours dans d'autres SI similaires à Quadrige a été effectuée.

Après traitement, le choix du jeu de données cible, le REPHY 2010-2020, s'est révélé pertinent, car il présentait des cas d'anomalies très divers de causes différentes, tout en étant représentatif de cas présents dans d'autres jeux de données du SI Quadrige, ceci dans une volumétrie de données raisonnable compte tenu de la durée de mes travaux. Par ailleurs, la progression dans mes travaux a été favorisée par la réactivité de la coordination du REPHY, qui a apporté des éléments m'ayant permis de consolider les tests, voire de les mettre en production. J'ai aussi sollicité le réseau d'experts Ifremer. Cette prise de contact a permis d'alimenter mes réflexions et d'orienter certaines recherches, comme celles des données « exceptionnelles » (*e.g.* distances entres les données, *cf*. § 3.3.4).

L'élaboration des tests a représenté une grande part des travaux. En effet, les tests ont été conçus les uns à la suite des autres. Pour chacun d'entre eux, il a été nécessaire d'investiguer dans les données afin de juger de leur pertinence, et à cette occasion des observations d'opportunité ont pu inspirer d'autres tests. Cet avancement en pas à pas a engendré de nombreux nouveaux tests et des bouleversements dans l'ordonnancement de l'ensemble des tests de la procédure. En effet, cette procédure a évolué selon les expériences qu'elle a apportées, et évoluera encore à l'avenir selon les cas de figure rencontrés et la direction que prendront les activités du SI Quadrige. C'est dans l'optique d'une évolution et une amélioration continues qu'une organisation standard de tests a été décrite, comprenant entre autres dans certains cas la gestion d'exceptions qu'il sera certainement nécessaire d'appliquer à l'usage, sous la forme de liste d'identifiants d'éléments à exclure des tests (*e.g* : lieu de surveillance à cheval sur le méridien de Greenwich dans le test d'inversion de signe dans les coordonnées des passages).

Les travaux que j'ai menés cette année ont permis d'identifier et ordonner des tests pertinents dans le cadre de la procédure de qualification biogéographique des données taxinomiques, présentés dans la Figure 17. J'ai pu produire les scripts de la plupart des tests, dont certains sont d'ores et déjà en production et permettent quotidiennement de contrôler la qualité des données du SI Quadrige. L'analyse des cas d'anomalie avec le support de la Cellule Quadrige et des experts thématiques a permis la correction d'informations dans les données. J'ai aussi défini les travaux à poursuivre, en proposant des préconisations de tests ou des évolutions du SI Quadrige à mettre en œuvre, qui sont présentés dans le paragraphe suivant.

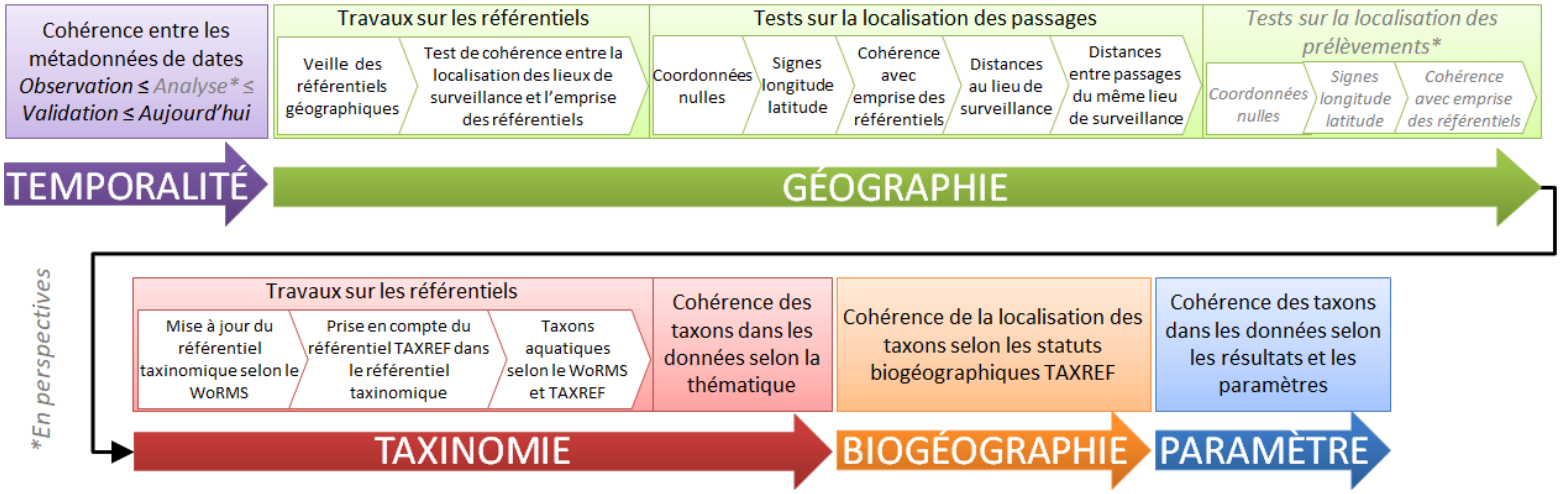


Figure 17 : Organisation actuelle de la procédure de qualité biogéographique des données taxinomiques mise en place.

5. Perspectives

Dans cette partie sont présentées différentes préconisations, inspirées de la mise en place de la

procédure de qualification biogéographique des données taxinomiques du SI Quadrige, et qui pourraient venir compléter la démarche qualité. Afin d'en faciliter le rapprochement avec la procédure, ces préconisations seront présentées de sorte à faire le lien avec les étapes de la procédure.

5.1 Standardisation de la donnée

Horton et al. (2021) présentent des recommandations pour la standardisation de l'Open Nomenclature (ON) (taxinomie) concernant l'appellation des taxons identifiés sur une image (*e.g. photographie d'un individu, utilisée pour identifier le taxon)*. Ces recommandations concernent les **qualificatifs** communément utilisés en cas d'identification imprécise (un niveau taxinomique moins précis que celui de l'espèce) tels que « *sp.* », « *spp.* », « *sp1* », « *stet.* » ou bien « *indet.* ». Leur démarche vise à établir une pratique commune de l'utilisation de ces qualificatifs pour assurer une interopérabilité, mais également sur la manière de stocker l'information expliquant la raison de l'imprécision de l'identification qu'il serait possible de consulter lors d'une éventuelle ré-identification ultérieure. L'identification de taxons sur image est une des techniques utilisées dans le cadre de l'acquisition des données taxinomiques du SI Quadrige. Cependant, actuellement le SI Quadrige ne permet pas de bancariser pas la raison de l'imprécision des identifications, ni de bancariser séparément les qualificatifs et les noms des taxons. Quant aux supports visuels d'identification, le SI Quadrige ne les relie pas aux données qui les concernent.

Ainsi, dans la perspective de se conformer à ces recommandations pour la standardisation de l'ON et de gagner en interopérabilité, le SI Quadrige devrait bancariser l'information permettant d'offrir la possibilité d'une ré-identification *a posteriori* des taxons sur image. Pour cela, il conviendrait :

- de bancariser le lien entre la donnée et le support multimédia ayant permis l'identification des taxons ;
- d'identifier ou de créer les éléments de l'architecture de la BD à utiliser pour bancariser les informations nécessaires.

5.2 Temporalité

Un des tests de la procédure consiste à vérifier la cohérence entre les différentes dates rattachées à une même donnée (*cf*. § 3.2.2). Ce test est inspiré d'un test du projet Kurator (#76) (Morris et al., 2018), qui cite la date de l'analyse (*cf*. Figure 9). A l'heure actuelle, cette date n'est pas bancarisée dans la BD Quadrige mais il s'agit d'une demande d'évolution. Une fois cette évolution effectuée, le test pourra être mis à jour afin d'intégrer la nouvelle information bancarisée que sera la date d'analyse. La date d'analyse devra être antérieure à la date de validation et postérieure à la date d'observation (*cf.* Figure 9). En outre, il pourrait être intéressant d'imposer le respect de cette chronologie dès la validation de la donnée, en bloquant la validation en cas d'anomalie.

5.3 Localisation

Lors de l'étude des emplacements des centroïdes de lieux de surveillance par rapport aux zonages des référentiels (*cf.* § 3.1.1), il a été mis en évidence que le référentiel des Zones Marines Quadrige n'était pas à jour sur certains points : entités manquantes en outre-mer (*e.g.* Mayotte), les délimitations du littoral non mises à jour suite à la parution de référentiels, plus précis, du trait de côte. Ainsi, la mise à jour du référentiel des Zones Marines Quadrige améliorerait la pertinence des résultats des tests de cohérence géographique car des lieux actuellement « hors zones » seraient réintégrés dans les zones marines.

Par la suite, lors de l'étude des distances entre passages et lieux de surveillance (*cf.* § 3.3.4.2), il a été mis en évidence que le SI Quadrige était dépourvu de l'information selon laquelle l'emprise géographique d'un passage a été héritée ou non du lieu de surveillance au moment de la saisie. En effet, les données indiquent seulement si un passage possède la même emprise géographique que son lieu de surveillance au moment de la consultation des données. Or, en cas de modification de l'emprise géographique d'un lieu de surveillance, cette modification n'est pas systématiquement effectuée sur les passages préexistants qui en héritent : ainsi, il est possible d'avoir une multitude de passages superposés avec une emprise géographique indiquée « Réelle », alors que celle-ci est à l'origine issue de celle du lieu de surveillance à leur création. Il est alors impossible de discerner les passages dont l'emprise géographique a été redéfinie au niveau du passage, et les passages dont l'emprise géographique est héritée d'une ancienne version du lieu de surveillance. Pour pallier ce problème, il conviendrait de mettre en place une historisation des

modifications faites sur les lieux.

De plus, les tests de localisation des données saisies n'ont été effectués qu'au niveau des passages. Certains tests pourraient être transposés pour s'appliquer à la géométrie indiquée au niveau du prélèvement, comme le test de cohérence de l'emplacement selon l'emprise des référentiels géographiques (*cf.* § 3.3.3). La vérification sur les prélèvements aurait lieu après celle sur les passages, afin d'éviter les doublons d'anomalies et fonctionnerait selon la même liste d'exceptions (lieu ou passages à écarter du processus).

5.4 Achèvement et automatisation de la procédure

Les tests décrits dans le présent rapport sont détaillés dans des fiches standards décrivant les outils informatiques nécessaires à leur exécution. La partie « traitement des résultats et règles de qualification » reste à définir pour certains tests. En effet, comme mentionné précédemment, c'est en analysant les cas d'anomalies que les causes sont identifiées et que les règles de résolution sont définies. Il reste donc à analyser les résultats de nombreux tests pour formaliser les règles de traitement à appliquer.

Enfin, pour optimiser la procédure et la rendre facilement reproductible périodiquement, il serait pertinent de développer un outil permettant de :

- Lancer automatiquement les tests sur des critères de sélection de données prédéfinis ;
- Générer automatiquement les fichiers d'anomalies pour les envoyer aux acteurs adéquats ;
- Intégrer en base les corrections et qualification selon les règles de traitement définies pour chaque test.

Cet outil pourrait s'intégrer aux plateformes existantes de qualification dites « automatique » et « experte » déjà en place ou à l'outil « AlerteAno ».

6. Bibliographie

Cellule Quadrige. (**2019**). La qualification : définition. Disponible sur : https://wwz.ifremer.fr/quadrige2_support/La-qualification-de-mes-donnees à « FicheQualif00_Definition » (Consulté le 19/07/2021)

Cellule Quadrige. (**2009**). Glossaire Quadrige2. Disponible sur https://wwz.ifremer.fr/quadrige2_support/Mon-support-Quadrige/Je-souhaite-ou-je-suis-une-formation/Documents-de-formation dans « Documents_formation » (Consulté le 19/07/2021)

Chapman AD. , Belbin L, Zermoglio PF, Wieczorek J, Morris PJ, Nicholls M, Rees ER, Veiga AK, Thompson A, Saraiva AM, James SA, Gendreau C, Benson A, Schigel D. (**2020**). Developing Standards for Improved Data Quality and for Selecting Fit for Use Biodiversity Data. Biodiversity Information Science and Standards 4: e50889. https://doi.org/10.3897/biss.4.50889

Gargominy, O. , Tercerie, S., Régnier, C., Ramage, T., Dupont, P., Daszkiewicz, P. & Poncet, L. (**2020**). TAXREF v14, référentiel taxonomique pour la France : méthodologie, mise en œuvre et diffusion. Muséum national d'Histoire naturelle, Paris. Rapport UMS PatriNat (OFB-CNRS-MNHN)

Horton T, Marsh L, Bett BJ, Gates AR, Jones DOB, Benoist NMA, Pfeifer S, Simon-Lledó E, Durden JM, Vandepitte L and Appeltans W. (**2021**). Recommendations for the Standardisation of Open Taxonomic Nomenclature for Image-Based Identifications. Front. Mar. Sci. 8:620702. doi: 10.3389/fmars.2021.620702

Ifremer. (**2021**). Présentation de l'Institut sur le site de l'Ifremer. Disponible sur : https://wwz.ifremer.fr/L-institut (Consulté le 21/07/2021)

Ifremer. (**2020**). Quadrige et la surveillance. Disponible sur : https://quadrige.eaufrance.fr/Quadrige-et-la-surveillance (Consulté le 21/07/2021)

Ifremer. (**2017a**). VIGIES. Disponible sur : https://wwz.ifremer.fr/Recherche/Departements-scientifiques/Departement-Oceanographie-et-Dynamique-des-Ecosystemes/VIGIES (Consulté le 18/07/2021)

Ifremer. (**2017b**). Les missions de la cellule d'administration. Disponible sur : https://wwz.ifremer.fr/quadrige2_support/La-Cellule-Quadrige/Les-missions-de-la-cellule-d-administration (Consulté le 19/07/2021)

INPN. (**2021**). Définition « Taxon ». Disponible sur : https://inpn.mnhn.fr/informations/glossaire/liste/t.Glossaire à la lettre T (Consulté le 20/07/2021)

Morris P, Hanken J, Lowery D, Ludäscher B, Macklin J, McPhillips T, Wieczorek J, Zhang Q (**2018**). Kurator: Tools for Improving Fitness for Use of Biodiversity Data. Biodiversity Information Science and Standards 2: e26539. https://doi.org/10.3897/biss.2.26539

Neaud-Masson Nadine. (**2020**). Quadrige² : Manuel de saisie pour les programmes REPHY et REPHYTOX Version 4. Disponible sur : ODE/VIGIES/20-02. https://archimer.ifremer.fr/doc/00440/55200/ (Consulté le 18/08/2021)

TDWG. (**s.d.**). Darwin Core. Disponible sur : https://dwc.tdwg.org/terms/ (Consulté le 18.08/2021)