

Éléments de contexte pour le contrôle qualité

La connaissance de la qualité des données, en sécurisant l'utilisateur, incite davantage à leur réutilisation.

Ce décryptage de la norme ISO 19157 a pour vocation de donner un cadre méthodologique pour qualifier les données lors de leur diffusion.

L'essor des données ouvertes et géolocalisées et la profusion d'usages existants et à venir nous rend tous progressivement producteur et utilisateur de données géographiques.

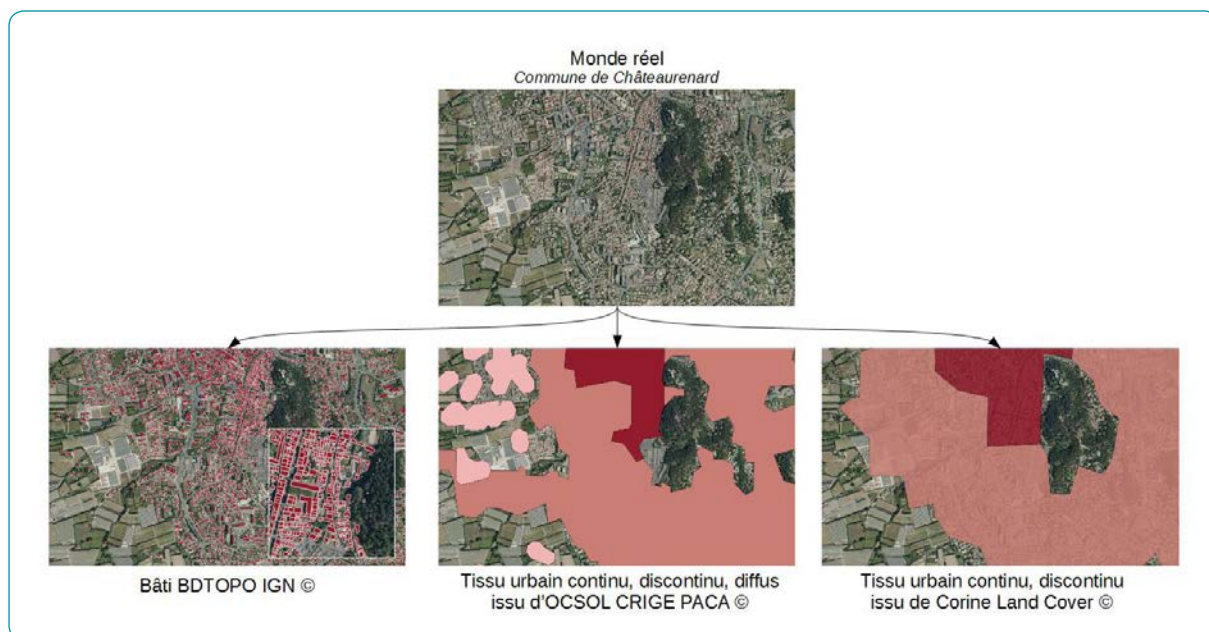
Les activités régaliennes ou les politiques publiques s'appuient sur de l'information maîtrisée où la qualité des données produites ou utilisées devient un entrant indispensable. Pour autant, tout le monde ne dispose pas des moyens des producteurs institutionnels de données et il paraît utile de fournir des recommandations et des méthodes plus adaptées au contexte de chacun, pour qualifier les données géographiques, communiquer sur les résultats obtenus voire savoir les interpréter. C'est l'objectif que s'est fixé le Cerema en proposant cette collection de fiches, à l'interface des productions et des usages.

Cette fiche rassemble l'ensemble des éléments communs aux contrôles des différents critères. Ce document vient compléter les fiches critères. En cas de nécessité, des éléments complémentaires peuvent se retrouver dans les fiches critères dès lors qu'ils dérogent aux règles communes ou qu'il y a nécessité de les préciser.



1. Terrain nominal versus monde réel

Un contrôle qualité de données géographiques s'effectue soit par rapport au monde réel, soit par rapport au terrain nominal.



Définitions

Le monde réel, comme son nom l'indique, représente la totalité des objets que l'on désire évaluer. Il faut donc disposer d'une donnée de référence qui recense la totalité des objets normalement présents dans l'emprise géographique concernée (tant pour un échantillon ou que pour une base entière).

Le terrain nominal se définit comme la restriction du monde réel au travers du filtre des spécifications. Avant de produire une base de données, celle-ci doit être spécifiée. Les spécifications définissent le contenu de la base de données et imposent à la fois des restrictions en termes de sélection et de contenu, ainsi que des exigences de qualité pour chaque classe d'objets de la base. C'est sur ces éléments que portent ensuite les contrôles qui sont effectués et dont il faut tenir compte.

Quand une base de données a fait l'objet de spécifications, le contrôle s'effectue nécessairement en référence à la description et aux simplifications proposées dans les spécifications.

Exemple de spécification : classe **RESERVOIR** de la **BD TOPO** ®

Les spécifications décrivent les données visées selon des caractéristiques a priori :

- définition : réservoir (eau, matières industrielles...) de plus de 10 m de diamètre ;
- topologie : simple ;
- genre : surface 3D ;
- sélection : tous les réservoirs de plus de 10 m de diamètre sont inclus sauf :
 - les réservoirs souterrains, les citernes sont dans la classe POINT_EAU,
 - les réservoirs d'eau non couverts (voir classe SURFACE_EAU, nature bassin).

Dans cet exemple, les spécifications précisent que les réservoirs retenus doivent avoir un diamètre supérieur à 10 m, doivent être positionnés en surface et être couverts. Cela exclut tous les autres.

Exemple de critère qualité : précision géométrique planimétrique de la BD TOPO®

Le tableau décrit la correspondance entre les sources de données et la valeur de la précision escomptée. Cette valeur apparaît ensuite sous forme d'un nombre réel dans l'attribut PREC_PLANI des objets décrits.

Ce tableau, issu des spécifications de la BD TOPO, nous permet d'apprécier la précision de position des objets en fonction de la source utilisée ou du mode de saisie employé.

Cet attribut PREC_PLANI commun à plusieurs classes n'est pas décrit pour chaque classe, et a obligatoirement l'une des valeurs de la troisième colonne du tableau.

Source de données	Précision	Traduction dans l'attribut « précision planimétrique » PREC_PLANI
Photogrammétrie, plan ou fichier métrique	0,5 à 1,5 m	1,5
Levé GPS dynamique, BD TOPO version antérieure, BD PARCELLAIRE recalée	1,5 à 2,5 m	2,5
Orthophotographie, plan ou fichier non métrique, levé terrain, BD PARCELLAIRE	2,5 à 5 m	5
Carte 1/25 000 (SCAN 25), image satellite	5 m à 10 m	10
BD CARTO, GEOROUTE	> à 10 m	30

Exemple de critère qualité : qualité sémantique de la BD TOPO® de l'IGN

La qualité sémantique¹ est la conformité des valeurs d'attributs des objets par rapport à la réalité.

Pour rendre compte de cette qualité sémantique, l'IGN calcule un indicateur (Confusions : taux de confusion) qui traduit en pourcentages la proportion d'erreurs sur une valeur d'attribut d'un objet (numéro de route par exemple).

Pour la BD TOPO® par exemple, l'attribut d'une classe ne doit pas présenter un taux d'erreur dû à la confusion entre deux classes (ex. tronçon de route/tronçon de voie ferrée) supérieur à 1 %.

Remarque : en l'absence de spécifications, la seule référence connue demeure le monde réel. Effectuer des contrôles par rapport à ce dernier peut nécessiter des réductions implicites ou explicites. On verra dans les chapitres suivants que l'existence d'une source de référence peut jouer le rôle de réduction du monde réel.

Même en cas d'existence de spécifications, cela n'exclut pas la possibilité d'effectuer le contrôle par rapport au monde réel pour conserver une vision plus complète.

Hormis les référentiels géographiques produits par des producteurs professionnels et les séries de données produites suivant des standards (CNIG ou Géostandards COVADIS par exemple), la plupart des données thématiques ou métiers sont produites sans spécification et nécessitent d'être contrôlées par rapport au monde réel.

¹ La terminologie qualité sémantique est celle retenue par l'IGN dans ses spécifications. Au regard de la norme ISO 19157, elle correspond au sous critère de qualité « justesse des attributs non quantitatifs ».

2. Contexte du Contrôle

Il faut disposer d'une base de données ou de documents de référence reconnus qui apportent une information de qualité théorique supérieure au lot de données à contrôler.

Pour l'exhaustivité par exemple, le jeu de contrôle doit recenser l'ensemble des données du monde réel, ou contenir au moins autant d'objets que les spécifications du jeu de données à contrôler le prévoient.

Pour la précision de position, le jeu de contrôle ou les méthodes mises en œuvre pour définir le monde réel ou nominal doivent être plus précises que la qualité attendue du jeu à contrôler.

Par exemple, l'arrêté du 16 septembre 2003² recommande un coefficient de confiance minimum de 2.

La difficulté consiste à trouver la donnée qui fait référence sur le sujet que l'on évalue. Ainsi, deux cas peuvent se présenter, du plus au moins favorable :

- on dispose d'une source servant de référence et reconnue comme telle ;
- il n'existe pas de source de contrôle pertinente.

2.1 On dispose d'une source servant de référence ou reconnue

La fiche du CNIG n° 82 de 2005 définit les référentiels géographiques et les données de références que nous rappelons ci-dessous.

Les données de référence sont clairement identifiées, définies et sont placées sous la responsabilité d'une structure publique clairement identifiée comme responsable de cette donnée. Les utilisateurs accordent aux données de référence un niveau de confiance très élevé, lié à la légitimité de l'organisme responsable de cette donnée. Elles offrent une couverture exhaustive du territoire.

La donnée de référence reste la propriété de l'organisme qui la génère et qui continue d'en assurer l'entretien. L'organisme qui produit et entretient la donnée de référence peut être distinct de celui qui l'intègre et de celui qui la diffuse : ces fonctions doivent être clairement distinguées. Les données de référence constituent la source à laquelle tout utilisateur se réfère pour valider ou obtenir une information.

Les référentiels géographiques, les référentiels « métier » et les données d'intérêt général sont constitués par des données de référence. En revanche, les autres catégories, données de contexte et données d'initiatives locales n'en font

pas partie, car elles ne respectent pas les exigences de référence. Certaines données métier peuvent néanmoins en faire partie.

Si l'on dispose d'une base de données ou d'un document de référence, la méthode est relativement simple et basée sur la comparaison entre source de contrôle et éléments à contrôler. On notera que la notion de source de référence ne signifie pas forcément une source unique mais peut être la combinaison de plusieurs entrants.

Cette comparaison, pour l'exhaustivité par exemple, revient à compter le nombre d'objets manquants et le nombre d'objets en trop. Il en est de même pour la précision thématique ou temporelle dès lors que les éléments à contrôler sont effectivement présents dans la source de contrôle.

Pour la précision de position, il importe que les éléments à contrôler soient présents, éventuellement dans une géométrie qui peut être différente. Les règles de mise en correspondance sont précisées dans la fiche n° 10 « Précision de position ».

2 Arrêté des classes de précision (pour plus de détails, se référer à la fiche « précision de position »).

2.2 Il n'existe pas de source de contrôle pertinente

Si aucune base de données de référence n'est connue ou accessible, la méthode est plus complexe, car il faut alors être en mesure d'évaluer les différents critères selon plusieurs angles plus ou moins subjectifs comme :

- la connaissance que l'on peut avoir du terrain ou d'une partie de celui-ci ;
- la connaissance du sujet traité ou de la base à évaluer ;
- la méthode utilisée pour la saisie des données (photogrammétrie, numérisation d'une carte ou d'un document existant, croisement de différentes bases...) qui permet d'apprécier un niveau de qualité sans pouvoir le fonder sur des éléments concrets et factuels. Savoir interpréter la généalogie d'un jeu de données demande une certaine expertise mais permet d'appréhender la qualité en général ;
- si la saisie des données s'est appuyée sur un référentiel connu ou plusieurs référentiels (référentiel géographique ou référentiel métier), l'analyse des spécifications du ou des référentiels utilisés est indispensable et aide à l'interprétation.

Cette situation regroupe aussi bien le cas où aucune source de contrôle n'est disponible que le cas où des sources partielles existent dont on ne maîtrise toutefois pas la pertinence.

Les méthodes préconisées en l'absence de source de référence :

■ Le contrôle terrain

- **exhaustivité, précision thématique, précision temporelle** : après avoir choisi l'échantillon qui doit être représentatif de la population à évaluer, un contrôle terrain permet de vérifier la présence effective des objets et tout attribut lié à la localisation, l'appellation, la caractérisation, la topographie...

A contrario, le contrôle terrain n'est pas une méthode optimale pour recenser les omissions ou évaluer toute caractéristique administrative, réglementaire...

- **précision de position** : le contrôle terrain est la méthode la plus efficace bien que nécessitant des moyens importants ;
- **cohérence logique** : sans objet.

Dans certains cas, le contrôle terrain peut être avantageusement remplacé par l'exploitation des orthophotographies à grande échelle dès lors que les objets à contrôler y sont facilement identifiables et que la prise de vue est suffisamment récente.

Par exemple, le recensement d'éoliennes peut être contrôlé sur des orthophotographies, car son interprétation ne permet pas de confusion. En revanche, contrôler la nature de certains bâtiments ou de zones d'occupation du sol est plus ambigu en dehors du terrain.

■ Le dire d'expert

Le dire d'expert (par la connaissance du terrain – ou d'une partie qui peut servir d'échantillon – ou de la thématique) : si une partie du territoire est facilement contrôlable par la connaissance qu'on en a sans avoir à se rendre sur le terrain et qu'elle est suffisamment importante pour être représentative de l'ensemble de la population, les éléments de qualité peuvent être estimés.



Attention : ces estimations ne sont pas fondées sur des mesures, comme cela est le cas pour le contrôle terrain. Elles portent obligatoirement une certaine part de subjectivité. Elles sont donc plus sujettes à être remises en cause .

C'est la raison pour laquelle ce type d'évaluation doit impérativement être commenté en même temps que la qualité estimée par cette méthode. Il faut ainsi être capable de juger l'incertitude qui accompagne la valeur fournie par l'expert et proposer un intervalle de confiance autour de cette valeur.

Ce type d'estimation est, certes subjectif, mais présente l'avantage d'être préférable à aucune qualification des données.

■ L'aspect visuel par analyse thématique

Les données géographiques offrent naturellement la possibilité d'être consultées visuellement. Cette caractéristique facilite la détection des erreurs possibles et les oublis. La sélection par analyse thématiques, en sélectionnant différents critères, rend aisée cette recherche visuelle.



La superposition du parcellaire des fichiers fonciers avec la photo aérienne permet de repérer des parcelles de formes carrées. Il s'agit des parcelles pour lesquelles seul le localisant est disponible.

■ L'exploitation des spécifications

Il s'agit, quand le jeu de données à contrôler a été produit en s'appuyant sur une base de données référentielle géographique, d'appuyer l'estimation des éléments de qualité sur les informations présentes dans les spécifications.

Par production sur un référentiel, on entend aussi bien la numérisation sur un référentiel image (Scan 25, BD Ortho) qu'une extraction d'un référentiel vecteur (sélection du réseau routier national de la BD Carthage).

L'intérêt de s'appuyer sur un référentiel est de conserver, par propagation, les éléments de qualité du référentiel origine.

Cette méthode est intéressante, car elle permet de qualifier tous les critères (hors cohérence logique) dès lors que des informations de qualité sont présentes dans les spécifications.

Dans tous les cas, la production obtenue ne peut être d'une qualité supérieure à celle prévue pour le référentiel géographique.

■ L'interprétation de la généalogie

La généalogie rassemble tout l'historique d'un jeu de données, depuis sa conception jusqu'au produit fini : levé initial, compléments thématiques, traitements géométriques et ou topologiques :

- le levé initial peut provenir de l'exploitation d'un référentiel géographique ou d'autres techniques, levé GPS, géocodage, numérisation, photogrammétrie, relevé lidar. Cette liste n'est pas exhaustive.
- cette méthode ne permet pas de fournir des chiffres quantifiés mais seulement une estimation des éléments de qualité.

Dans la norme ISO 19113, la « généalogie » était un critère de qualité à part entière. Elle ne figure plus dans la norme ISO 19157 comme un critère de qualité mais simplement comme une métadonnée de la norme ISO 19115. La raison principale est que l'on trouve rarement cette métadonnée bien renseignée. Sa description reste textuelle et ne permet qu'une estimation de la qualité. En revanche, il ne faut pas la sous-estimer car, dans certains cas, elle demeure la seule possibilité d'évaluer la qualité, aussi embryonnaire et incomplète soit-elle.

Ce qu'il faut retenir

La méthode de contrôle à employer dépend de trois facteurs :

- existence de spécifications ou non ;
- disponibilité de sources de contrôle pertinentes ou non ;
- critère qualité à évaluer.

Il n'est pas possible de définir de règles systématiques. Chaque jeu de données nécessite, en fonction de ce qu'il est important d'évaluer au vu des usages pressentis, une analyse préalable du contexte pour définir dans quelle configuration on se trouve :

- contrôle s'appuyant sur des données de référence ;
- contrôle terrain ;
- méthodes plus subjectives – à dire d'expert, analyse de la documentation – qui, malgré leurs imperfections, seront toujours préférables à l'absence d'information sur la qualité des données.

Série de fiches « Qualifier les données géographiques »

Fiche n° 01	Connaitre la qualité d'une donnée géographique fiabilise son utilisation
Fiche n° 02	Généralités sur la qualité des données géographiques
Fiche n° 03	Éléments de contexte pour le contrôle qualité
Fiche n° 04	Éléments statistiques
Fiche n° 05	Méthodes d'échantillonnage
Fiche n° 06	Modes de représentation
Fiche n° 07	Critère de cohérence logique
Fiche n° 08	Critère d'exhaustivité
Fiche n° 09	Critère de précision thématique
Fiche n° 10	Critère de précision de position
Fiche n° 11	Critère de qualité temporelle



Contributeurs ●●●

Fiche réalisée sous la coordination de Gilles Troispoux et Bernard Allouche (Cerema Territoires et ville)

Rédacteurs

Yves Bonin (Cerema Méditerranée), Arnauld Gallais (Cerema Ouest)

Contributeurs

Mathieu Rajerison, Silvio Rousic (Cerema Méditerranée)

Relecteurs

Benoît David (Mission information géographique MTES/CGDD), Stéphane Rolle (CRIGE PACA), Magali Carnino (DGAC), Stéphane Lévêque (Cerema Territoires et ville)

Maquettage

Cerema Territoires et ville
Service édition

Impression

Jouve
Mayenne

Date de publication 2017
ISSN : 2417-9701
2017/57



Contact ●●●

accueil.dtectv@cerema.fr

Boutique en ligne : catalogue.territoires-ville.cerema.fr

La collection « Connaissances » du Cerema

Cette collection présente l'état des connaissances à un moment donné et délivre de l'information sur un sujet, sans pour autant prétendre à l'exhaustivité. Elle offre une mise à jour des savoirs et pratiques professionnelles incluant de nouvelles approches techniques ou méthodologiques. Elle s'adresse à des professionnels souhaitant maintenir et approfondir leurs connaissances sur des domaines techniques en évolution constante. Les éléments présentés peuvent être considérés comme des préconisations, sans avoir le statut de références validées.

© 2017 - Cerema
La reproduction totale ou
partielle du document doit
être soumise à l'accord
préalable du Cerema.

Aménagement et développement des territoires - Ville et stratégies urbaines - Transition énergétique et climat - Environnement et ressources naturelles - Prévention des risques - Bien-être et réduction des nuisances - Mobilité et transport - Infrastructures de transport - Habitat et bâtiment