

*Qualifier les données géographiques
Un décryptage de la norme ISO 19157*

Modes de représentation

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 existant 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 propose des méthodes pour exprimer la qualité d'un jeu de données, en tenant compte, d'une part, du contexte de sa création et de son évaluation, d'autre part, du public à qui on s'adresse.

Plus que de nouvelles représentations, elle propose d'étudier dans quelles conditions les 3 méthodes évoquées sont les plus pertinentes :

- la fourniture des mesures de la qualité par écart moyen quadratique, taux d'excédent ou de déficit...
- la représentation simplifiée par « toile d'araignée » ;
- la représentation schématique (smiley, systèmes d'étoile, feux...).



1. Les méthodes d'expression de la qualité des données

Il est proposé dans la suite de s'intéresser à trois modes de représentation adaptés à la retranscription d'une évaluation de la qualité d'un lot de données.

Ces trois méthodes sont :

- la fourniture des mesures de la qualité ;
- la représentation simplifiée par « toile d'araignée » ;
- la représentation schématique.

1.1 Fourniture des mesures de qualité (forme complexe)

Cette méthode revient à fournir un ensemble de mesures, plus ou moins nombreuses en fonction des critères évalués, sous forme de chiffres correspondant aux résultats du contrôle qualité. L'expression de la qualité des données à travers un ensemble de mesures peut être :

- **simple** (par exemple la mesure de la précision de position exprimée par une valeur unique) ;
- **multiple** (par exemple l'exhaustivité exprimée via des taux de déficit et d'excédent) ;
- **complexe** (par exemple l'exactitude thématique exprimée sous forme matricielle).

L'expression de la qualité des données est dans ce cas particulièrement riche en s'adressant principalement à des spécialistes.

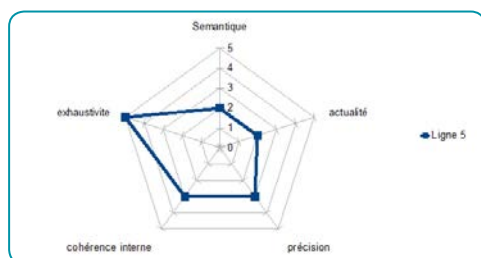
Son principal intérêt est d'être univoque.

1.2 Représentation simplifiée

Cette représentation revient à remplacer les différentes mesures de contrôle par un rendu normalisé. Quelle que soit la forme originelle de la mesure, sa restitution est rendue uniforme.

Le diagramme polaire (également appelé « toile d'araignée ») fait partie de cette famille des représentations simplifiées. Elle présente pour chaque critère de qualité une valeur sur une échelle.

Parmi les autres possibilités envisageables, il existe les histogrammes, les codes couleur, etc.



■ Les atouts

Une telle représentation apporte une lecture d'ensemble rapide de la qualité d'un jeu de données. Aussi, elle permet de comparer différents jeux de données simplement et graphiquement.

Enfin, dans l'hypothèse où l'on disposerait d'une « toile témoin » correspondant à la traduction des exigences de qualité nécessaires et suffisantes pour un usage donné, il est possible de comparer le résultat des mesures d'un jeu de données avec cette « toile témoin » servant de référence.

■ Les inconvénients

La grande difficulté de cette méthode est de définir les graduations de chaque échelle. S'il paraît assez simple de définir une gamme d'échelle (ou de notes) pour une unique variable quantitative telle que la précision de position¹, l'application semble plus complexe pour l'expression d'un critère tel que la cohérence interne. En effet, il convient de mixer des variables aussi diverses que le respect de la structure d'une table, le respect du système de référence, etc.

De plus, ce travail de graduation doit être conduit de telle sorte qu'une même valeur corresponde aux mêmes degrés de qualité sur chaque branche critère.

Le rendu par diagramme polaire peut être potentiellement utilisé (cf. infra) pour restituer la qualité d'un jeu de données par rapport au monde réel (absence de spécifications), pour quantifier le respect des exigences de qualité, ou pour évaluer la qualité par rapport à un usage donné. En fonction de l'objectif visé, une note sur une branche n'aura pas le même impact.

Remarque : Le diagramme polaire ne pouvant se suffire à lui-même, il doit être accompagné des éléments de contexte (informations sur le contrôle) ayant permis de le produire : date et organisme de contrôle, population de départ et taille de l'échantillon, les moyens de contrôle utilisés (contrôle terrain, dire d'expert, calculs statistiques), mesures de certains sous-critères (déficit, excédent versus exhaustivité).

1 Selon que l'on travaille à très grande échelle ou à l'échelle de territoires, la notion d'une « très bonne » précision de position ne sera pas la même.

1.3 Représentation synthétique

Il s'agit d'aller encore plus loin dans la simplification de l'information sur la qualité d'un jeu de données en réduisant l'ensemble des mesures à une seule information.

Les représentations possibles sont multiples :

- note (sur 5, sur 20...) ;
- smiley ;
- feu (vert, orange, rouge) ;
- étoiles (1 à 5 étoiles en fonction de la qualité du lot).

Concernant les méthodes graphiques, on veillera à ne pas dépasser 5 niveaux pour que l'exercice conserve son intérêt simplificateur.

■ Les atouts

Les atouts de cette représentation sont :

- la lecture immédiate ;
- la compréhension aisée du résultat.

■ Les inconvénients / difficultés

Cette simplification à l'extrême n'est pas sans contrepartie :

- la méthodologie de passage d'un ensemble de mesures à une note unique nécessite de définir des règles de construction univoques (un même lot de mesures doit toujours aboutir à la même note) ;
- indiquer comment interpréter une note nécessite de donner accès aux méthodes de construction via une explication sur la notation ;
- la notation peut s'avérer trop synthétique et de ce fait très réductrice.

Un autre intérêt de la représentation synthétique est de fournir une information sur l'adaptation du lot de données à des usages type (analyse statistique, cartographie, consultation, analyse spatiale, géomarketing, positionnement, navigation, etc.).

Exemple : un jeu de données régulièrement actualisé et complet mais de précision géométrique médiocre se verrait affecter une bonne note pour l'usage cartographie ou géomarketing mais une mauvaise pour l'usage analyse spatiale ou positionnement-navigation.

1.4 Méthodologie des contrôles

Quelle que soit la représentation employée pour restituer le bilan qualité, il est impératif de l'accompagner d'informations sur la méthode de contrôle utilisée (échantillon, tests, ...).

Comme pour tous les critères, il importe de commenter la note attribuée avec une description détaillée de la méthode mise en œuvre. En complément des éléments d'évaluation de la qualité temporelle, on rappellera systématiquement :

- la référence utilisée le cas échéant en tant que source de contrôle : producteur, date, spécifications, système de référence temporel, objectifs de précision temporelle éventuellement annoncés dans les spécifications, etc.
- la méthode utilisée dans le cas d'un contrôle par échantillonnage, la taille de l'échantillon, les effectifs de chaque classe de l'échantillon et la valeur de l'intervalle de confiance.

2. Quand utiliser ces méthodes ?

Une des questions à se poser est : dans quel cadre utiliser telle représentation plutôt qu'une autre ? Pour répondre à cette question, il faut définir son besoin : souhaite-t-on une évaluation absolue d'un lot de données (par rapport à des spécifications ou en l'absence de spécifications), une adéquation d'un lot à un usage particulier, etc. ?

Plusieurs situations sont à envisager :

- le jeu de données a été produit en référence à des spécifications éventuellement accompagnées d'exigence de qualité ;
- le jeu de données ne répond pas à des spécifications partagées (elles existent peut-être mais ne sont pas diffusées en même temps que le jeu

de données) mais il existe un « référentiel » de production ou d'exploitation produit par ailleurs (norme, standard...). C'est un cas que l'on trouvera régulièrement dans la diffusion de données open data ;

- le jeu de données ne fait référence à aucun document permettant d'en mesurer la qualité interne (donnée trop marginale pour avoir été standardisée, pas de communauté d'intérêt...) ni au respect de règles de saisie ou de sélection. C'est le cas de nombreux lots de données disponibles sur les portails (portail data.gouv.fr par exemple), ou présents dans le patrimoine des services (les spécifications ont peut-être existé mais ne sont plus connues).

Il convient également de s'interroger sur le contexte dans lequel s'inscrit l'analyse de la qualité :

- en tant que producteur ou commanditaire : évaluation de la conformité du lot de données ;
- en tant qu'utilisateur : pertinence du lot de données vis-à-vis des usages pressentis.

Pour le premier point de vue, l'approche sera différente selon les situations citées précédemment.

Pour l'approche utilisateur, la pertinence des données par rapport à l'usage importe davantage que la présence de spécifications. Par exemple, un lot BD Carto© de précision décimétrique peut être conforme aux spécifications de saisie et de qualité. Pour autant, il ne sera d'aucun intérêt pour une collectivité qui travaille à l'échelle du corps de rue.

2.1 Existence de spécifications

■ Présence d'exigences qualité

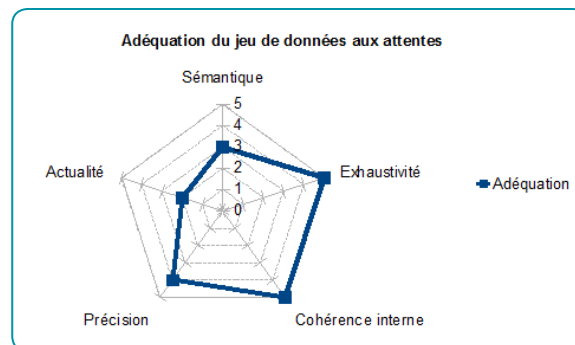
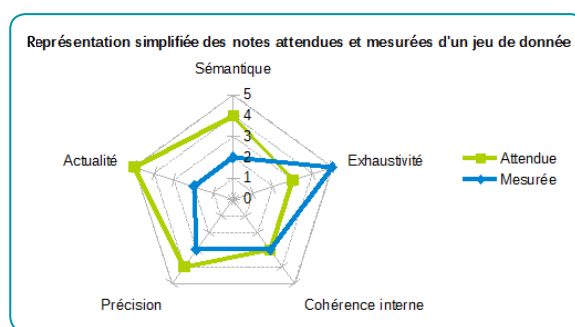
C'est certainement le cas le plus simple puisque l'on dispose, d'une part, d'une définition du contenu attendu, d'autre part, des objectifs qualitatifs qui ont été fixés. Ces derniers sont généralement établis en fonction des exploitations qui seront faites.

La représentation de la qualité peut alors prendre les 3 formes :

- la forme **complexe** qui consiste à lister l'ensemble des résultats des contrôles sous forme chiffrée avec la mise en perspective des résultats de mesures avec les exigences qualité ;
- la forme **simplifiée** avec la double représentation des exigences et des mesures, ou une forme comparative ;
- la forme **synthétique** avec deux résultats possibles (point de vue commanditaire) : lot conforme => meilleure note (feu vert par exemple), lot non conforme => moins bonne note (feu rouge par exemple).

La forme complexe est à privilégier dans une relation contractuelle (commanditaire-prestataire) ainsi que la forme synthétique car celle-ci donne efficacement l'information de conformité ou non.

La forme simplifiée (avec double représentation) est utile comme outil d'information aux utilisateurs (utilisations prévues ou autres usages) qui ne sont pas nécessairement intéressés par le détail du contrôle ni par le bilan contractuel mais pour qui la présentation des niveaux d'exigence qualité attendus et obtenus par critère a du sens.



■ Absence d'exigence qualité

La base de données est évaluée par rapport aux spécifications, mais il n'existe pas de critère ou de seuil permettant de définir la conformité ou non-conformité du jeu de données. Cette notion de conformité explicite ou implicite n'a donc pas lieu d'être.

L'existence de spécifications permet cependant de définir l'échantillon de contrôle. C'est, d'une certaine façon, le cas de la majorité des géostandards COVADIS (Commission interministérielle de validation des données pour l'information spatialisée).

La forme complexe est celle qui apporte naturellement le moins d'ambiguïté. Mais comme pour les autres cas, on peut considérer qu'elle ne présente un intérêt qu'en deuxième niveau de lecture pour ceux qui souhaitent un complément d'information.

La forme simplifiée semble la plus adaptée car elle permet de représenter les résultats des mesures sous forme graphique tout en maintenant le niveau de détail par critère. Cette représentation ne porte pas de jugement sur la conformité du lot.

La forme synthétique apportant une dose de subjectivité par rapport à une conformité théorique, elle n'est pas adaptée. On l'évitera dans une telle situation.

2.2 Absence de spécifications

Deux situations peuvent se présenter.

■ Pas de spécifications mais existence d'un cadre technique

L'existence de normes ou de standards pour certaines thématiques (comme des profils UML ou équivalents par exemple) est de nature à satisfaire les spécifications techniques d'un lot de données (définition des tables, listes énumérées existantes, contraintes topologiques...).

Exemple : le format Google Trafic (GTFS) ou le profil Neptune pour les données sur les réseaux de transport collectif.

Cependant, cette situation favorable ne peut remplacer systématiquement des spécifications. L'existence d'un tel cadre technique apporte des réponses pour le critère cohérence logique. Pour les autres critères (exhaustivité, précision de position, exactitude thématique...) on se retrouve dans la même situation qu'en absence de spécifications.

L'absence de spécifications privilégie alors une approche utilisateur. Le choix des modes de représentation est guidé par les règles suivantes :

- la forme complexe est à proscrire ;
- la forme simplifiée est la plus intéressante si les échelles de mesure sont bien comprises ;
- la forme synthétique est intéressante si elle est mise en perspective avec des usages type.

Dans tous les cas, il conviendra d'informer sur le cadre technique utilisé.

■ Absence complète de spécifications

Cette situation est similaire au cas précédent avec un degré de complexité supplémentaire lié à l'absence d'éléments techniques (système de référence, modèle de données, etc.).

Dans un premier temps, le problème qui se pose est de savoir si la donnée est exploitable aisément. Par « aisément », on entend : qui ne nécessite pas d'analyse approfondie du contenu pour connaître le système de référence ou le sens des attributs et de leur contenu et/ou ce qui est manipulable par les outils SIG standards.

L'usage de la représentation synthétique est opportun pour exprimer si le lot de données est techniquement exploitable ou pas.

Exemple 1 :

Feu vert pour un lot utilisable tout de suite et au contenu non ambigu.

Feu orange pour un lot utilisable mais qui demande une certaine expertise.

Feu rouge pour un lot jugé non exploitable ou requérant trop de traitements préalables.

Exemple 2 : le système de référence n'est pas indiqué et les données ne sont pas superposables aux autres lots de données.

Dans un second temps, si le lot de données est exploitable, on l'analysera et on en exprimera la qualité, ce qui permettra, in fine, de reconstruire une partie des spécifications et du modèle de données.

Ce qu'il faut retenir

Mesurer la qualité d'un jeu de données, c'est bien, savoir la communiquer c'est mieux.

Ce document propose trois formes de présentation, de la plus riche à la plus simple.

La forme complexe, consistant à présenter dans le détail les contrôles et les résultats obtenus, s'adresse prioritairement aux producteurs de données et aux commanditaires et trouve tout son intérêt quand des spécifications existent.

La forme simplifiée qui présente les résultats sur un nombre réduit de critères agrégés est une solution intermédiaire qui intéressera autant les producteurs que les utilisateurs.

Sa représentation polaire, avec des graduations normalisées, permet, outre une vision globale de la qualité des critères agrégés, de comparer des jeux de données entre eux ou par rapport à une référence attendue.

La troisième forme, dite synthétique, consiste à réduire l'ensemble des mesures à une valeur unique.

Cette valeur peut être numérique mais sera avantageusement représentée sous forme graphique dans une gamme limitée de niveaux (généralement 5).

Cette forme synthétique est également adaptée à une utilisation mettant en valeur l'adéquation d'un jeu de données à des usages type.

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



Contact

accueil.dtectv@cerema.fr

Date de publication 2017
ISSN : 2417-9701
2017/60

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

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