Cerema

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

# Critère d'exhaustivité

*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 parait 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 éléments de méthode pour utiliser le critère d'exhaustivité ainsi que des indicateurs pour rendre compte de la mesure de cet élément de qualité.**

## 1. Les définitions utilisées

L'exhaustivité est un critère de la norme ISO 19 157 qui se définit comme la présence et l'absence d'entités, de leurs attributs et de leurs relations. Il se décompose en deux sous-critères :

- **l'excédent** : données excédentaires présentes dans le jeu de données ;
- **l'omission** : données absentes d'un jeu de données.

**Remarque :** ce critère ou ses deux sous-critères peuvent se mesurer pour chaque classe d'objets, attribut ou relation que l'on désire évaluer.

**Fiche n° 08 -** décembre 2017

## 2. Description des mesures possibles à réaliser

Les deux sous-critères sont abordés dans ce qui suit plus en détail, notamment par la description de mesures qui leur sont rattachées dans la norme.

Les différentes définitions de ces mesures sont rappelées dans le **tableau 1** et les types de valeurs qu'elles doivent fournir sont indiquées pour chacune.

| | Mesure | Définition | Type de valeur |
|---|---|---|---|
| **Excédent** | Élément en excès | Indication qu'un élément figure à tort dans les données. | Booléen<br>(la valeur « true » indique que l'élément est de trop). |
| | Nombre d'éléments en excès | Nombre d'éléments d'un jeu de données ou d'un échantillon qui n'auraient pas dû y figurer. | Nombre entier |
| | Taux d'éléments en excès | Nombre d'éléments en excès dans le jeu de données ou dans l'échantillon par rapport au nombre d'éléments qui auraient dû être présents. | Nombre réel |
| | Nombre d'instances d'entités dupliquées | Nombre total de duplications d'objets au sein des données. | Nombre entier |
| **Omission** | Élément manquant | Indicateur qui montre qu'un élément spécifique est absent des données. | Booléen<br>(la valeur « true » indique l'absence d'un élément) |
| | Nombre d'éléments manquants | Comptage de tous les éléments qui auraient dû figurer dans le jeu de données ou dans l'échantillon et qui en sont absents. | Nombre entier |
| | Taux d'éléments manquants | Nombre d'éléments manquants du jeu de données ou de l'échantillon par rapport au nombre d'éléments qui auraient dû y figurer. | Nombre réel |

Tableau 1

## 3. Indicateur retenu

Dans l'optique de fournir un élément de qualité **unique** pour l'exhaustivité d'un jeu de données, et à des fins de simplification, il est proposé d'utiliser une mesure supplémentaire (non décrite dans la norme) désignée par le **taux d'exhaustivité** :

Définition ▪ Nombre total d'éléments en excès ou manquant dans le jeu de données par rapport au nombre total d'objets du jeu de données.

Description ▪ Soient :

$N_m$ : nombre d'objets constatés dans l'échantillon ou dans le jeu de données.

$Nb_E$ : nombre d'objets en excédent dans l'échantillon ou dans le jeu de données.

$Nb_O$ : nombre d'objets manquant dans l'échantillon ou dans le jeu de données.

$N_t = N_m + Nb_O - Nb_E$ : nombre d'objets attendus dans l'échantillon ou dans le jeu de données.

Alors, le taux d'exhaustivité vaut :

$$1 - \frac{(Nb_O + Nb_E)}{N_t}$$

Type de valeur ▪ Nombre réel (souvent exprimé en pourcentage).

**Exemple :** le jeu de données comprend 50 objets, 5 sont en excédents, 2 sont manquants. Le jeu de données devrait donc théoriquement contenir 50 + 2 - 5 = 47 objets.

Le taux d'exhaustivité est alors égal à : 1 – (7/'47) soit : 85 %

**Remarque :** on notera que ce critère qui ajoute les éléments manquants et les éléments en excès. Lorsque l'on veut conserver une information complète de l'exhaustivité, il conviendra de conserver la connaissance des deux sous-critères (taux d'éléments en excès et taux d'éléments manquant) en complément du taux d'exhaustivité.

# 4. Méthode de contrôle

Mesurer l'exhaustivité d'une classe d'objets consiste à évaluer le nombre d'objets manquants ou le nombre d'objets en excédent dans la classe d'objets, ou plus généralement les deux.

Les méthodes employées pour qualifier l'exhaustivité étant proches de celles relatives à la qualification de la précision thématique du jeu de données, il est recommandé de mener ces contrôles en parallèle.

À l'instar de la qualification de la précision thématique, les méthodes diffèrent selon que l'on dispose ou pas d'un jeu de données de référence.

## 4.1 Existence d'une référence reconnue

Lorsque l'on dispose d'une base de données de référence, la méthode est relativement simple et consiste à compter le nombre d'objets manquants et le nombre d'objets en excédent par rapport à cette référence.

**Remarque :** la base de données de référence peut prendre des formes diverses : base de données, fichier tableur, listing, définition administrative, arrêté, site Internet... Cette liste n'est pas exhaustive et il convient de considérer comme référence reconnue toute source d'informations faisant foi dans son secteur même si elle ne provient pas d'une source publique ou officielle.

## 4.2 Contrôle par rapport au terrain nominal

Une base de données peut faire l'objet de spécifications de contenu. La méthode consiste alors à compter le nombre d'objets manquants et le nombre d'objets en excédent par rapport au terrain nominal défini par ces spécifications.

Cela n'exclut pas que l'on puisse le faire également par rapport au monde réel pour recueillir une vision plus complète et plus générale de l'exhaustivité.

L'évaluation de l'exhaustivité doit rester indépendante de l'objectif annoncé dans les spécifications de qualité. Elle reflétera uniquement le taux d'exhaustivité mesuré dans le lot contrôlé.

**Remarque :** En cas d'exigences de qualité annoncées dans les spécifications, la vérification de leur respect relève davantage de la conformité (respect du contrat) que d'une évaluation de la qualité intrinsèque. La conformité ou non-conformité d'un jeu de données n'est pas prise en compte dans les règles de représentation ou de notation de la qualité.

## 4.3 Absence de source de contrôle

En l'absence de spécifications et si aucune base de données de référence n'est connue ou accessible, le seul contrôle possible est la comparaison avec le monde réel. Les méthodes à envisager sont donc :

- le contrôle terrain ;
- le dire d'expert ;
- l'exploitation des référentiels supports de saisie ;
- l'interprétation de la généalogie.

### ■ Le contrôle terrain

Le contrôle terrain est la seule méthode réellement objective, mais elle demande à être appliquée avec rigueur pour garantir la fiabilité et la confiance dans les résultats obtenus.

Le contrôle terrain s'effectue, a priori, sur un échantillon selon le procédé de l'échantillonnage orienté surface[1]. Le recensement des excédents revient à vérifier que l'objet présent dans le jeu de données

1 Pour plus d'information sur l'échantillonnage, voir la fiche n° 5 « Méthodes d'échantillonnage ».

n'existe pas réellement dans la réalité et que sa présence dans la base n'est pas conforme avec le monde réel ou le terrain nominal.

Le recensement des omissions revient à parcourir et analyser la totalité des surfaces de l'échantillon, sans se limiter aux seuls secteurs où des objets apparaissent dans le jeu de données.

Le contrôle terrain peut dans certains cas être avantageusement remplacé par l'exploitation des ortho-photographies à grande échelle lorsque les objets à contrôler sont facilement identifiables et localisables.

### ■ Le contrôle à dire d'expert

Le contrôle à dire d'expert s'appuie sur l'avis d'un sachant ou d'un collège de sachants. Il ne peut être considéré comme valide que si les excédents et omissions sont quantifiés précisément. En cas de doute ou uniquement d'estimation, le dire d'expert ne peut être retenu pour qualifier l'exhaustivité du jeu de données qu'accompagné d'un intervalle de confiance ou d'un niveau d'incertitude. Dans tous les cas, il doit être commenté.

Accompagné d'un commentaire ou d'un intervalle de confiance, le dire d'expert demeure préférable au fait de ne disposer d'aucune information.

**Remarque :** Le contrôle terrain ou le recours à un ou plusieurs experts, sur un échantillon, permet d'obtenir une valeur précise de l'exhaustivité sur la zone étudiée.

### ■ L'exploitation des référentiels support de saisie

Le (ou les) référentiel(s) utilisés lors de la saisie des données a (ont) une importance non négligeable sur la qualité finale de la base de données produite et en particulier sur l'exhaustivité des données.

Il est donc indispensable de consulter les spécifications de chacun des référentiels utilisés si ceux-ci en possèdent. En fonction des données concernées, ces spécifications peuvent constituer une source d'information précieuse.

### ■ L'interprétation de la généalogie

La méthode utilisée pour l'acquisition des données (photogrammétrie, numérisation, lever terrain...) conditionne probablement l'exhaustivité des données et la capacité à mesurer cette dernière. En revanche, elle ne permet pas, à elle seule, de pouvoir l'évaluer avec suffisamment de certitude. L'interprétation de la méthode d'acquisition permet simplement de compléter une des méthodes utilisées et décrites dans ce document. Elle est également très liée au type des objets dont on désire mesurer l'exhaustivité. C'est cependant un élément à prendre en compte lors de la réflexion.

**Remarque :** en plus de ces différentes techniques pour pallier l'absence de sources de contrôle, les retours des utilisateurs peuvent constituer un bon indice de l'exhaustivité (par exemple si plusieurs utilisateurs font remonter le manque de données dans un secteur en particulier). Cela nécessite l'existence d'une procédure pour que les utilisateurs puissent rapporter les erreurs.

## 4.4 Précautions sur le dénombrement

Le simple dénombrement du jeu de données par rapport à une source de contrôle ne suffit pas pour évaluer l'exhaustivité. En effet, excédents et omissions peuvent se compenser.

Le comptage nécessite une phase de comparaison élément par élément dont la méthodologie dépend essentiellement de la forme disponible des données de référence (numériques ou analogiques, centralisées ou réparties à plusieurs endroits, directement utilisables ou nécessitant une interprétation). Il n'y a pas de règle standard ou de méthode systématiquement reproductible pour cette comparaison.

L'objectif affiché est de disposer in fine pour chaque classe d'objets de l'effectif initial du jeu de données $N_m$ et du nombre d'objets manquants $Nb_O$ ou en excès $Nb_E$ afin de calculer le taux d'exhaustivité.

## 4.5 Contrôle par échantillonnage

En cas de contrôle sur un échantillon, les règles standards d'échantillonnage s'appliquent (cf. la fiche n° 5 « Méthodes d'échantillonnage »).

La mesure d'exhaustivité obtenue sur l'échantillon peut être extrapolée à la population entière à la condition que l'échantillon soit suffisamment représentatif.

Une telle valeur doit être accompagnée de son intervalle de confiance.

Pour un taux d'exhaustivité, l'intervalle de confiance s'exprime par la formule :

$$p \pm t \sqrt{(p \frac{(1-p)}{n})}$$

où p est le taux d'exhaustivité évalué de l'échantillon, n la taille de l'échantillon et t le coefficient de confiance prenant les valeurs suivantes :

| Intervalle de confiance | 90 % | 95 % | 96 % | 98 % | 99 % |
|---|---|---|---|---|---|
| t | 1,64 | 1,96 | 2,05 | 2,33 | 2,58 |

En cas de contrôle sur un échantillon, il convient de préciser dans les métadonnées tous les détails qui ont conduit à ces valeurs : la taille de l'échantillon, la valeur de l'intervalle de confiance choisie et les résultats du calcul.

**Exemple :** soit un taux d'exhaustivité mesuré de 92 % sur un échantillon de 50 objets pour un intervalle de confiance de 95 %. L'intervalle de confiance est alors de :

$92\ \% \pm 1{,}96 \times \sqrt{\frac{0{,}92 \times (1-0{,}92)}{50}}$ soit 92 % ± 8 %

## 5. Représentation - Notation

Pour un critère donné, la notation doit rester unique, indépendante du contexte, des différents usages d'un lot de donnée ou de la méthode utilisée.

Le choix des différentes valeurs du tableau ci-dessous peut sembler surprenant, mais elles n'ont pas été choisies au hasard. Disposer de 50 % d'une information peut, dans certains cas bien précis, déjà être considéré comme intéressant même s'il n'est pas possible, par exemple, d'en tirer des enseignements statistiques. Ces valeurs ont été choisies en relation avec des exemples de situations qui ont été proposés par des services utilisateurs qui représentent bien une réalité terrain.

| Taux d'exhaustivité | Note sur 5 |
|---|---|
| De 95 % à 100 % | 5 |
| De 90 % à 95 % | 4 |
| De 75 % à 90 % | 3 |
| De 50 % à 75 % | 2 |
| Taux < 50 % | 1 |

## Ce qu'il faut retenir

Le critère d'exhaustivité est primordial pour la qualification des données géographiques.

La directive Inspire recommande son utilisation dans 22 thèmes sur les 34 référencés, soit dans 65 % des cas.

La norme propose deux sous-critères et sept mesures. L'indicateur préconisé ici pour la qualification n'est pas présent dans la norme ISO 19157. Il représente le taux total d'objets en déficit et en excès par rapport au nombre total d'objets de la base. Il est censé être plus représentatif de la notion d'exhaustivité.

Les contrôles portant sur l'aspect « sémantique » du lot de données (exhaustivité et précision thématique) étant assez proches quant à leur méthodologie, leur réalisation et leur rapportage, il est recommandé de les mener en parallèle.

En présence de données de référence, ce critère peut être simple à mesurer. En revanche, en leur absence son évaluation devient plus complexe et plus difficile. On aura alors recours aux différentes méthodes préconisées comme le contrôle terrain et le recours aux avis d'experts en prenant soin de toujours évaluer le taux d'incertitude qui s'attache à une telle évaluation.

Enfin, la note affectée à ce critère devra toujours être accompagnée des éléments de méthode utilisés pour sa détermination.

## Série de fiches « Qualifier les données géographiques »



## 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), Yvan Bédard (Professeur Honoraire à l'université Laval, CEO d' Intelli³).



**Maquettage**
Cerema Territoires et ville
Service édition

**Impression**
Jouve
Mayenne

Date de publication 2017
ISSN : 2417-9701
2017/62


## 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**


**Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement - www.cerema.fr**
**Cerema Territoires et ville - 2 rue Antoine Charial - 69426 LYON cedex 03 - Tél. +33 (0)4 72 74 58 00**
*Siège social :* Cité des mobilités - 25 av. François Mitterrand - CS 92803 - 69674 BRON Cedex - Tél. +33 (0)4 72 14 30 30