

Généralités sur la qualité des données géographiques

*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 première fiche, introductive, a pour ambition de dresser le panorama des principales notions relatives à la qualité des données géographiques en s'appuyant fortement sur la norme ISO 19157, référence internationale pour la qualification des données géographiques. C'est ainsi que sont présentés les principaux critères qualité, les mesures utilisables et les principes d'évaluation.

Tous ces éléments présentés de manière synthétique sont repris plus en détail dans les fiches de la collection « Qualifier les données géographiques ».

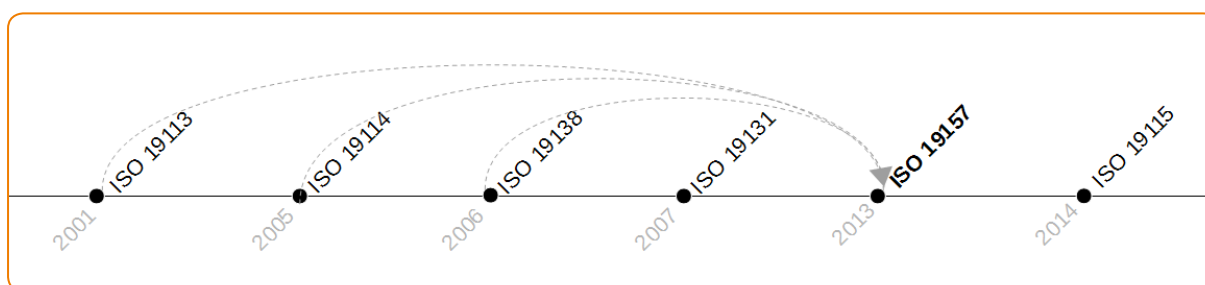


1. Les normes ISO de la qualité des données géographiques

La qualité des données géographiques est particulièrement importante en termes de prise de décision et de risques encourus lorsque la qualité des données utilisées est mal connue ou non maîtrisée. Chaque base de données devrait systématiquement faire l'objet d'une qualification avant d'être utilisée dans le cadre d'un usage afin d'en connaître les limites.

Afin de contrôler la qualité de leurs bases de données, les instituts nationaux responsables de la production des données géographiques de référence dans leur pays ont rédigé plusieurs normes permettant de décrire, contrôler et mesurer la qualité de leurs données. Ces normes, reconnues internationalement, demeurent la référence en matière de qualité des données.

Historiquement, trois normes ont été produites sur la qualité des données géographiques, comme le montre la frise temporelle ci-dessous :



ISO 19113 (2001)

Cette norme établissait les principes de description de la qualité des données géographiques. Elle définissait les différents composants ou critères destinés à décrire la qualité des données.

ISO 19114 (2005)

Norme destinée à proposer un cadre de méthode pour l'utilisation des différents critères définis dans la norme ISO 19113.

ISO 19138 (2006)

Elle définissait un ensemble de mesures de qualité des données destinées à l'évaluation des critères de la norme ISO 19113 et à la mise en place de rapports sur la qualité de données.

Ces trois normes ne sont plus en vigueur et ont été agrégées pour donner naissance à une nouvelle norme, la **norme ISO 19157**. Cette norme constitue la référence sur le sujet de la qualité des données géographiques. Une version traduite en français est disponible auprès de l'AFNOR. C'est une norme non obligatoire au regard du CEN et de l'AFNOR.

ISO 19157 (2013)

Elle établit les principes de description de la qualité des données géographiques, définit des critères destinés à décrire la qualité des données, spécifie des méthodes, décrit des procédures générales d'évaluation de la qualité des données géographiques, pose les principes de la description de la qualité des données dans des rapports. Elle définit également un ensemble de mesures de qualité des données destinées à l'évaluation des différents critères.

Elle s'applique aux producteurs de données fournissant des informations relatives à la qualité pour décrire et évaluer la façon dont un jeu de données répond à sa spécification de produit et aux utilisateurs cherchant à déterminer si des données géographiques spécifiques sont, ou non, de qualité suffisante pour une application particulière.

Cette norme sert de fil conducteur à la collection des fiches Cerema « qualifier les données géographiques » qui y font référence à de nombreuses reprises sans s'interdire toutefois de s'en éloigner quand les prescriptions normatives apparaissent trop déconnectées des besoins et usages concrets des utilisateurs.

Ainsi, la définition des critères qualité, la sélection des indicateurs pour les qualifier et la description des stratégies d'échantillonnage sont repris majoritairement depuis les recommandations de la norme ISO 19157.

Deux autres normes, sont étroitement liées à la qualité des données : la **norme ISO 19131** relative aux spécifications d'un lot de données et la **norme ISO 19115** sur les métadonnées.

ISO 19131 (2007)

Elle décrit les exigences relatives à la spécification de contenu informationnel géographique, en s'appuyant sur les concepts présentés dans les autres normes internationales ISO 19100. Elle apporte également une aide en matière de création de spécifications de contenu informationnel, de sorte qu'elles puissent être comprises et utilisées dans le cadre pour lequel elles ont été prévues.

ISO 19115 (2014)

Elle définit le schéma requis pour décrire des informations géographiques et des services au moyen de métadonnées. Elle fournit des informations concernant l'identification, l'étendue, la qualité, les aspects spatiaux et temporels, le contenu, la référence spatiale, la représentation des données, la distribution et d'autres propriétés des données géographiques numériques et des services.

Ces deux normes apportent des éclairages complémentaires à l'ISO 19157 sur les phases amont (Spécifications pour l'ISO 19131) et aval (Métadonnées dont le rapportage de la qualité pour l'ISO 19115) mais n'ont pas d'incidence directe sur l'évaluation de la qualité des données géographiques et sont par conséquent peu présentes dans la collection de fiches.

2. La norme ISO 19157

Les premières parties de la norme, les plus intéressantes pour une première prise de contact, s'appuient sur le triptyque « critère qualité ; mesure ; méthode d'évaluation » qui permet d'obtenir les résultats de l'évaluation qualité.

Ces trois notions sont reprises en détail dans les fiches ad hoc de la présente collection et une première présentation synthétique est proposée ci-dessous.

2.1 Les différents critères « qualité »

Le point d'entrée principal est la définition des différents critères qualité qui vont permettre de qualifier toutes les composantes d'un jeu de données.

Au nombre de cinq, ils sont présentés dans la norme comme des ensembles de sous-critères (on parle également d'éléments de qualité), chaque sous-critère apportant un éclairage plus ciblé sur un aspect particulier.

La présentation proposée ci-après déroge légèrement à l'ordre retenu par la norme et privilégie une logique basée conjointement sur la chronologie d'évaluation et l'importance des critères.

■ La cohérence logique

Ce premier critère regroupe tous les éléments de qualité qui vont attester qu'un lot de données est exploitable. L'accent est mis essentiellement sur les aspects techniques (la forme) contrairement

aux quatre autres critères qui traitent préférentiellement du contenu (le fond). Ce critère est généralement évalué en premier pour s'assurer que les données sont conformes aux attendus techniques avant d'évaluer les autres critères. Il regroupe quatre sous-critères :

- la cohérence conceptuelle ;
- la cohérence des domaines de valeurs ;
- la cohérence du format ;
- la cohérence topologique.

■ **l'exhaustivité**

Ce second critère, l'exhaustivité, est celui qui est généralement considéré comme le plus important en terme de qualité de contenu. D'ailleurs, c'est celui qu'on retrouve le plus fréquemment dans les règlements techniques de la Directive Inspire quand des exigences de qualité sont précisées. Il qualifie la présence ou l'absence de données qui sont traduites par les deux sous-critères :

- excédent ;
- omission .

■ **La précision thématique**

Ce troisième critère est la suite logique du critère d'exhaustivité. Après s'être assuré que les objets attendus sont effectivement présents dans le jeu de données, on s'attache à vérifier si les informations qu'ils portent sont exactes. On retrouve alors trois sous-critères qui tiennent compte des différentes formes possibles de confusion entre objets et de la typologie des informations portées :

- justesse du classement ;
- justesse des attributs non quantitatifs ;
- précision des attributs quantitatifs ;

▷ **La précision de position**

Ce quatrième critère évalue non plus la qualité du contenu comme les critères précédents mais la justesse de localisation. Bien que présenté en quatrième position, c'est un critère qui peut s'avérer fondamental pour certains types de données ou d'usage. Il est décomposé en trois sous-critères :

- précision absolue ou externe ;
- précision relative ou interne ;
- précision de position de données matricielles.

Les deux premiers sous-critères s'adressent aux données sous forme vectorielle. Le troisième sous-critère est identique, dans sa finalité, au sous-critère de précision absolue mais concerne plus spécifiquement les données sous forme matricielle (grille, image, modèle numérique...), ce qui induit des méthodes d'évaluation et des indicateurs différents.

■ **La qualité temporelle**

Ce dernier critère présente moins d'enjeu car il ne concerne que les jeux de données comportant des informations de datation, ce qui n'est pas une généralité. Il ne faut pas le confondre avec l'actualité des données qui, elle, relève des métadonnées générales. On retrouve trois sous-critères :

- exactitude de la mesure temporelle ;
- cohérence temporelle ;
- validité temporelle.

■ **Utilisabilité**

Il existe un sixième critère dans la norme ISO 19157 qui n'est pas traduit sous forme de mesure ou d'indicateur car fortement dépendant de l'usage qui est fait des données géographiques. Ce critère d'utilisabilité (usability dans la version anglaise) est fonction des exigences de chaque utilisateur qui ne peuvent être décrites en utilisant les éléments de qualité listés ci-dessus. Ce critère ne fait pas l'objet d'une fiche spécifique. En revanche, cette notion d'adéquation au besoin se retrouve en filigrane dans plusieurs fiches et est reprise sous une forme différente dans le §3 de ce document.

2.2 Les mesures

Pour chaque sous-critère, la norme propose une série de mesures ou d'indicateurs qui permettent de l'évaluer. Ces mesures dépendent de la nature du sous-critère, comme, par exemple, la présence ou non d'un élément, le caractère qualitatif ou quantitatif d'un attribut, la précision absolue, etc.

Il est fréquent de trouver plusieurs mesures possibles pour un même sous-critère :

- **indicateur d'erreur** (opérateur booléen : la valeur est juste ou fausse) ;

- **nombre d'erreurs** : nombre total d'erreurs mesurées sur l'échantillon ;
- **taux d'erreur** : pourcentage d'erreurs relevées sur l'échantillon évalué ;
- **moyenne** ;
- **écart type**.

D'autres mesures sont proposées mais d'utilisation plus marginale et ne feront pas l'objet d'approfondissement.

De manière générale, dès qu'il s'agira de comptabiliser des nombres d'erreurs, on privilégiera les indicateurs sous forme de taux aux dénombrements. Leur lecture et leur comparaison entre jeux de données est plus aisée.

2.3 Les méthodes d'évaluation

Indépendamment du critère de qualité et de la mesure retenue, la troisième composante essentielle est la méthode d'évaluation utilisée pour obtenir un résultat. Les méthodes possibles dépendent très sensiblement du contexte dans lequel s'inscrit la démarche de qualification des données et plus particulièrement de deux facteurs : l'existence ou non de spécifications, et la disponibilité ou non de sources de contrôle plus précises ou du moins pouvant faire fonction de référence.

De manière schématique, une méthode d'évaluation de la qualité des données décrit les procédures et les traitements appliqués aux données pour parvenir à un résultat de la mesure de la qualité.

Il n'est pas rare d'avoir, pour un même jeu de données, à mixer des méthodes d'évaluation différentes si on cherche à évaluer plusieurs critères ou si le jeu de données présente d'importantes hétérogénéités.

Remarque : Il est important de préciser la méthode d'évaluation de la qualité pour chaque mesure appliquée au moment du rapportage. Cela apporte un éclairage complémentaire aux résultats bruts du contrôle qualité.

La norme ISO 19137 ne détaille pas les méthodes possibles d'évaluation, ce qui est pourtant une information essentielle pour quiconque se lance dans une démarche de qualification des données. La collection de fiches Cerema s'attache à compenser ce manque en proposant différentes méthodes et en précisant dans quels contextes elles sont les plus pertinentes.

Par contre, la norme est très riche sur l'aspect échantillonnage, essentiel dès lors qu'il n'est pas possible ou, plus prosaïquement, économiquement raisonnable d'évaluer l'ensemble d'un jeu de données. Le bon choix de l'échantillon par sa taille et sa représentativité est une condition indispensable pour assurer la confiance dans les résultats qualité obtenus.

En conclusion, l'évaluation d'un élément de qualité – ou sous-critère – est décrite au moyen :

- d'une mesure : le type d'évaluation ;
- d'une méthode d'évaluation : la procédure utilisée pour évaluer la mesure ;
- d'un résultat : le produit de l'évaluation.

3. Les expressions de la qualité

En approfondissant un peu la notion de qualité, on constate qu'elle est plus complexe que la simple fourniture de chiffres. Plusieurs approches complémentaires existent. A la qualité interne, qui peut être assimilée aux critères techniques de la qualité présentés dans le chapitre précédent, on associe souvent la qualité externe ainsi que la qualité perçue. Le critère d'utilisabilité, présent dans la norme mais non approfondi, est à la frontière de ces deux approches.

3.1 La qualité interne

Elle exprime la qualité d'un lot de données produit au regard de nécessaires spécifications, rédigées au préalable pour décrire précisément ce qui doit être produit (cahier des charges). Ces spécifications détaillent la structure et le contenu de la base de données à produire en précisant la qualité exigée.

Les méthodes de contrôle de la qualité interne sont décrites dans la norme ISO 19157. La qualité interne ainsi évaluée correspond alors à l'écart entre les spécifications initiales et ce qui a été réellement produit.

3.2 La qualité externe

Elle se définit plutôt comme l'adéquation à un besoin exprimé. Dit différemment, c'est l'aptitude d'un jeu de données à satisfaire un usage donné. Il faut bien entendre ici un type d'usage générique comme une exploitation cartographique, un usage à but statistique, ou la constitution d'un observatoire.

En absence de spécifications qui, généralement, sont là pour traduire le besoin par rapport aux usages envisagés, il n'est pas possible d'évaluer la conformité d'un jeu de données. Qualifier revient par conséquent à donner une liste de résultats bruts sans statuer sur leur valeur et leur pertinence. Pour autant, tout professionnel de la géomatique est en capacité de juger, certes de manière subjective, si un jeu de données l'intéresse ou non.

Cette notion est encore peu utilisée et très peu documentée, la difficulté venant sans doute de la complexité à traduire scientifiquement chaque famille d'usage en exigences de qualité et sortir des approches uniquement subjectives. C'est une thématique qui relève encore du domaine de la recherche.

Pour autant, et c'est un des aspects novateurs de cette collection de fiches, une première appréciation de la qualité externe est apportée au travers des méthodes de représentation simplifiée ou synthétique de la qualité présentées dans la fiche éponyme et qui s'affranchissent de l'existence de spécifications.

3.3 La qualité perçue pour un usage spécifique

Elle exprime la perception d'un utilisateur pour un besoin spécifique. Cette approche, plutôt marginale il y a encore une quinzaine d'années, a été démocratisée avec l'essor de l'Internet et la multiplication des systèmes de notation par les communautés d'internautes. Elle s'est développée de manière exponentielle ces dernières années avec l'ubérisation de la société où chacun d'entre nous devient potentiellement « notateur » ou « noté ». Il n'y a pas de raison que l'information géographique échappe à cette évolution et pourquoi ne pas imaginer à terme, sur les géoportails, une notation des jeux de données résultant de l'agrégation d'avis d'utilisateurs, qu'ils soient particuliers ou professionnels.

La norme ISO 19157 n'aborde pas ces questions et ces dernières ne sont pas développées plus avant dans la collection de fiches Cerema. C'est un sujet qui dépasse très largement la géomatique car relevant plutôt du champ de la sociologie ou de l'anthropologie.

Ce qu'il faut retenir

La norme ISO 19157 fédère et réunit les éléments normatifs existants sur la qualité des données dans un seul document auto-porteur.

Tout ce qu'il est utile de connaître pour qualifier un jeu de données s'y trouve. Il en est ainsi de la description des critères de qualité et de leurs sous-critères, des mesures utilisables et des méthodes d'évaluation, autant d'éléments qui sont repris en détail dans la collection de fiches Cerema « Qualifier les données géographiques ».

Pour autant, il ne faut pas oublier que la norme a été produite par des professionnels de la géomatique, pour leurs besoins propres en premier lieu. Elle ne répond donc pas à toutes les situations. La norme apporte des éléments méthodologiques intéressants qu'il convient d'aborder avec des approches plus empiriques que les qualités externe ou perçue, car ces dernières comprennent une part de subjectivité.

Le cas idéal qui consiste à disposer de spécifications et à évaluer la conformité de la production n'est pas une généralité et on se retrouve plus fréquemment à devoir juger de l'adéquation d'un jeu de données non documenté à un besoin spécifique.

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), Gilles Troispoux (Cerema Territoires et ville)

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
2017/56

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



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.

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