Bordeaux, le 29 Novembre 2023

CRITÈRES DE QUALITÉ DONNÉES

Critères retenus pour les deux niveaux de la qualité des Métadonnées des jeux de données du Catalogue PIGMA

Membres de l'équipe PIGMA
Guillaume Blanchard – Administrateur de la plateforme PIGMA
Héloïse Deschamps – Animatrice technique de la plateforme PIGMA

1. Analyse des critères de qualité de la donnée analysables

Les jeux de données (ou JDD) du catalogue PIGMA se caractérisent par une fiche métadonnées, avec des champs obligatoires à renseigner et d'autres optionnels et une ou plusieurs ressources associées. La plupart des jeux de données ont une ressource principale ainsi que des ressources. Différents défauts peuvent affecter la qualité de ces données : topologies invalides, nommage des champs incorrects, absence d'identifiant, incohérence des attributs temporels, omission d'éléments, etc… Afin de déterminer la qualité des données du catalogue PIGMA, on analyse, pour chacun des 5 critères de la qualité des données géographiques, les sous-critères analysables et les analyses qui peuvent être menées.

1.1 La norme ISO 19157

L'Organisation Internationale de Normalisation, l'ISO, a publié en 2013 la norme 19157 sur la qualité des données géographiques. Cette norme établit des principes et définit des critères de description des données et de leur qualité, en spécifiant des méthodes de contrôle et en décrivant les procédures d'évaluation de cette qualité. Plus spécifiquement, elle pose 5 grands critères de qualité, que le Cerema a synthétisé sous forme de fiches en 2017 :

- **Cohérence logique** : Degré de cohérence interne des données, attestant que le jeu de données est exploitable
- **Exhaustivité** : Présence en excès ou absence d'éléments dans le jeu de données
- **Précision thématique** : Exactitude des informations portées par les entités
- **Précision de position** : Justesse de la position géographique, absolue ou relative
- **Qualité temporelle** : Qualité des attributs et des relations temporels entre les entités

Chacun de ces critères comporte plusieurs sous-critères, qu'on analyse un par un, en précisant les défauts qu'on peut analyser, la métrique utilisée pour mesurer ce défaut (nombre d'erreurs, taux d'erreurs ou conformité stricte) et les remarques qui peuvent être faites quant à la faisabilité de l'analyse.

1.2 Cohérence logique

Nom du critère	Défauts analysables	Représentation	Remarques
Cohérence Conceptuelle	**Non-conformité au schéma conceptuel** : • Identifiant non-unique • Absence de champ identifiant • Identifiant non-renseigné pour une entité • Nom de champs incorrect (espace, accent, caractère spécial, etc.)	Nombre d'erreurs, taux d'erreurs ou conformité stricte	**Critère important** pour l'exploitation. Identification automatique des champs identifiants difficile mais **faisable**. Respect du nommage des champs faisable via des expressions régulières.
Cohérence au Domaine de Valeurs	**Non-conformité au domaine de valeurs** : • Valeurs en dehors de la plage définie (pH d'un sol > 15) • Type erroné (population de 15 250,3 hab) • Valeur d'attribut non prévue par les spécifications	Nombre d'erreurs, taux d'erreurs	Hors-domaine de valeurs **difficile à établir** pour des valeurs numériques → détermination par les écarts-types uniquement si les valeurs sont très concentrées autour d'une valeur moyenne. Hors-domaine impossible pour les valeurs textuelles, sans disposer des spécifications. D'une manière général, nécessité de disposer des plages de valeurs admises pour contrôler ce critère.
Cohérence de Format	**Conflit de structure physique** : • Format de fichier incohérent avec le format renseigné • Absence de fichiers obligatoires (ex : shx dans un shapefile) • Format non-validé par un validateur (JSON invalide) • Format de date erroné (JJ-AAAA-MM) • Format de code incomplet (ex : INSEE_COM = 1300 au lieu de 01300)	Nombre d'erreurs, taux d'erreurs ou conformité stricte	Cohérence entre format réel/format renseigné facilement analysable. Idem pour les fichiers du Shapefile. Usage d'un validateur possible pour certains formats (CSV, XML, JSON) Analyse du format des codes possibles uniquement pour des champs dont le type ou le nom permet d'être quasi-certain qu'il s'agisse de code (ex : type varchar(5), nom : code_commune, etc.)
Cohérence Topologique	**Défauts de topologie** : 1. Connexion arc-nœud erroné 2. Erreur de connexion (undershoot, overshoot) 3. Micro-polygones 4. Pointe (spike)	Nombre d'erreurs, taux d'erreurs ou conformité stricte	Tous les défauts de topologie cités sont analysables. Pour les couches ponctuelles, il reste à déterminer si l'on souhaite que deux points proches soient accrochés (coordonnées

	5. Polygones troués 6. Auto-intersection (polygones papillons) 7. Auto-chevauchement (intersection entre entités mitoyennes)		identiques) ou parfaitement distincts (coordonnées différentes). Les auto-chevauchements se détectent bien ; les lacunes (creux entre entités mitoyennes) sont plus difficiles à détecter. Les défauts 1 et 2 concernent davantage les données de réseaux, les défauts 3 et 4 faussent les calculs de surface. Les défauts 5, 6 et 7 gênent voire invalident les analyses spatiales qu'on peut réaliser.

La Cohérence au Domaine de Valeurs ne pourra vraisemblablement pas être contrôlée. Une notation et une pondération de ces différents sous-critères restent à établir.

Y:

1.3 Exhaustivité

Nom du critère	Défauts analysables	Représentation	Remarques
Excédent	**Éléments en excès dans le jeu de données** : • Éléments non-concernés par le jeu de données • Éléments obsolètes	Nombre d'erreurs, taux d'erreurs	Comparaison par rapport à une donnée référentielle, faisant foi, par rapport à une étendue géographique définie (si un point d'eau du SDIS 16 est situé dans le sud-Dordogne → excédent) **Impossibilité** de déterminer et **d'accéder automatiquement à une donnée de référence** pour un jeu de données donné. Éventuellement détermination des excédents par comparaison spatiale
Omission	**Éléments manquants dans le jeu de données** : • Éléments concernés par le jeu de données mais non-inclus • Éléments supprimés par erreur • Comparaison d'éléments entre deux versions d'un même jeu de données	Nombre d'erreurs, taux d'erreurs	**Nécessité absolue de disposer d'une donnée référentielle.** La donnée OSM n'est pas toujours assez exhaustive et/ou précise pour servir de référence générique Éventuellement comparaison par rapport à une version antérieure de la donnée et rapport de différence remonté au producteur

L'Exhaustivité ne sera vraisemblablement pas analysable sauf sur deux éléments :

- La **comparaison par rapport à une zone géographique définie** (via les métadonnées ou l'étendue spatiale des données), avec un tampon de largeur à définir
- La comparaison par rapport à une version antérieure d'un jeu de données, permettant de détecter des entités potentiellement supprimées par erreur

Une notation et une pondération de ces différents sous-critères devraient également être établies.

Y:

1.4 Précision thématique

Nom du critère	Défauts analysables	Représentation	Remarques
Justesse du Classement	**Classement erroné** : • Valeurs d'attribut qui ne correspond pas à la valeur réelle (ex : Caserne référencée comme Hôpital)	Nombre d'erreurs, taux d'erreurs ou conformité stricte	Comparaison par rapport à la réalité donc impossible à déterminer automatiquement. **Nécessité d'une donnée référentielle, d'un dire d'expert, etc.** Seul le producteur ou un utilisateur attentif peut attester de la Justesse du Classement.
Justesse des Attributs Non-Quantitatifs	**Attributs qualitatifs erronés** : • Valeurs d'attribut qui ne correspond pas à la valeur réelle (ex : Caserne référencée comme Hôpital) • Orthographe de l'attribut incorrecte (ex : surface de type « *phorêt* »)	Nombre d'erreurs, taux d'erreurs	Déterminer la justesse nécessite une donnée référentielle ou un dire d'expert. Seule la **vérification de l'orthographe** peut être faite automatiquement, via des bibliothèques Python (ex : pyspellchecker)
Justesse des Attributs Quantitatifs	**Conflit de structure physique** : • Valeur d'attribut qui ne correspond pas à la valeur réelle (ex : pommier de hauteur 150m)	Nombre d'erreurs, taux d'erreurs ou conformité stricte	Comparaison par rapport à la réalité donc impossible à déterminer automatiquement. Nécessité d'une donnée référentielle, d'un dire d'expert, etc. Seul le producteur ou un utilisateur attentif peut attester de la Justesse des Attributs Quantitatifs

Seule la vérification de l'orthographe dans certains champs peut être effectuée. Une notation et une pondération de ces différents sous-critères devraient également être établies.

Y:

1.5 Précision de position

Nom du critère	Défauts analysables	Représentation	Remarques
Précision Absolue	**Proximité des coordonnées** : • Écart des coordonnées absolues par rapport aux coordonnées réelles	Nombre d'erreurs, taux d'erreurs, valeur moyenne des incertitudes de position	Comparaison par rapport à la réalité donc impossible à déterminer automatiquement. Nécessité d'une donnée référentielle, d'un dire d'expert, etc. Seul le producteur ou un utilisateur attentif peut attester de la Précision Absolue.
Précision Relative	**Proximité des coordonnées** : • Écart des coordonnées relatives par rapport aux coordonnées réelles	Nombre d'erreurs, taux d'erreurs, valeur moyenne de l'erreur verticale/horizontale	Comparaison par rapport à la réalité donc impossible à déterminer automatiquement. Nécessité d'une donnée référentielle, d'un dire d'expert, etc. **Théoriquement faisable** pour des données routières ou de réseaux dont la distance entre les différentes bornes sont connues. Seul le producteur ou un utilisateur attentif peut attester de la Précision Relative.
Présence dans une emprise	**Comparaison par rapport à une emprise de référence** : • Éléments en dehors de l'emprise (avec ou sans tampon)	Nombre d'erreurs, taux d'erreurs ou conformité stricte	**Comparaison spatiale faisable** → analyse similaire à la recherche d'excédents **mais difficulté à caractériser l'erreur** : si une entité est hors-zone → excédent ou écart de position ?

La présence dans une emprise peut être analysée, ce qui pose alors le problème de caractérisation de l'erreur (excédent ou écart de position).
Une notation et une pondération de ce sous-critère devraient également être établies.

Y:

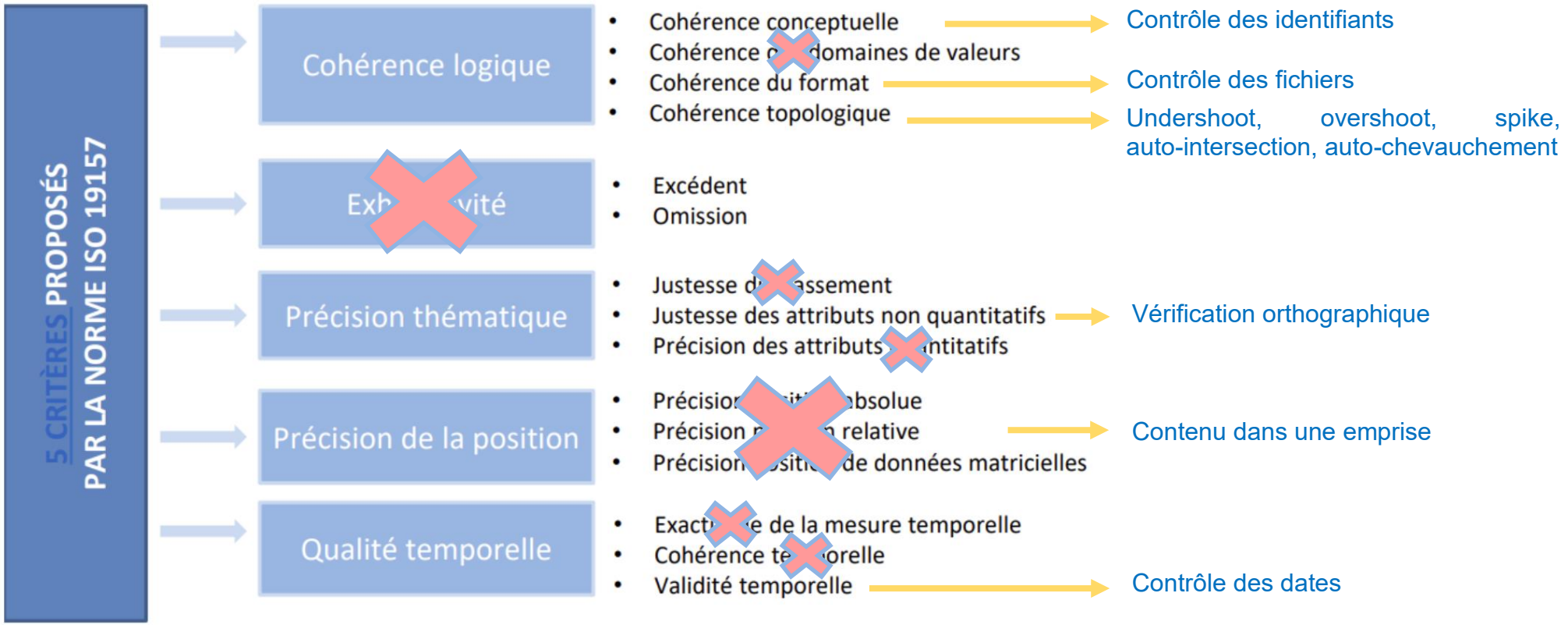
1.6 Qualité Temporelle

Nom du critère	Défauts analysables	Représentation	Remarques
Exactitude de la précision temporelle	**Proximité des attributs temporels** : • Écart des attributs temporels par rapport aux valeurs réelles	Nombre d'erreurs, taux d'erreurs, valeur moyenne des écarts	Comparaison par rapport à la réalité donc impossible à déterminer automatiquement. Nécessité d'une donnée référentielle, d'un dire d'expert, etc. Seul le producteur ou un utilisateur attentif peut attester de l'Exactitude de la précision temporelle.
Cohérence temporelle	**Respect de la chronologie des dates** : • Chronologie des dates/horaires incohérentes (ex : jour de fermeture antérieur au jour d'ouverture)	Nombre d'erreurs, taux d'erreurs, valeur moyenne de l'erreur verticale/horizontale	Comparaison entre deux attributs date, time ou varchar(4). Théoriquement possible de contrôler la chronologie. Avoir à disposition deux attributs date est peu fréquent dans les données du catalogue. En pratique, déterminer le champ date antérieur et le champ postérieur nécessite d'analyser sémantiquement le nom des champs (et y trouver des valeurs comme 'ouverture/fermeture', 'début/fin', 'entrée/sortie', etc.
Validité temporelle	**Dates invalides** : • Dates au mauvais format (JJ-AA-MM) • Dates incongrues (13-13-0221, 30-02-2007, etc.)	Nombre d'erreurs, taux d'erreurs ou conformité stricte	**Défauts de format temporels analysables** (DateTimeConverter, expressions régulières)

Le respect de la chronologie est théoriquement analysable mais reste difficile en pratique. Le critère de Validité temporelle peut être analysé, en contrôlant le format des attributs temporels. Une notation et une pondération de ce sous-critère devraient également être établies.

Y:

2. Synthèse des critères retenus



PIGMA / GIP ATGeRi - 6 Parvis des Chartrons- 33 075 BORDEAUX CEDEX - Tél. 05 57 85 40 42 - Fax 05 57 85 40 26
contact@gipatgeri.fr - www.pigma.com/www.pigma.org

3. Notations et Pondération des critères retenus

3.1 Cohérence logique

Nom du critère	Coeff	Type de critère	Utilité et Qualité attendue	Notation
Identifiant unique		Cohérence conceptuelle	La donnée doit posséder au moins un champ identifiant. Les identifiants doivent être uniques et renseignés pour toutes les entités, sans exception ni doublon.	**Critère exclusif :** La présence d'un doublon ou l'absence d'identifiant pour une entité de la donnée exclut le jeu de données du Label Optimal
Noms des champs		Cohérence conceptuelle	Le nommage des champs exclut les caractères spéciaux, les espaces et les signes de ponctuation.	**Critère exclusif :** La présence d'un ou plusieurs champs mal nommés exclut le jeu de données du Label Optimal
Cohérence et Conformité du Format	**5 et 5**	Cohérence du format	Le format déclaré de la donnée correspond au format réel (Cohérence). La donnée est conforme aux spécifications du format (JSON valide, fichiers indispensables du Shapefile).	**Note sur la cohérence :** **0 :** Les formats réel et déclaré ne concordent pas **10 :** Les formats réel et déclaré concordent **+** **Note sur la Conformité :** **ERREUR :** La donnée n'est pas conforme **5 :** La donnée est conforme mais ne contient que les fichiers indispensables **10 :** La donnée est conforme et contient plus de fichiers que ceux indispensables

Y:

Codes valides	**3**	Cohérence du format	Toutes les valeurs des champs contenant des codes (type code culture, code insee, etc.) correspondent au format du code, avec un seuil d'erreur inférieur à 5%.	**Note sur la cohérence :** **0 :** Au moins un champ code comporte des erreurs, avec un taux d'erreurs supérieur à 5% **5 :** Au moins un champ code comporte des erreurs, avec un taux d'erreurs inférieur à 5% **OU** La donnée ne comporte aucun champ code **10 :** Tous les champs codes sont valides

Y:

3.2 Cohérence logique - Topologie

Nom du critère	Coeff	Type de critère	Utilité et Qualité attendue	Notation
Géométries nulles		Cohérence Topologique	La donnée ne comporte aucune géométrie nulle. Chaque entité possède une géométrie.	**Critère exclusif :** La présence d'une géométrie nulle pour une entité de la donnée exclut le jeu de données du Label Optimal
Micro-géométrie	**5**	Cohérence Topologique	La donnée ne contient pas de micro-géométries (micro-polygones, scories de découpage).	**Note sur la cohérence :** **0 :** La donnée comporte plus de 5% de micro-géométries. **5 :** La donnée comporte moins de 5% de micro-géométries. **10 :** Aucune micro-géométrie
Topologie	**5**	Cohérence Topologique	La donnée ne possède pas de défaut de topologie. Les géométries sont correctement accrochées, sans écart, ni chevauchement. Lorsque deux entités linéaires se chevauchent, ces entités doivent partager un sommet commun. La donnée ne contient pas de pics, des erreurs de connexion des segments d'une géométrie.	**Note sur l'accrochage :** **0 :** La donnée contient au moins une erreur **10 :** La donnée ne contient aucune erreur

Y:

3.3 Précision de la position

Nom du critère	Coeff	Type de critère	Utilité et Qualité attendue	Notation
Emprise de référence	**3**	Précision absolue	Toutes les entités de la donnée sont contenues dans une emprise spatiale de référence (France ou Région Nouvelle-Aquitaine)	**Note sur la position :** **0 :** La donnée comporte au moins une entité hors de l'emprise. **5 :** La donnée est comprise en France. **10 :** La donnée est comprise en Nouvelle-Aquitaine

3.4 Critères non déployés

Nom du critère	Coeff	Type de critère	Utilité et Qualité attendue	Notation
Orthographe	**1**	Précision thématique	Les valeurs dans les champs de typage (type, libellé, etc.) sont correctement orthographiées (ex : type « forêt », au lieu de « phorêt »)	**Note sur la position :** **0 :** Au moins une erreur dans les champs de typage **5 :** La donnée ne comporte pas de champs de typage **10 :** Aucune erreur dans les champs de typage
Dates	**3**	Qualité temporelle	Les valeurs dans les champs « date » sont valides (pas de dates impossibles comme 30 février ou absurdes comme 01/01/0222) et correspondent aux formats suivants : JJMMAAAA, AAAAMMJJ (avec ou sans tiret ou barre oblique)	**Note sur la position :** **0 :** Au moins une date est impossible ou absurde **5 :** Les dates sont valides mais le format des dates n'est pas valide **10 :** Le format des dates est valide et aucune erreur n'est présente

Y: