SIMM
système
d'information sur
le milieu marin

SAR
SERVICE D'ADMINISTRATION
DES RÉFÉRENTIELS MARINS

Note technique : Principes généraux de qualification des données du SIMM

23 février 2021 Version : 1

SIMM

Historique du document

Date de création : 19 juin 2020

Dernière modification : 24 février 2021

Version	Date	Auteur	Relecture	Commentaires
0.1	19/06/2020	C.Rabévolo	S.Piel	
0.2	22/07/2020	C.Rabévolo	L.Coudercy, A. Huguet	Rajouts / Modifications
0.3	04/09/2020	C.Rabévolo	A.Rouyer	Corrections mineures
0.4	30/11/2020	C.Rabévolo	A.Rouyer, S.Piel	Ajouts et modifications suite aux remarques du GP Langage Commun
0.5	08/01/2021	C.Rabévolo	A.Rouyer, S.Piel	Ajout d'exemples (O&M, SIE, etc), corrections mineures
1	23/02/2021	C.Rabévolo	A.Rouyer, S.Piel	Ajouts et modifications suite aux remarques du GP Langage Commun

Contacts des membres du Service d'Administration des Référentiels (SAR) :

Nom	Téléphone	Mail	Organisme
Steven Piel	0298338745	steven.piel@ofb.gouv.fr	Office français de la biodiversité Direction surveillance, évaluation, données Pôle maritime de Brest 16 quai de la Douane 29229 Brest
Armelle Rouyer	0298224058	armelle.rouyer@ifremer.fr	Service SISMER – Ifremer centre Bretagne ZI de la pointe du Diable 29280 Plouzané
Clémence Rabévolo	0298224695	clemence.rabevolo@ifremer.fr	Service SISMER – Ifremer centre Bretagne ZI de la pointe du Diable 29280 Plouzané

Table des matières

1. Objet de ce document et contexte 3

2. Rôle des producteurs et banques du SIMM dans la qualification des données diffusées 4

2.1. Communication sur les procédures d'assurance et de contrôle qualité 4

2.2. Qualification des données du SIMM par les producteurs et banques 6

3. Communication des informations de qualification dans le cadre du SIMM 6

3.1. A l'échelle du lot de données 7

3.2. A l'échelle de la donnée 9

4. Conclusion 14

1. Objet de ce document et contexte

Le Système d'Information sur le Milieu Marin (SIMM) a pour objectif de faciliter le partage et la diffusion des données sur le milieu marin en offrant un point d'accès centralisé à des informations stables, tenues à jour et facilement compréhensibles. Ainsi, il a un rôle essentiel dans le rapportage aux politiques publiques (DCSMM, DCPEM, etc.).

Un des objectifs du SIMM cité dans le Schéma National des Données sur le Milieu Marin (SDNMM) est de permettre un accès au public à une **information fiable**. Afin de répondre à cette exigence, le SIMM doit mettre en place des processus qualité sur les données diffusées. Ainsi, un des buts métier identifié dans le dossier d'urbanisation du SIMM est de « **définir et mettre en œuvre la qualité des données** ». Ce but métier est décliné en la fonction F04 : « Qualifier les données et les interfaces ». Celle-ci s'articule autour de trois services :

Service	Processus
S09-Définir des critères de qualité des données et des flux	- Etablir des critères de qualité à appliquer aux données du SIMM par thématiques - Etablir des spécifications de services rendus pour les flux du SIMM
S10-Qualifier les données et les flux	- Mettre en œuvre les critères de qualité du SIMM sur les données - Mettre en œuvre les critères de qualité du SIMM sur les flux
S11-Produire des guides de bonnes pratiques	- Rédiger des guides de bonnes pratiques à destination des producteurs de données - Diffuser des guides de bonnes pratiques à destination des producteurs de données

Contrairement au Système d'Information sur l'Eau, il n'est pas prévu pour l'instant dans le SDNMM de mettre en place un service dédié à la définition de règles de bonne pratique pour l'assurance qualité comme c'est le cas d'Aquaref pour le SIE. C'est donc le **Service d'Administration des Référentiels** (SAR) qui a été mandaté[1] par le Groupe de Pilotage du Langage Commun pour travailler sur cette thématique.

23 systèmes d'information métiers, existants ou à créer, diffuseront leurs données via le SIMM à terme. Chacun de ces systèmes d'information métiers rassemble les données de plusieurs banques, **le SIMM diffusera donc des données issues d'un grand nombre de**

[1] Mandat du SAR concernant la qualification des données du SIMM

banques, et de producteurs. A l'échelle du SIMM il est en conséquence très complexe de faire des recommandations pour l'ensemble des étapes de qualification et des différentes thématiques.

Une journée de réflexion autour de la qualification des données du SIMM a été organisée le 30 janvier 2020. Il a été décidé à l'issue de cette réunion de travailler à la création du présent document proposant des principes généraux de qualification des données du SIMM.

Cette note technique est à caractère informatif et ne constitue pas un document de spécification. Au fur et à mesure du document, des idées d'outils, des recommandations, et des bonnes pratiques seront mises en avant sous la forme d'encarts orangés.

2. Rôle des producteurs et banques du SIMM dans la qualification des données diffusées

2.1. Communication sur les procédures d'assurance et de contrôle qualité

Il est important que les producteurs et les banques communiquent sur les procédures mises en place pour l'assurance et le contrôle qualité des données qu'ils diffusent.

L'assurance qualité est assurée par des actions en amont, sur toutes les étapes de la collecte jusqu'à la concentration des données. Cela peut concerner :

- des agréments ou habilitations d'organismes pour certains actes (mesures, prélèvements, …)
- des procédures de contrôle au moment de la production
- la mise en œuvre de guides de bonne pratique,
- l'application de normes garantissant la qualité,
- des éléments de contrôles de cohérence et de conformité à l'entrée des outils de concentration.

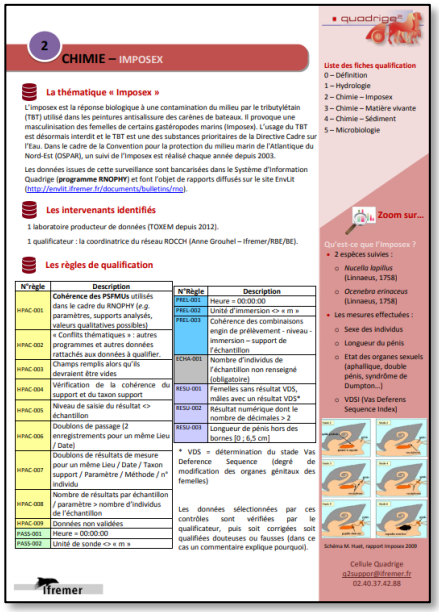
Communiquer sur ces éléments permet d'assurer aux utilisateurs un certain niveau d'information sur la qualité des données qu'ils souhaitent utiliser. Ces informations permettent une réutilisation des données la plus appropriée possible car les utilisateurs peuvent le faire en connaissance de cause.

La communication sur l'assurance qualité peut s'effectuer via des documents au format pdf identifiés détaillants les actions mises en place. Ces documents pourront avoir des identifiants pérennes de type URI [2] afin d'être intégrés dans les métadonnées. Il est également possible de pointer vers le site du producteur sur lesquels l'utilisateur pourra retrouver des informations concernant l'assurance qualité, comme les certificats [3] des normes ISO 9001, Core Trust Seal, etc. Toutefois, il faudra être vigilant à la pérennité de ces pages lorsqu'elles sont signalées dans des métadonnées si celles-ci ne sont pas identifiées par des URI.

Il est vivement recommandé aux producteurs et aux banques de communiquer également sur leurs protocoles de contrôles qualité.

Comme pour les informations relatives à l'assurance qualité, les critères de contrôles peuvent faire l'objet de documents explicatifs au format pdf identifiés par des URIs[2], ou de pages informatives sur le portail des producteurs ou banques.

A titre d'exemple, la banque de données Quadrige propose sur son portail une page « La qualification de mes données » diffusant des fiches synthétiques et thématiques concernant la qualification des données. *Exemple ci-contre d'une fiche spécifique au processus de qualification d'un type de données présent dans Quadrige.*


2 CHIMIE – IMPOSEX

La thématique « Imposex »

L'imposex est la réponse biologique à une contamination du milieu par le tributylétain (TBT) utilisé dans les peintures antisalissure des carènes de bateaux. Il provoque une masculinisation des femelles de certains gastéropodes marins (Imposex). L'usage du TBT est désormais interdit et le TBT est une des substances prioritaires de la Directive Cadre sur l'Eau. Dans le cadre de la Convention pour la protection du milieu marin de l'Atlantique du Nord-Est (OSPAR), un suivi de l'Imposex est réalisé chaque année depuis 2003.

Les données issues de cette surveillance sont bancarisées dans le Système d'Information Quadrige (**programme RNOPHY**) et font l'objet de rapports diffusés sur le site EnvLit (http://envlit.ifremer.fr/documents/bulletins/rno).

Les intervenants identifiés

1 laboratoire producteur de données (TOXEM depuis 2012).

1 qualificateur : la coordinatrice du réseau ROCCH (Anne Grouhel – Ifremer/RBE/BE).

Les règles de qualification

N°règle	Description
HPAC-001	**Cohérence des PSFMUs** utilisés dans le cadre du RNOPHY (e.g. paramètres, supports analysés, valeurs qualitatives possibles)
HPAC-002	« Conflits thématiques » : autres programmes et autres données rattachés aux données à qualifier.
HPAC-003	Champs remplis alors qu'ils devraient être vides
HPAC-004	Vérification de la cohérence du support et du taxon support
HPAC-005	Niveau de saisie du résultat <> échantillon
HPAC-006	Doublons de passage (2 enregistrements pour un même Lieu / Date)
HPAC-007	Doublons de résultats de mesure pour un même Lieu / Date / Taxon support / Paramètre / Méthode / n° individu
HPAC-008	Nombre de résultats par échantillon / paramètre > nombre d'individus de l'échantillon
HPAC-009	Données non validées
PASS-001	Heure = 00:00:00
PASS-002	Unité de sonde <> « m »

N°Règle	Description
PREL-001	Heure = 00:00:00
PREL-002	Unité d'immersion <> « m »
PREL-003	Cohérence des combinaisons engin de prélèvement - niveau - immersion – support de l'échantillon
ECHA-001	Nombre d'individus de l'échantillon non renseigné (obligatoire)
RESU-001	Femelles sans résultat VDS, mâles avec un résultat VDS*
RESU-002	Résultat numérique dont le nombre de décimales > 2
RESU-003	Longueur de pénis hors des bornes [0 ; 6,5 cm]

* VDS = détermination du stade Vas Deference Sequence (degré de modification des organes génitaux des femelles)

Les données sélectionnées par ces contrôles sont vérifiées par le qualificateur, puis soit corrigées soit qualifiées douteuses ou fausses (dans ce cas un commentaire explique pourquoi).

quadrige²

Liste des fiches qualification
0 – Définition
1 – Hydrologie
2 – Chimie – Imposex
3 – Chimie – Matière vivante
4 – Chimie – Sédiment
5 – Microbiologie

Zoom sur...

Qu'est-ce que l'imposex ?
- 2 espèces suivies :
  - *Nucella lapillus* (Linnaeus, 1758)
  - *Ocenebra erinaceus* (Linnaeus, 1758)
- Les mesures effectuées :
  - Sexe du individus
  - Longueur du pénis
  - Etat des organes sexuels (aphallique, double pénis, syndrôme de Dumpton...)
  - VDSI (Vas Deferens Sequence Index)

Schéma M. Huet, rapport imposex 2009

Cellule Quadrige
q2support@ifremer.fr
02.40.37.42.88

Ifremer

[2] Uniform Resource Identifier : https://fr.wikipedia.org/wiki/Uniform_Resource_Identifier

[3] Certificat ISO 9001 diffusé par le Sandre sur son portail

Globalement le SIMM recommande aux producteurs et banques de données de communiquer de la manière la plus transparente et exhaustive possible sur l'ensemble des actions mises en place concernant la qualité des données (et ce, à toutes les étapes : données brutes, données agrégées/produits, données qualifiées, données qualifiées à postériori).

2.2. Qualification des données du SIMM par les producteurs et banques

Le principe général n°5 du dossier d'urbanisation du SIMM « Validation et qualification » énonce ceci : *Les producteurs de données sont en charge de l'acquisition, de la validation et de la qualification thématique des données. Les banques mettent à disposition des outils de contrôle de conformité et assurent une qualification selon des règles métiers.*

I**l n'est donc pas du ressort de l'équipe du SIMM de procéder à la qualification des données**. Le SIMM ne donne pas de préconisations concernant les moyens de contrôle de la qualité des données. La qualité est définie vis-à-vis des besoins des missions de service public qui président à la collecte des données. **La qualification est effectuée par les producteurs et les banques selon leurs procédures habituelles.** Toutefois, le SIMM veille à ce que les utilisateurs puissent avoir accès aux informations de qualification nécessaires à la bonne réutilisation des données diffusées.

Ainsi, le SIMM proposera des moyens de communication des informations de qualification des données qui soient facilement compréhensibles par les utilisateurs, et accompagnera les producteurs et banques dans la diffusion de ces informations.

3. Communication des informations de qualification dans le cadre du SIMM

La réflexion sur les moyens de communication des informations de qualification des données dans le cadre du SIMM se fait à deux échelles : au niveau du lot de données (jeu de données plus ou moins homogènes) et au niveau de la donnée (enregistrement par enregistrement). La suite de ce document présente les différents outils envisageables pour assurer une bonne diffusion des informations de qualité à ces différents niveaux.

3.1. A l'échelle du lot de données

Les producteurs et banques de données renseignent les fiches de métadonnées liées aux lots de données de la manière la plus complète possible. Cela représente un premier niveau indispensable aux utilisateurs pour avoir une idée du degré de qualité du lot de données.

Le principe général n°6 du dossier d'urbanisation du SIMM « - Description des données - » énonce ceci : *Toute donnée est associée à des métadonnées. Ces dernières peuvent se trouver au niveau des banques, des lots de données ou sur la donnée elle-même. Elles comportent une indication de la qualité de la donnée*. **Toute donnée n'est ainsi diffusable que si elle est accompagnée de ses métadonnées**.

Les données du SIMM sont actuellement regroupées et diffusées via son catalogue qui utilise l'API Sextant. Le catalogue diffuse les fiches de métadonnées en conformité avec la norme ISO 19115, une norme internationale communément employée pour les informations géographiques. **Cette norme intègre plusieurs champs liés à la qualité des données :**

- « généralité sur la provenance » ou « généalogie » retrace l'historique des données en décrivant les principales phases de production, traitement et de qualification de la donnée, ce qui permet de disposer d'une information minimale sur la qualité.
- « événements » décrit les traitements opérés sur la donnée étape par étape.
- dates de création et de mise(s) à jour (versions)
- résolution spatiale et/ou temporelle
- etc.

Ces champs n'ont pas pour vocation de juger si les données sont de bonne ou de mauvaise qualité mais de permettre aux utilisateurs de vérifier que le niveau de qualité proposé est en adéquation avec leurs besoins.

Le SAR envisage de construire une « **charte de description des métadonnées** » listant le minimum de description des métadonnées nécessaire pour qu'un lot de données puisse être diffusé dans le cadre du SIMM. Cela assurera aux utilisateurs de bénéficier des informations suffisantes sur les données pour pouvoir les réutiliser. L'équipe du SIMM devra toutefois accompagner les producteurs et banques de données dans le renseignement de ces champs.

Pour certains jeux de données géographiques homogènes, il est possible d'aller plus loin dans le renseignement des métadonnées, avec des critères plus poussés sur la qualité des données.

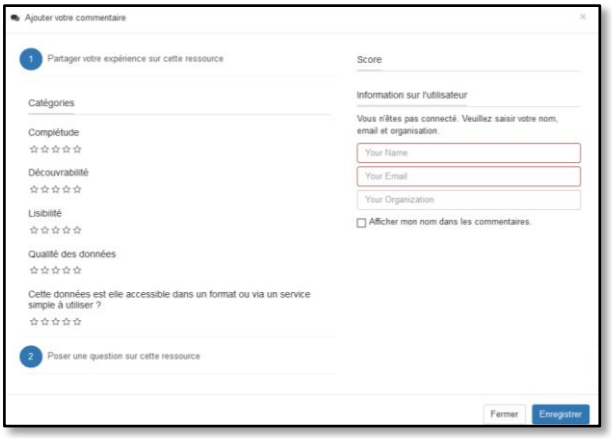
La **norme ISO 19157**[4] sur la qualité des données géographiques propose des critères standardisés pour qualifier un jeu de données géographiques et en communiquer les résultats. Il pourrait être possible de recommander l'utilisation de certains de ces critères, notamment en utilisant les outils développés par le GT-QuaDoGéo (CNIG), groupe de travail spécialisé du CNIG relatif à la qualité des données géographiques.

En complément des informations communiquées par les producteurs et les banques, il peut être envisagé de prévoir des moyens permettant d'avoir une qualification dite « externe » par les utilisateurs eux-mêmes, c'est-à-dire « une qualification par les pairs ». Le retour utilisateur à l'échelle d'un jeu de données sera surtout orienté sur la réutilisation des données et sur l'avis global de l'utilisateur (complétude, mise à jour, qualité globale, etc.). Il peut être intéressant pour le producteur d'avoir un retour sur la qualité d'un jeu à l'usage afin de l'améliorer, mais aussi sur ses réutilisations, qui peuvent être bien différentes de l'objectif initial de la création des données.

L'outil « retour utilisateur » de Géonetwork implémente le standard Geospatial User Feedback (GUF) de l'Open Geospatial Consortium (OGC). Il permet à l'utilisateur de noter un jeu de données selon plusieurs critères et de laisser un commentaire (précision sur la notation, réutilisation des données). Les critères par défaut sont : complétude, découvrabilité, lisibilité, qualité, accessibilité ; mais peuvent être modifiés. Les notes et avis sont visibles par tous les utilisateurs. Un retour au moyen d'un « workflow » vers le producteur est nécessaire afin d'assurer que les avis laissés lui parviennent, avec retour vers l'utilisateur si nécessaire.

[4] Pour en savoir plus sur la norme ISO 19157, le Cerema a créé une collection de fiches de décryptage de la norme : https://www.cerema.fr/fr/actualites/serie-fiches-cerema-qualifier-donnees-geographiques

Exemple de formulaire de retour utilisateur proposé dans Géonetwork avec l'outil issu de l'implémentation de la norme GUF. Il s'agit ci-contre du formulaire du catalogue des parcs nationaux français.



3.2. A l'échelle de la donnée

Les niveaux de qualité d'une donnée et les informations associées sont une réelle plus-value pour qu'un tiers sache s'il peut utiliser ou non une donnée au vu de l'usage qu'il souhaite en faire.

C'est pourquoi des informations concernant la qualité devraient être associées à chaque enregistrement. Tout comme pour les lots de données, **des informations de type « métadonnées » sont associées aux données, par exemple la date et l'heure de l'enregistrement, le lieu de la mesure, l'outil de mesure utilisé, etc.** Ce sont les informations essentielles sans lesquelles il est impossible de réutiliser les données convenablement.

Il est également recommandé **de renseigner pour chaque donnée des informations concernant le ou les niveaux de qualité (si différents critères) calculés par le producteur et/ou la banque** : la date de cette qualification, l'organisme ayant effectué cette qualification (banque, producteur), et le processus de qualification (explication de la qualification, lien vers les règles de qualification de la banque, contenu de l'historique, etc.). Si un processus de qualification a été effectué par le producteur, puis par la banque, il est recommandé de stocker les niveaux de qualité des deux contrôles afin d'éviter de perdre de l'information.

Si une donnée possède un **historique de qualification** (plusieurs niveaux de qualité attribués au cours du temps, issus par exemple d'une qualification automatique puis d'une qualification manuelle par un expert), **seul le niveau de qualité à l'instant t sera diffusé**

dans le cadre du SIMM. Il sera considéré comme le niveau le plus à jour aux vues des connaissances scientifiques et techniques et celui qui capitalise le plus d'expertise.

Lorsque des données ont été produites il y a plus d'une dizaine d'années et qu'elles n'ont pas encore été qualifiées, il faut admettre qu'elles ne le seront probablement jamais. En effet, il sera trop difficile d'obtenir toutes les informations nécessaires à leur qualification, qui n'ont pas forcément été bien renseignées ou ont pu être perdues.

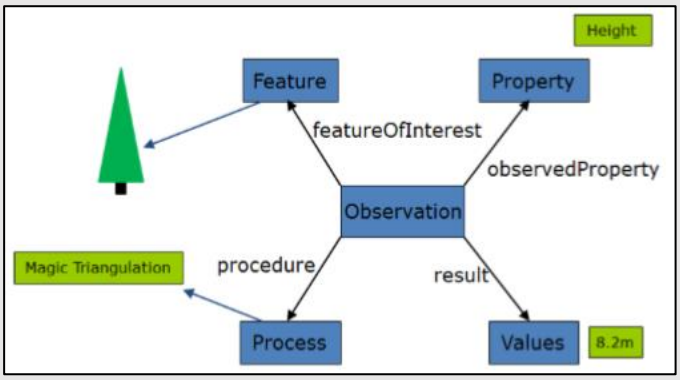
Le Service d'Administration des Référentiels (SAR) s'engage à produire des standards qui renseigneront sur la manière de structurer les données pour les diffuser dans le cadre du SIMM. Ces standards seront définis en fonction des besoins, selon les domaines métiers ou les rapportages aux politiques publiques. Le SAR pourra y inclure des champs de type « métadonnées » et de type « qualification » comme cités précédemment afin d'assurer la communication des informations sur la qualité des données à l'échelle de chaque enregistrement. **Les producteurs devront être accompagnés dans le renseignement de ces champs afin de s'assurer de leur complétude.** Le SAR pourra pour cela créer des tutoriels voire des formations à destination des producteurs, et proposer des outils pour faciliter l'utilisation des standards : test d'un fichier d'échange en ligne, générateur de base de données, etc.

Le SAR pourra utiliser, lors de la création des standards du SIMM, les standards internationaux, développés notamment par l'OGC, ISO et INSPIRE. En effet, ils intègrent des attributs sur la qualité des données, comme par exemple Observation and Measurement de l'OGC (ci-contre).

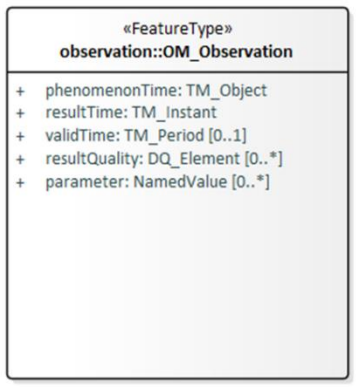
Exemple :

Dans le **standard Observation and Measurement**, l'observation est l'élément central du modèle et correspond à l'action de mesure d'une valeur.

Elle est reliée à différents éléments qui vont venir apporter les informations nécessaires à la compréhension des données (le résultat, l'objet d'intérêt, la méthode employée, la propriété mesurée, etc).



Représentation schématique d'un exemple d'utilisation de O&M



Un des attributs de la classe Observation est « **resultQuality** » et doit être renseigné lorsqu'il n'y a qu'une valeur (un résultat) ou que le jeu de données est homogène du point de vue de la qualité. Dans le cas de différences de qualité entre les résultats, alors le niveau de qualité doit être attribué pour chaque valeur au niveau de la classe « résultat ». Une dernière possibilité offerte par O&M est également de stocker les informations liées à la qualité dans l'attribut « Parameter » de l'observation.

Le nom de la personne ayant fait la mesure / l'observation ou la qualification des données ne doit pas apparaitre dans les données diffusées par le SIMM, même si cela peut donner une idée sur la qualité de la donnée (observateur expérimenté ou non). En effet, cette information n'est pas compréhensible pour un panel large d'utilisateurs, et est inexploitable dans le temps. De plus, cela ne respecte pas les principes du Règlement Général sur la Protection des Données (RGPD) et de la loi Informatique et Libertés.

Pour aller plus loin, il peut être envisagé de **standardiser les niveaux de qualité à renseigner dans les formats nécessaires à la diffusion des données** dans le cadre du SIMM. Ces valeurs standardisées rendent les niveaux de qualité comparables entre eux. Elles permettent également pour des utilisateurs non experts, qui ne sont pas en mesure d'effectuer une requalification des données, de savoir si une donnée peut être utilisée ou non.

Lors du brainstorming de janvier 2020, il est ressorti que la majorité des participants utilisent une sélection de tags inspirés des codes qualité de la NOAA[5]. Si SeaDataNet[6] propose une dizaine de tags qualité différents, les valeurs revenant le plus souvent dans les SI et banques participantes sont les suivantes :

- **non qualifié** : *la mesure n'a pas subi de processus de contrôle de sa qualité.*
- **bon** : *le protocole a été respecté (assurance qualité) et la mesure a été validée par les processus de contrôle qualité effectués par les producteurs ou la banque de données.*

[5] https://www.nodc.noaa.gov/GTSPP/document/codetbls/gtsppcodes/gtspp_qual.html

[6] Quality Control Standards for SeaDataNet (p°29) : https://www.seadatanet.org/content/download/596/file/SeaDataNet_QC_procedures_V2_(May_2010).pdf

- **douteux :** *non-respect du protocole (assurance qualité), non-confiance dans la valeur enregistrée (par exemple : valeur non habituelle)*. Ce niveau de qualité devrait idéalement être assorti d'un commentaire précisant pourquoi la donnée a été qualifiée comme étant douteuse. Cela permet éventuellement à l'utilisateur de conserver ou non les données dites douteuses selon l'usage qu'il compte en faire.
- **faux** : *valeur manifestement erronée, erreur détectée lors d'un contrôle qualité.*

Le SIMM pourra s'en inspirer, avec certains ajustements selon les domaines métiers, si nécessaire (par exemple, discrimination de plusieurs cas où la donnée peut être jugée comme douteuse).

Station	Date	Support	Fraction	Paramètre	Résultat de l'analyse	Unité	Qual.
01002228 - LA TERNOISE A TILLY CAPELLE (62)	07/12/2017	3 - Eau	23 - Eau brute	2911 - PBDE154	< 1.5E-4	µg/L	
01002228 - LA TERNOISE A TILLY CAPELLE (62)	07/12/2017	3 - Eau	23 - Eau brute	1600 - 4ClToluène	< 0.5	µg/L	
01002228 - LA TERNOISE A TILLY CAPELLE (62)	07/12/2017	3 - Eau	23 - Eau brute	1727 - 12DCEtn T	< 0.5	µg/L	
01002228 - LA TERNOISE A TILLY CAPELLE (62)	07/12/2017	3 - Eau	23 - Eau brute	1864 - Carbosulfa	< 0.1	µg/L	
01002228 - LA TERNOISE A TILLY CAPELLE (62)	07/12/2017	3 - Eau	23 - Eau brute	2982 - Difenacoum	< 0.02	µg/L	
01002228 - LA TERNOISE A TILLY CAPELLE (62)	07/12/2017	3 - Eau	23 - Eau brute	5743 - Dioxacarb	< 0.02	µg/L	
01002228 - LA TERNOISE A TILLY CAPELLE (62)	07/12/2017	3 - Eau	23 - Eau brute	1172 - Dicofol	< 0.001	µg/L	
CAPELLE (62)	07/12/2017	3 - Eau	23 - Eau brute	1152 - Déméton-S	< 0.01	µg/L	
CAPELLE (62)	07/12/2017	3 - Eau	23 - Eau brute	1159 - Dichlofent	< 0.005	µg/L	
CAPELLE (62)	07/12/2017	3 - Eau	23 - Eau brute	1360 - Dichloflua	< 0.005	µg/L	

0 – qualification non définissable
1 – correcte
2 – incorrecte
3 – incertaine
4 – non qualifié

Exemple d'utilisation des [valeurs de qualité dans le SI métier Naïades] du SIE. La couleur associée à chaque valeur informe d'un coup d'œil l'utilisateur sur la qualité de la donnée (ici vert = correct).

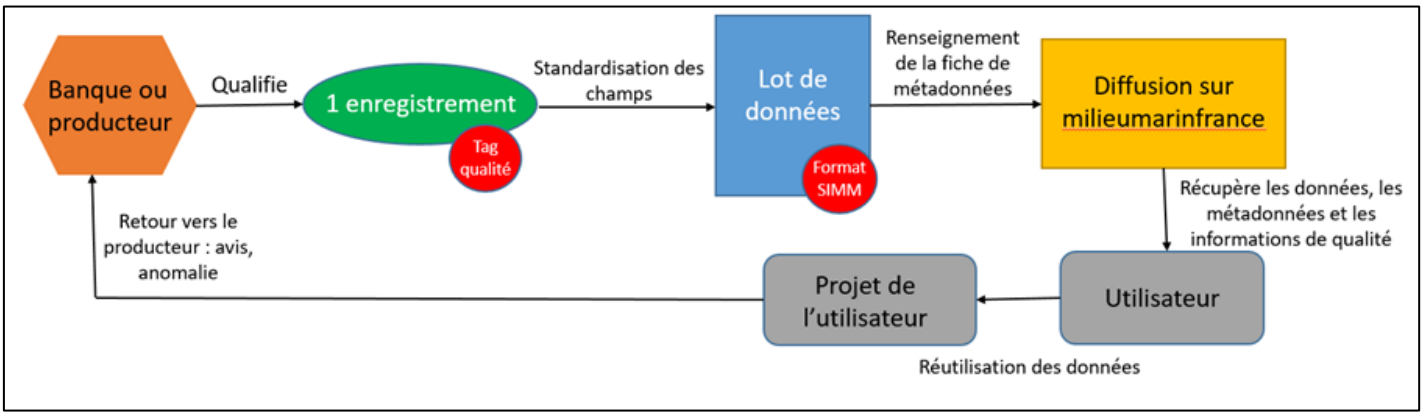
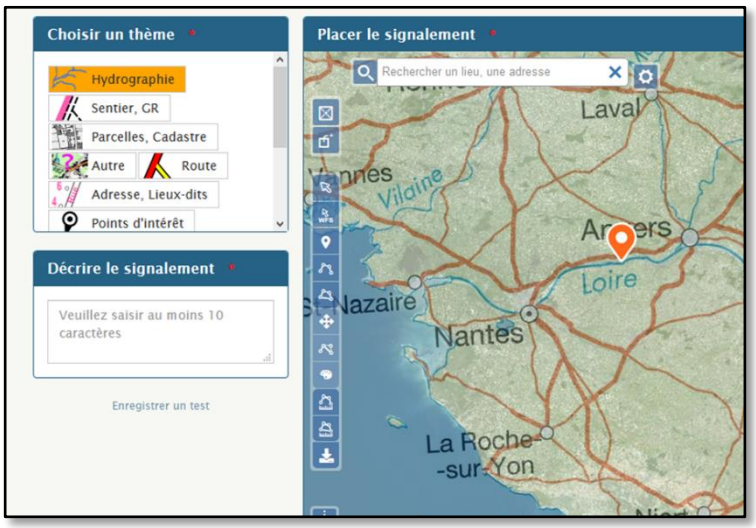


Schéma synthétique du cycle de qualification d'une donnée

Une réflexion pourra être menée sur la manière de communiquer les informations relatives à la qualité de la manière la plus complète possible pour les utilisateurs « experts », mais également d'une manière suffisamment accessible pour le « grand public ».

De la même manière qu'à l'échelle du lot de données, il est possible de travailler sur la qualification des données de l'utilisateur vers le producteur. En complément d'un outil de retour utilisateur au niveau de la fiche de métadonnée, il peut donc être mis en place un outil similaire au niveau de la donnée. Cela est notamment plus adapté pour signaler une anomalie sur un enregistrement spécifique. Le producteur pourra alors corriger cette anomalie et améliorer la qualité des données diffusées, ou bien modifier la valeur de qualité. Ce retour utilisateur par donnée assure un dernier « contrôle qualité » par les usagers eux-mêmes au moment de la réutilisation des données pour d'autres applications. Ainsi, plus des données seront utilisées, plus leur niveau de qualité augmentera grâce à la contribution des utilisateurs et les consolidations éventuelles du producteur.

Différentes plateformes proposent des outils de signalement d'anomalie comme l'espace collaboratif de l'IGN[7] ou l'outil Infonaut du Shom[8]. Le logiciel Géonetwork propose également ce genre d'outils pour effectuer un signalement « géoréférencé » sur une carte.



Exemple de l'outil de signalement de l'espace collaboratif de l'IGN.

[7] https://espacecollaboratif.ign.fr/

[8] https://data.shom.fr/infonaut/

4. Conclusion

La qualité des données dans le cadre du SIMM est à penser à plusieurs niveaux. Tout d'abord, l'équipe du SIMM s'assure que les producteurs et banques de données donnent suffisamment d'informations sur les données diffusées. Cela passe par le renseignement des fiches de métadonnées associées aux lots de données (généalogie du jeu, source, etc.) et au renseignement des champs de type métadonnée dans les formats d'échange (date du prélèvement, paramètres, etc.). C'est une première étape pour que les utilisateurs aient en leur possession suffisamment d'informations pour réutiliser les données à bon escient.

Ensuite, des protocoles de remontée et de diffusion des informations sur la qualification des données peuvent être mis en place afin d'informer les utilisateurs sur le niveau de qualité des données, et la manière dont a été faite la qualification. Pour cela, le SIMM proposera des solutions standardisées, notamment via l'utilisation de tag de qualité. Il accompagnera également les banques et producteurs dans la diffusion de ces informations. Il est notamment recommandé de communiquer le plus possible sur les protocoles qualité et les protocoles de qualification des données.

Enfin, des moyens de récolter les retours des utilisateurs seront mis en place afin d'améliorer la qualité des données via une qualification dite « externe ».