Quadrige² - Référentiel National de gestion des données de la surveillance littorale

Emilie GAUTHIER, Antoine HUGUET, Catherine BELIN, Didier CLAISSE, Magali DUVAL

Juillet 2014

QUALIFICATION DES DONNEES DU SYSTEME D'INFORMATION Quadrige²

Ifremer

SOMMAIRE

1. Glossaire

DCE	Directive Cadre sur l'Eau
DCSMM	Directive Cadre Stratégie pour le Milieu Marin
DDTM	Direction Départementale des Transports et de la Mer
DYNECO/VIGIES	Unité DYNamique de l'Environnement COtier / service Valorisation de l'Information pour la Gestion Intégrée Et la Surveillance
IFREMER	Institut Français de Recherche pour l'Exploitation de la Mer
LER	Laboratoire Environnement Ressources (Laboratoire Ifremer)
NODC	National Oceanographic Data Center (http://www.nodc.noaa.gov/)
R	Logiciel R (http://www.r-project.org/)
REMI	REseau de surveillance MIcrobiologique
REPHY	REseau de surveillance du PHYtoplancton et des PHYcotoxines
RNO	Réseau National d'Observation des contaminants
ROCCH	Réseau d'Observation des Contaminants Chimiques (suite du RNO)
SANDRE	Service d'Administration Nationale des Données et Référentiels sur l'Eau
SI	Système d'Information
SQL	Structured Query Language
Surval	Outil de diffusion des données Quadrige² (http://envlit.ifremer.fr/resultats/surval)

2. Contexte

2.1. Le Système d'Information Quadrige²

Le Système d'Information (SI) Quadrige² contient des données issues de la surveillance sanitaire, environnementale et des ressources aquacoles des eaux littorales, dans le cadre de réseaux locaux et nationaux. Q² contient notamment toutes les données acquises pour l'évaluation DCE de la qualité des masses d'eau littorales.

Cette base de données contient *principalement* des données :

- Hydrologiques : température, salinité, nutriments, et autres paramètres physicochimiques
- Microbiologiques : qualité microbiologique des coquillages
- Biogéochimiques : contaminants chimiques dans les sédiments, le biote et l'eau
- Biologiques : suivi de la mortalité des huîtres, diversité et toxines phytoplanctoniques, faune et flore benthique.

Ces données peuvent être des résultats élémentaires d'analyse ou d'identification, mais aussi des couches cartographiques, des photographies, ou des fichiers de mesure (fichiers issus d'appareils de mesure ou de logiciels produisant un grand nombre de mesures sur un même échantillon).

Toutes ces données sont publiques, et doivent être diffusées :

- Aux gestionnaires du milieu marin
- A l'ensemble de la communauté scientifique
- Au grand public.

Un niveau de qualité est donc indispensable pour que chaque type d'utilisateur sache ce qu'il peut faire ou non des données.

2.2. Historique de la qualification Quadrige²

L'outil Quadrige (depuis 1996) puis Quadrige² (depuis 2008) permet aux utilisateurs de cette base d'attribuer différents niveaux de qualité à leurs données au cours du temps : contrôle, puis validation, puis qualification. Le contrôle et la validation sont effectués en continu par les utilisateurs depuis la mise en service de Quadrige.

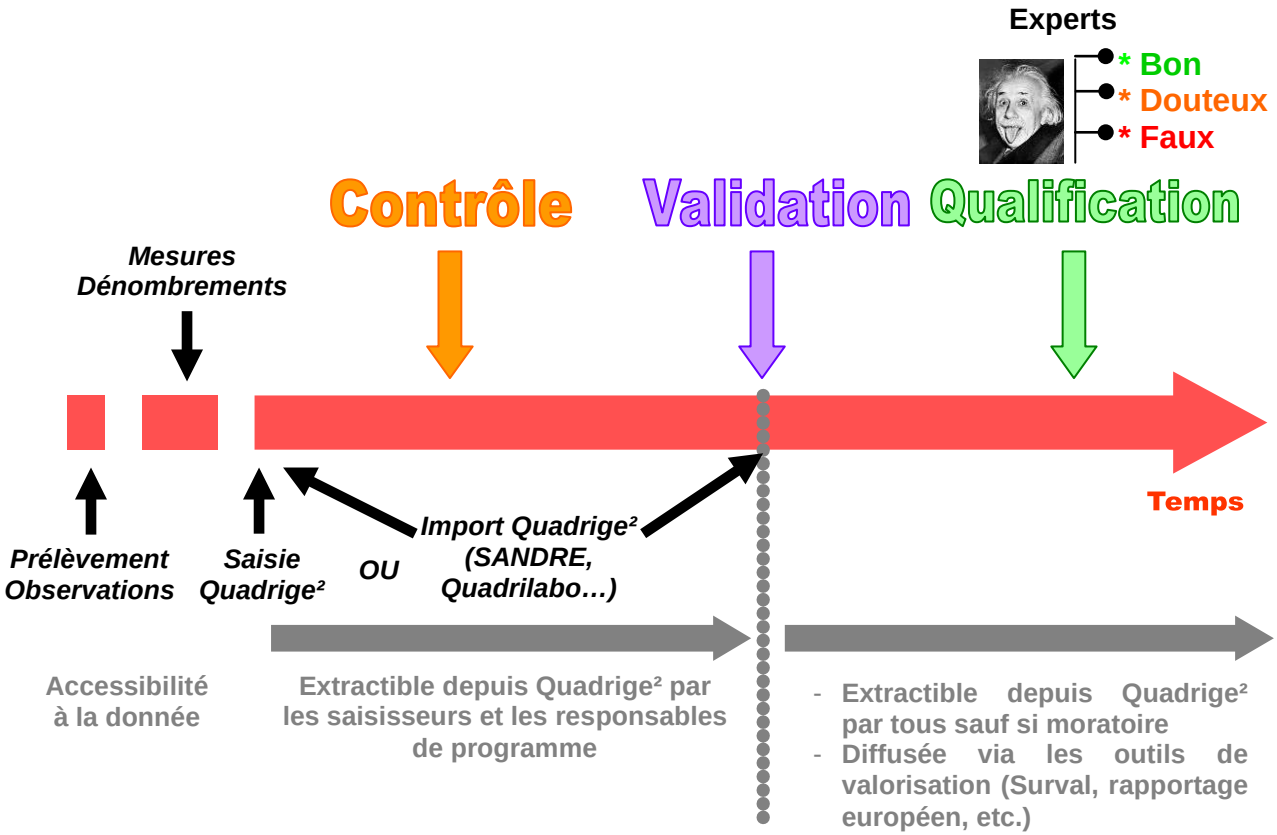
Les travaux de qualification des données de Q² ont été réalisés depuis de nombreuses années sous différentes formes :

- Années 2000 : échanges entre les coordinateurs du réseau REMI (microbiologie) et les laboratoires producteurs de données (LERs) pour corriger / qualifier les données douteuses. Les LERs ont fourni de nombreux retours sous forme de cahiers papier.
- 2004-2005 : travaux de qualification des données de contamination chimique (réseau RNO) : réalisation de programmes informatiques d'édition de graphiques automatisée, et qualification de presque 112 000 résultats.
- 2007 : utilisation de l'outil de qualification de l'application Quadrige² pour qualifier des données de contamination chimique (plus de 130 000 résultats qualifiés).
- 2009 : préparation d'un processus de qualification automatique des données de toxines phytoplanctoniques (réseau REPHY).
- 2009-2012 : mise en œuvre d'un processus de qualification « automatisé » pour qualifier les données hydrologiques et microbiologiques.
- 2010 : audit du processus de qualification « automatisé » par des « Data Manager » en recherche clinique : validation du processus et amélioration de la traçabilité.
- 2013 : qualification en routine des données hydrologiques et microbiologiques
- 2014 : qualification via le processus « automatisé » de l'ensemble des métadonnées chimie (réseau RNO/ROCCH), et qualification par expertise des résultats d'analyse chimique (même réseau).

La communauté Quadrige² a ainsi acquis au fil du temps une expérience intéressante en termes de qualification des données. La suite de ce document détaille l'organisation actuelle de l'ensemble des étapes d'attribution des niveaux de qualité aux données Quadrige².

3. Définition des étapes de contrôle, validation et qualification

Les données de Quadrige² ont un cycle de vie commun à toutes les thématiques :



Les données sont acquises sur le terrain et/ou en laboratoire, puis saisies dans la base de données Quadrige² via l'application du même nom. Le **contrôle** consiste à éditer les données saisies (résultats et métadonnées) pour vérifier leur **cohérence vis-à-vis du cahier de paillasse** (ou des fiches terrain).

Une fois ce contrôle effectué et les corrections éventuelles apportées, les données sont **validées** :

- Confirmation de la **validité technique** de la donnée (elle correspond bien au résultat de l'analyse réalisée)
- **Verrouillage** de la donnée (elle n'est plus modifiable, même par son saisisseur)
- **Diffusion** de la donnée : les données validées sont extractibles par tous les utilisateurs Quadrige² ayant accès à la base de données, et diffusées via Surval[1] (sauf si la donnée est protégée par un moratoire).

La qualification intervient après ce premier processus de vérification de la donnée. Elle correspond à :

- la **recherche des données douteuses** voire aberrantes du point de vue scientifique
- la **correction** des données lorsque cela est possible
- **l'attribution d'un niveau de qualification aux données** : ce niveau est
  - **bon** : la donnée a un sens scientifiquement parlant : son analyse sera pertinente,

[1] http://envlit.ifremer.fr/resultats/surval__1

- **douteux** : la donnée est peut-être fausse : sa prise en compte risque de biaiser l'analyse qui en sera faite
- **faux** : la donnée ne doit pas être intégrée aux analyses de données car elle est aberrante ou présente un problème connu (ex : mauvaise série analytique et impossible de la refaire).

Le niveau de qualification correspond au niveau de confiance des données. Il conditionne les modalités de diffusion de ces données (seules les données qualifiées « Bonnes » et « Douteuses » sont diffusées vie Surval[1]), et l'utilisation dans le cadre de traitement spécifique.

La qualification se décompose en 2 grandes étapes :

1) une **qualification « automatique »** qui consiste à rechercher des erreurs « grossières » et facilement identifiables. Exemples : erreur de paramètre, de support analytique, erreur de saisie de nombre (température de 100°C au lieu de 10°C par exemple) ou incohérences (données saisies sur le niveau « surface » avec une immersion de 20 m par exemple). Ces erreurs peuvent être décelées informatiquement en définissant des règles de contrôle simples (ex : heure = 00:00:00).
2) Une **qualification « experte »** qui consiste à mettre en évidence les données statistiquement aberrantes via des méthodes adaptées (séries temporelles, tests statistiques…). Par exemple, pour déceler qu'une salinité de l'eau est anormale pour la saison à laquelle elle a été mesurée, il faut replacer la donnée dans une série temporelle et tenir compte de la saisonnalité.

La qualification automatique aboutit à l'attribution (éventuellement provisoire) d'un niveau de qualité aux données (Bon, Douteux ou Faux). Seules les données qualifiées bonnes ou douteuses sont utilisées pour la qualification experte.

A l'issue de cette qualification experte, les données sont qualifiées de la façon suivante :

Niveau initial \ Niveau final		Qualification experte		
		BON	**DOUTEUX**	**FAUX**
Qualification automatique	**BON**	X	X	X
	DOUTEUX	Possible, mais rare (cas des données dont l'analyse statistique peut lever le doute)	X	X
	FAUX			

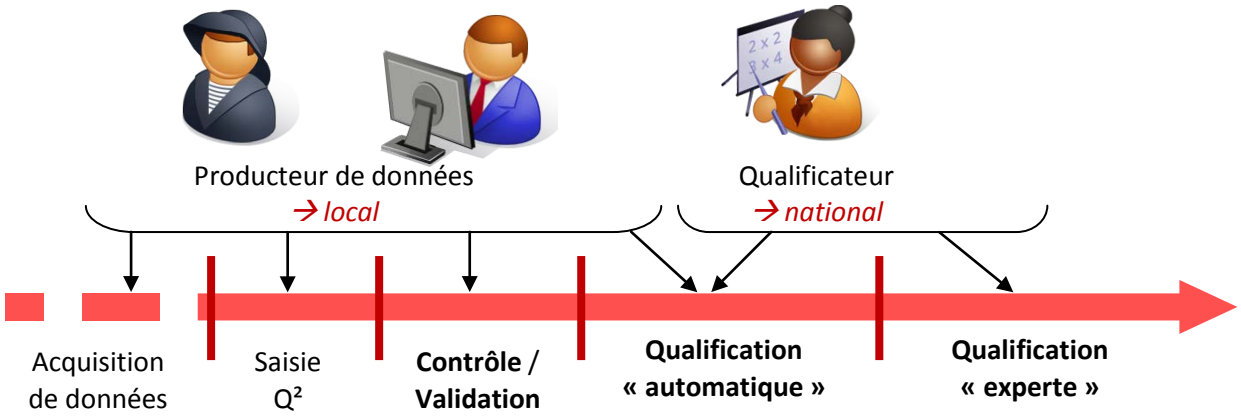
cas non concerné

4. Répartition des rôles

Le contrôle est de la responsabilité du saisisseur de la donnée et/ou des personnes ayant accès aux fiches terrain et de laboratoire (pour vérifier la cohérence des saisies). La validation est opérée par les mêmes opérateurs que le contrôle à la demande des coordinateurs de réseaux : les personnes de terrain / laboratoire sont ceux qui valident techniquement la donnée et confirment la véracité des saisies. La diffusion de la donnée, qui est sous la responsabilité des coordinateurs de réseaux, est déléguée implicitement aux laboratoires saisisseurs par cette validation (cf. 3).

La qualification est de la responsabilité d'experts thématiques, ayant les connaissances scientifiques nécessaires à l'interprétation des données. Les qualificateurs doivent être identifiés par les coordinateurs et/ou les maîtres d'ouvrage des réseaux de surveillance. L'Ifremer qualifie les données dont la chaîne d'acquisition est sous maîtrise d'œuvre interne. Les données acquises par les partenaires et/ou sous-traitants de l'Institut peuvent être qualifiés par les experts Ifremer s'ils estiment avoir suffisamment d'informations pour le faire.

L'échelle de contrôle / validation / qualification est d'abord locale, puis nationale :



Les producteurs de données sont des laboratoires Ifremer, des universitaires, des bureaux d'études, des associations, des services déconcentrés de l'Etat (DDTM), et toute autre structure qui acquiert de la donnée de surveillance environnementale.

Dans le cas des données qualifiées « en routine » actuellement à l'Ifremer, ce sont principalement les Laboratoires Environnement Ressources (LERs) qui sont concernés.

Les qualificateurs sont les coordinateurs de réseaux de surveillance, assistés par des experts thématiques de laboratoires de recherche. Ces experts sont reconnus internationalement.

5. Liste des niveaux de qualité avec définitions

Les niveaux de qualité utilisés dans la base de données Quadrige² font partie d'une liste standardisée[2] contenant 10 niveaux de qualité :

- 0 = « non qualifiée » (valeur par défaut au chargement de données en base)
- 1 = « bonne » (respect du protocole et aucune erreur détectée)
- 2 = « hors statistique »
- 3 = « douteuse » (non respect du protocole, non confiance dans la valeur enregistrée)
- 4 = « fausse » (erreur détectée à l'issue d'un test)
- 5 = « corrigée »
- 8 = « incomplète »
- 9 = « absente » (valeur manquante)

Dans Quadrige², seuls les niveaux 0 (Non qualifié), 1 (Bon), 3 (Douteux) et 4 (Faux) ont été retenus. La qualification d'une donnée est composée de trois informations :

- Niveau de qualité (cf. ci-dessus)
- Date de qualification
- Commentaire de qualification : obligatoire si le niveau est « Douteux » ou « Faux ».

[2] Cette liste de niveaux de qualité provient d'un standard reconnu au niveau international (standard NODC : http://www.nodc.noaa.gov/GTSPP/document/codetbls/gtsppcode.html#QUAL). Cette grille de niveaux de qualité est également utilisée dans le Système d'Information Halieutique (SIH).

Dans l'application Quadrige², on reconnaît les données qualifiées par un carré vert à côté de leur symbole (Figure 1).

Remarque : la couleur verte n'indique pas que le niveau de qualité est « Bon », mais uniquement qu'un niveau de qualité a été attribué.

Figure 1 : Exemple de symbologie de données qualifiées dans l'application Quadrige² : en violet, les données uniquement validées (Non qualifiées), et en vert les données qualifiées.

6. Qualification par donnée ou par jeu de données ?

Dans Quadrige², chaque donnée élémentaire est qualifiable :

- Métadonnées : couple lieu/date (nommé « passage »), prélèvement et échantillon. Chaque niveau de métadonnée peut être qualifié. Par exemple, les informations du passage (lieu, date, heure, hauteur d'eau sous le bateau…) peuvent être qualifiées « Bonnes », mais un souci lors de la conservation de l'échantillon peut le faire qualifier « Douteux ».
- Résultats : tous les résultats d'analyse ou d'observation, qu'ils soient de nature physique, chimique, biologique, ou même sous forme de fichier (couche cartographique, fichier de sonde de mesure, fichiers issus de capteurs automatiques, photos, etc.) sont qualifiés individuellement. Chaque résultat porte son propre niveau de qualité.

Le processus de qualification permet en revanche d'attribuer ces niveaux de qualité par lot de données (généralement lots annuels), sans avoir à saisir manuellement chaque niveau de qualité un par un.

La qualification experte consiste à analyser statistiquement les données par lots plus grands (séries temporelles), car l'analyse de tout l'historique permet plus facilement de discriminer les données qui sortent des valeurs classiques. Seuls les résultats sont traités en qualification experte.

7. Outils utilisés (procédure automatique ? Expertise ?)

Quadrige² permet de qualifier les données de différentes façons en fonction du besoin : qualification « ponctuelle » de quelques résultats, qualification d'un jeu de données homogène, qualification en routine de données intégrées régulièrement dans la base.

7.1. Qualification ponctuelle de quelques résultats

Exemples : on sait qu'un échantillon a été mal conservé et que les résultats de son analyse seront douteux, ou on a détecté un problème analytique et on sait que les résultats sont faux.

Dans ce cas, on peut qualifier ces quelques données via des interfaces dédiées dans l'application Quadrige² (Figure 2).

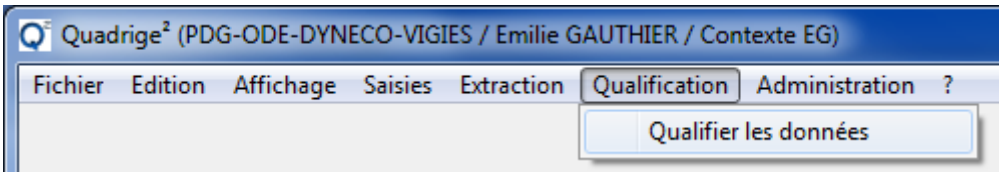
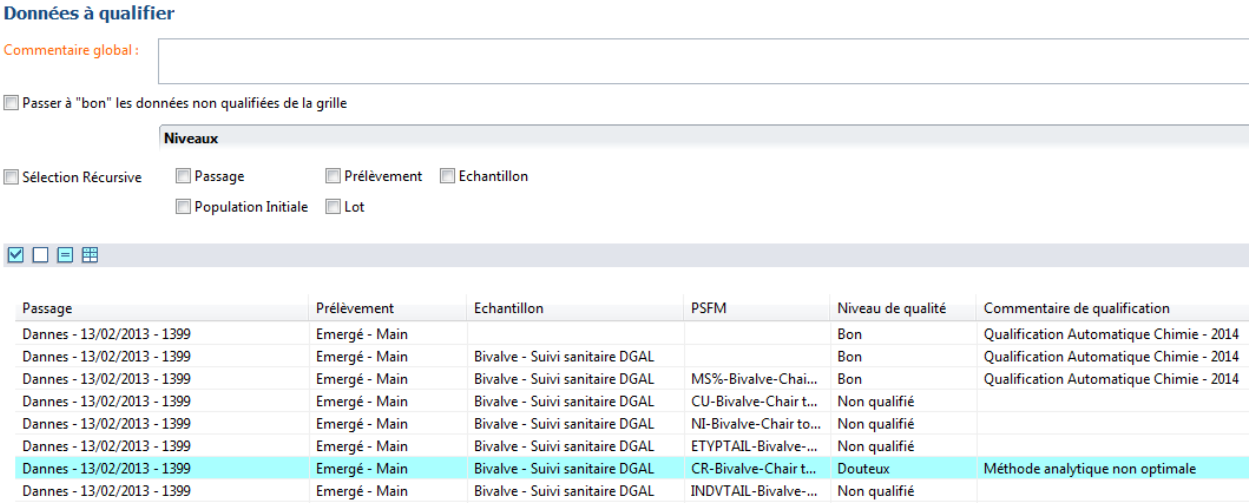


Figure 2 : Menu d'accès à l'outil de qualification de l'application Quadrige².

Cet outil permet de choisir des critères de requête pour sélectionner les données à qualifier, puis d'afficher ces données dans une grille permettant d'attribuer un niveau de qualité aux données (Figure 3).



Passage	Prélèvement	Echantillon	PSFM	Niveau de qualité	Commentaire de qualification
Dannes - 13/02/2013 - 1399	Emergé - Main			Bon	Qualification Automatique Chimie - 2014
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL		Bon	Qualification Automatique Chimie - 2014
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL	MS%-Bivalve-Chai...	Bon	Qualification Automatique Chimie - 2014
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL	CU-Bivalve-Chair t...	Non qualifié	
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL	NI-Bivalve-Chair to...	Non qualifié	
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL	ETYPTAIL-Bivalve-...	Non qualifié	
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL	CR-Bivalve-Chair t...	Douteux	Méthode analytique non optimale
Dannes - 13/02/2013 - 1399	Emergé - Main	Bivalve - Suivi sanitaire DGAL	INDVTAIL-Bivalve-...	Non qualifié	

Figure 3 : Grille de qualification de données dans l'outil de l'application Quadrige².

Pour accéder à ce menu, il faut disposer des droits de qualificateur dans Quadrige². Concrètement, ce sont les experts thématiques qui sont les qualificateurs (coordinateurs de réseaux de surveillance, qui gèrent les données de leur réseau dans Quadrige²). Ils réalisent l'opération de qualification généralement sur la demande d'un producteur de données.

7.2. Qualification d'un jeu de données homogène

7.2.1. Cas des « reprises » de données

Certains jeux de données sont intégrés dans Quadrige² sous forme de lots importants, migrés dans la base via des scripts informatiques. Selon la provenance de ces jeux de données, le fournisseur / producteur de la donnée peut définir le niveau de qualité de ses données.

Ex : données de surveillance des mammifères marins, ayant fait l'objet d'expertises précises par le producteur de données : ces données peuvent être intégrées dans Quadrige² avec un niveau de qualité « Bon ».

Ex : intégration de données historiques dont certaines informations essentielles sont manquantes (le protocole analytique par exemple) : intégration de ces données avec un niveau de qualité « Douteux ».

Le niveau de qualité est défini par le producteur de données en accord avec la cellule d'administration Quadrige² qui gère l'intégration des données. Dans les deux exemples ci-dessus, le niveau de qualité est déterminé dans le script informatique de migration dans Quadrige².

7.2.2. Cas des expertises sur jeux de données particuliers

Si un expert thématique travaille sur un jeu de données précis, que ce soit pour un rapport d'étude, une publication scientifique, ou toute autre analyse statistique, il peut transmettre les résultats de son expertise à la cellule d'administration Quadrige², et des niveaux de qualité peuvent être attribués aux données.

Dans ce cas, l'expert définit le périmètre du jeu de données avec la cellule d'administration Quadrige². Il définit les niveaux de qualité à attribuer, et la cellule exécute les requêtes de qualification des données (langage informatique : SQL).

Ex : qualification des données de zooplancton acquises dans le cadre du programme IGA (Impact des Grands Aménagements), travaux de thèse sur une thématique pour une période donnée, etc.

7.3. Qualification en routine des données intégrées régulièrement

Cette qualification concerne les données acquises dans le cadre des réseaux de surveillance pérennes comme le REPHY, le REMI, le ROCCH, etc[3]. Un processus de **qualification dit « automatique »** a été mis en place depuis 2009 pour qualifier chaque année les données des années précédentes.

Le principe est le suivant :

1) Les qualificateurs = experts thématiques définissent des « anomalies » à rechercher dans les données (ex : température hors des bornes [0 ;30°C]).
2) La cellule Quadrige² effectue une extraction des données de Quadrige² à qualifier (fichier au format .csv) et lance des programmes informatiques (développés sous le logiciel R) pour rechercher ces anomalies dans le jeu de données.
3) Les données sans anomalie sont qualifiées « Bon » dans la base immédiatement.
4) Les données avec anomalie (potentielle) sont envoyées aux producteurs de données (laboratoires côtiers) pour correction / qualification (format .csv).
5) Les retours des producteurs de données sont centralisés par la cellule d'administration Quadrige² qui les envoie au(x) qualificateur(s) pour validation (format .csv).
6) Le(s) qualificateur(s) renvoie le fichier .csv d'anomalies corrigées / qualifiées à la cellule Quadrige² qui intègre les corrections / qualifications en base de données via un script R (qui exécute des requêtes en SQL).

Suite à cette qualification « automatique », une qualification dite « experte » est réalisée. Elle consiste à analyser les résultats selon des modèles statistiques définis avec l'équipe de biostatisticiens du service DYNECO/VIGIES de l'Ifremer[4] (service auquel appartient la cellule d'administration Quadrige²).

Ces analyses statistiques sont réalisées via des programmes R, éditant des graphiques au format .pdf et des tableaux de données associées au format .csv. Le principe est le suivant :

1) Les qualificateurs et les biostatisticiens de DYNECO/VIGIES définissent les analyses statistiques à réaliser pour identifier les données potentiellement douteuses ou fausses.
2) La cellule Quadrige² effectue une extraction des données à qualifier (format .csv) et lance les programmes informatique R éditant les sorties graphiques et les tableaux de données à qualifier.

[3] Présentation des réseaux de surveillance : http://envlit.ifremer.fr/surveillance/presentation

[4] Cf.

3) Les qualificateurs analyses ces fichiers, et attribuent un niveau de qualité aux données (soit elles sont confirmées bonnes, soit elles sont qualifiées douteuses ou fausses avec un commentaire expliquant pourquoi). Les fichiers de données qualifiées (format .csv) sont renvoyés à la cellule Quadrige².
4) La cellule Quadrige² intègre les niveaux de qualité dans la base de données.

Un exemple d'édition graphique de qualification experte est présenté sur la Figure 4.

8. Les différents processus en fonction des thématiques

Pour les qualifications « automatique » et « experte », les règles de recherche d'anomalie diffèrent selon les données. Par exemple, une des anomalies recherchées consiste à vérifier la cohérence entre les engins de prélèvement utilisés, les niveaux de prélèvements (surface, fond…), et la profondeur à laquelle est réalisé le prélèvement. En hydrologie, la combinaison « Bouteille Hydrobios – Fond – 25m » est cohérente (ce n'est pas une anomalie), alors que cette même combinaison serait aberrante pour des données de qualité microbiologique des coquillages (on ne prélève pas des coquillages avec une bouteille hydrologique).

Pour des données biodiversité, il en irait de même : une espèce peut présenter une forte variation d'abondance au cours de l'année si elle prolifère dans certaines conditions (« blooms »), alors que d'autres espèces ont une abondance beaucoup plus stable.

La qualification des données est **à l'initiative des responsables thématiques des données (coordinateurs de réseaux de surveillance)**. Toute donnée intégrée dans Quadrige² pourrait donc être qualifiée selon un des processus mentionnés dans ce document, y compris les données de biodiversité (invertébrés benthiques, macroalgues, phanérogames, coraux, etc.).

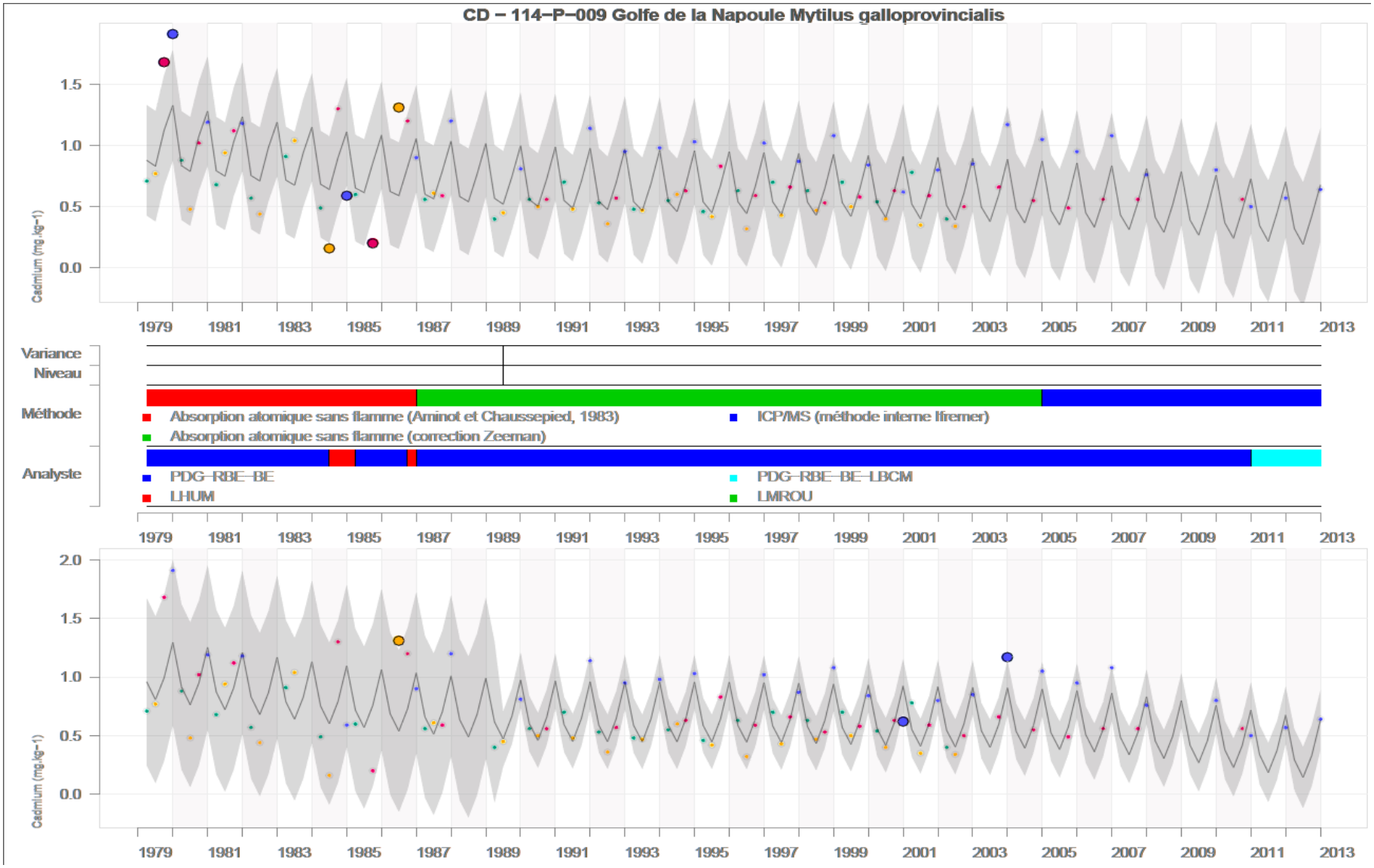


Figure 4 : Exemple de sortie graphique pour la qualification experte des données de DDTpp' (réseau RNO). Les points plus gros cerclés de noir sont les données identifiées comme "outliers" (données potentiellement douteuses ou fausses à vérifier).