Cerema

Direction territoriale Ouest
Département Villes et territoires

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
CNIG
Conseil national de l'information géographique

Groupe de Travail Qualité des Données Géographiques
Commission RMO du CNIG

Compte-rendu de réunion du 7 octobre 2021

Participants :

Nom Prénom	Organisme	Présent	Excusé
Geoffrey ALDEBERT	Etalab - data.gouv.fr	x	
Bernard ALLOUCHE	Cerema / DG / supervision des données		x
Jean-Marie ARSAC	AZIMUT	x	
Mathieu BECKER	ISOGEO	x	
Jérôme BOUTET	Idéo Ternum Bourgogne-Franche-Comté	x	
Benjamin CHARTIER	Consultant		x
Olivier DISSARD	CGDD/DRI/MIG		x
Arnauld GALLAIS	Cerema Ouest	x	
Thomas GRATIER	Etalab - data.gouv.fr	x	
Guillaume GRECH	UMS Patrinat, OFB - MNHN	x	
Marie LAMBOIS	IGN / Cellule normalisation	x	
Mathieu Le Moal	Axes Conseil	x	
Hélène LEUFROY	Géo17		x
Jocelyne MARC	IGN / Mission Qualité	x	
Thierry MARTIN	Association OpenDataFrance	x	
Suzanne NICEY	Idéo Ternum Bourgogne-Franche-Comté		
Steven PIEL	OFB / Service des référentiels marins	x	
Nicolas PY	IGN Centre Est	x	
Clémence RABEVOLO	IFREMER		x
Mathieu RAJERISON	Cerema Méditerranée	x	
Gessica REYNAUD	Géomap - Imagis		x
Stéphane ROLLE	CRIGE PACA	x	
Benoît SEGALA	Consultant	x	
Pierre VERGEZ	IGN / Mission CNIG	x	

Ordre du jour :

- Validation du précédent compte-rendu. Points d'info et d'actu.
- Qualité des données dans data.gouv.fr *(G. Aldebert et T. Gratier)*
- Classification des usages de l'info géographique *(J-M. Arsac)*
- Révision des fiches méthodologiques "précision de position" et "précision thématique" *(B. Segala)*
- Modèle de parties qualité et métadonnées des géostandards *(A. Gallais)*
- Outil de restitution de la qualification des données *(M. Rajerison)*
- Point divers

Prochaine réunion : **20 janvier 2022 à 9h30**

Les documents relatifs à cette réunion sont disponibles ici.

1. Validation du précédent CR - Points d'info et d'actu

Point d'infos & actus :

Le GT QuaDoGéo accueille G. Aldebert et T. Gratier (Etalab) et T. Martin (OpenData France). T. Gratier succède à J. Leboeuf, G. Aldebert a été l'initiateur de schema.data.gouv, et T. Martin a pour mission de développer l'open-data et sa standardisation des données au sein des collectivités territoriales.

Le thème de la qualité des donnés a le vent en poupe ! Que ce soit dans les ateliers de réflexions sur le nouveau CNIG, dans la feuille de route du MTE avec un important volet sur la qualité de la donnée, mais aussi par exemple dans "Transformer l'action publique par la donnée" la feuille de route 2021-2023 du ministère de la Transformation et de la Fonction publiques, dont on peut extraire, page 22 :

> Les jeux de données de qualité entraînent naturellement une plus grande réutilisation et de nouveaux usages. Il est par conséquent nécessaire de valoriser ces données et les efforts des administrations afin de favoriser leur réutilisation et inciter les autres acteurs à produire des jeux de données de qualité. Dans cette perspective, Etalab développera un «label qualité» sur la plateforme data.gouv.fr qui valorisera les jeux de données et producteurs vertueux. Les étapes pour développer ce label seront les suivantes :
> - En concertation avec les parties prenantes, identification des métriques d'évaluation de la qualité d'un jeu de données ;
> - Développement d'outils d'analyse automatique de la qualité d'un jeu de données ;
> - Création d'un label qualité sur data.gouv.fr qui valorise les données répondant aux critères de qualité ;
> - Valorisation des jeux de données de qualité dans les résultats de qualité
> **Action A24 :** Accompagner et outiller les administrations dans la production de données de qualité et créer un label de la qualité de la donnée. 1er semestre 2022 DINUM

Etalab a pour sa part dressé une synthèse des commentaires sur data.gouv.fr, dont la plupart en rapport avec la qualité des données, ainsi qu'un très riche atelier de réflexions sur la qualité des données.

- M. Lambois nous informe des avancées des travaux de révision de la norme ISO 19157 :

  19157-1 (modèle de qualité) : les travaux techniques sont finalisés (au stade DIS) et seront publiés en 2022 (ISO 19157-1:2022)

  19157-2 (encodage) : est publié.

  19157-3 (registre) : les travaux débutent. L'intention est de publier un registre en ligne, en collaboration avec l'OGC. Outre la disparition de la partie "useability", le report des mesures sur la qualité dans cette partie 3 offrira une meilleure lisibilité à la norme.

- T. Martin informe que le webinaire « Data dans les territoires » en collaboration entre OpenData France et le Cerema n'a pas été renouvelé au deuxième semestre, mais pourrait l'être en 2022.

- La formation qualité basée sur les fiches méthodologiques du CEREMA, initialement annoncée en juillet 2021 se déroulera les 25 novembre et 6, 8, 13, 15 décembre 2021 en format webinaire de 2h30 mêlant la théorie et la pratique. Cette formation ouverte à tous est pilotée par S.Rolle et M. Rajerison.

Les contenus des cinq modules sont décrits ici : module 1 : 25 novembre, module 2 : 6 décembre, module 3 : 8 décembre, module 4 : 13 décembre, module 5 : 15 décembre.

Revue des actions :

Le compte-rendu du précédent GT CNIG QuaDoGéo est relu et validé.

Actions réalisées :

- présentation du GT QuaDoGéo à la journée des utilisateurs Sextant le 22 juin (A.Gallais)
- Fourniture de jeux-tests de données de qualification ISO 19157 au format XML (M. Lambois)

Actions à lancer ou poursuivre :

- étudier le GUF de GeoNetwork, et tester la restitution sur une plateforme GeoNetwork 4 pour développer une démonstration. Les premiers retours de M. Lambois avaient révélé que l'outil s'avère a priori assez compliqué à installer et à mettre en œuvre.

2. Qualité des données dans data.gouv.fr

Par G. Aldebert. La présentation est disponible ici.

L'initiative schema.data.gouv.fr est portée par l'équipe data.gouv.fr d'Etalab. L'accompagnement vise en particulier les producteurs de schéma de données et les producteurs de données afin qu'ils atteignent un bon niveau de qualité. L'équipe développe pour ce faire les outils schema.data.gouv.fr, publier.etalab.studio et data.gouv.fr visant l'amélioration de la qualité des données publiées en open data.

L'accompagnement comprend actuellement :

- deux guides sur la qualité des données et sur la création de schémas de données ;
- l'infolettre sur data.gouv.fr informe des actualités, ainsi que des événements ponctuels;
- le site éditorial schema.data.gouv.fr donnant accès à la documentation correspondant à un schéma spécifique ;
- un forum autour des schémas de données publiés ou en cours de conception sur le dépôt Github, avec des réunions régulières pour conseiller les producteurs de schéma sur la stratégie de publication.

Le processus de conception d'un schéma prévoit :

- la publication d'une annonce préalable auprès de la communauté ;
- l'organisation de groupes de travail avec plusieurs acteurs pour la conception ;
- l'implémentation technique du schéma réalisée sur un dépôt Git public.

Avant toute publication d'un nouveau schéma sont vérifiés : l'intérêt public des données produites ; la conformité du schéma par rapport à un format (tableschema, jsonschema...) et aux pratiques préconisées (noms des champs, exemples, descriptions claires…) ; et enfin le fait que le schéma fasse bien consensus entre plusieurs acteurs.

T. Gratier cite l'exemple classique du schéma de la Base Adresse Locale. M. Rajerison cite celui des arrêtés de circulation pour le transport de marchandises publié par le Cerema en partenariat avec OpenDataFrance et la région PACA.

Questions / débats

A. Gallais remarque que le CNIG et Etalab sont de fait deux producteurs et diffuseurs de schémas et standards et ont naturellement vocation à collaborer en bonne connaissance de leurs actions respectives afin d'éviter de produire des schémas ou standards en doublon.

A. Gallais demande si les schémas Etalab peuvent gérer plusieurs tables, G. Aldebert

indique que ce n'est pas encore le cas compte tenu de l'utilisation du format json mais qu'une réflexion est en cours à ce sujet.

N. Py fournit à Etalab l'exemple du validateur GPU qui vérifie la conformité des données au standard CNIG PLU c'est à dire dans une structure relationnelle non limitée à une seule table.

T. Martin signale qu'il rejoint l'équipe OpenDataFrance (ODF) en charge du projet OpenDataFactory qui inclut la stratégie de standardisation des données portée par l'association ODF. Comme indiqué par G. Aldebert, ODF et Etalab entament une série de réunions pour faire converger leurs travaux sur les sujets de la gouvernance de la standardisation. Ces travaux vont également porter sur la convergence des initiatives en matière d'outillage pour la production des données par les producteurs (D-Lyne) et leur contrôle (Validata) puisque cet outil va évoluer vers le traitement de données json.

ODF intègre également dans sa démarche la priorisation du développement des usages. Ceci renforce les enjeux de la standardisation, de la qualité et de l'agrégation de données à des échelles territoriales. ODF relance pour ce faire un nouveau volet du projet OpenDataLocale d'accompagnement des collectivités à l'open-data sur un périmètre élargi mais centré sur l'accompagnement à la production et à la publication des données du SCDL qui va être très largement dynamisé par le projet OpenDataFactory. Il reste néanmoins à mobiliser une communauté d'acteurs sur ce dernier point.

3. Classification des usages de l'info géographique

Par J-M. Arsac.

Le groupe souhaite abonder la nomenclature des usages établies en 2020 par J-M Arsac dans le cadre de la méthodologie pratique pour qualifier des données, dans un mode collaboratif en faisant appel à l'intelligence collective en s'appuyant sur des réseaux tels que Géotribu, Géotamtam, Géorezo. Le besoin de contribution pourra notamment être relayé par le CNIG.
Cette nomenclature constitue un entrant indispensable à la maquette de restitution graphique de qualification de la donnée géographique (cf. §6)

J-M. Arsac rapporte qu'un appel à contributions lancé sur Géorezo a suscité de premiers échanges. Cet appel renvoie vers le document collaboratif en ligne (framapad) que la communauté des utilisateurs peut abonder avec les différents usages identifiés de l'information géographique.

A. Gallais suggère d'alimenter également la nomenclature avec de nombreux items de la nomenclature des usages exposée par l'IGN sur son site Géoservices, qui ont le mérite de renvoyer à des usages très concrets.

La difficulté de l'exercice semble d'établir le bon compromis entre une trop grande généricité des usages (qui seraient passe-partout) et des usages trop précis (pas assez généraux).

Le groupe de travail reconnaît la nécessité d'adopter deux entrées : l'une par la thématique concernée, l'autre par l'usage.

S. Rolle suggère de structurer la nomenclature de façon hiérarchique, à l'instar de la nomenclature Corine Land Cover.

La publication sera élargie à d'autres publics, en dehors de la sphère géomatique.

Décision / Actions :
- Abonder la nomenclature des usages de l'information géographique (tous)
- La consolider (J-M. Arsac, S. Rolle, J. Boutet)
- Relayer la démarche via Etalab, TeamOpenData (T. Gratier) et le CNIG (P. Vergez)
[Hors réunion] : fait pour TeamOpenData et Twitter.

4. Révision des fiches méthodologiques "précision de position" et "précision thématique"

Par B. Segala. Avec présentation des fiches révisées : "précision de position" et "précision thématique"

Précision de position

La révision de B. Segala propose la nouvelle mesure "Classement des objets contrôlés en fonction des écarts de positionnement constatés, selon un ou deux seuil(s)" qui est de type « dénombrement des écarts à la norme » pour qualifier les géométries “complexes“, linéaires ou polygonales, qui ne relèvent pas de travaux topographiques (donc pas de l'arrêté du 16 septembre 2003).
Cette mesure évite de devoir de mesurer point par point la précision du positionnement, et permet de classer les objets contrôlés en fonction des éventuels écarts de positionnement qu’ils présentent, ce qui allège grandement le travail de contrôle. La mesure comporte soit un seuil unique (telles les mesures 30 et 31 de la norme 19157) ou mieux : deux seuils pour cadrer la dispersion des incertitudes.

Précision thématique

La révision met en valeur la mesure du “taux global de bon classement“ ou “taux d’accord global“ pour qualifier la justesse d’un classement. Elle reconnaît la « matrice de classement erroné » (MCM, ou “matrice de confusion“) comme mesure de base pour la raison que l'on peut en déduire tous les autres indicateurs, au lieu (dans la première version de la fiche) d'une des deux matrices relatives de classement erroné (RMCM).
Enfin, elle apporte des avertissements quant à l’utilisation du coefficient kappa, et ne recommande plus cette mesure.

Décision / Actions :
- Relecture les deux fiches, jusqu'au 31 octobre. Silence vaudra acceptation (tous)
- Vérification de la conformité ISO19157 pour la mesure avec deux seuils (M. Lambois)
- A l'issue du processus de relecture : envoi des fiches au service de PAO (A. Gallais)
- Actualisation du registre des mesures de qualité des données géographiques en conséquence (A. Gallais)

5. Modèle de parties qualité et métadonnées des géostandards

Par A. Gallais. Avec présentation du document.

Cette action avait été évoquée lors de précédents GT QuaDoGéo. A la demande de P. Vergez pour les besoins de l'élaboration du standard CNIG StaR-Elec, A. Gallais a rédigé un modèle de parties qualité et métadonnées des géostandards.
Inspiré de géostandards CNIG et COVADIS existants, il comprend le « tronc commun des géostandards » où l'on retrouve :
- la fiche d’identification du standard

- des recommandations pour la saisie ou collecte des données
- le chapitre "Qualité des données" avec l'énoncé des objectifs de qualité et les mesures préconisées faisant référence au registre des mesures de qualité des données géographiques
- des règles d'organisation et de codification des identifiants, des attributs de type date, etc.
- le chapitre "Métadonnées", qui s'est vue actualisé et complété (en coordination avec M. Lambois) d'un §3.9 "Autres mesures qualité" permettant d'ajouter toutes les mesures qualité préconisées au chapitre "Qualité des données".

Questions / débats

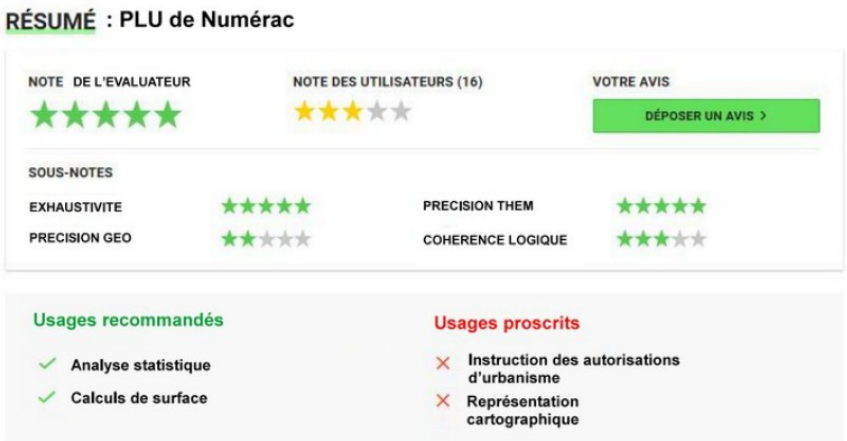
N. Py propose :
- d'ajouter que le critère d'exhaustivité est relatif à la sélection des objets du terrain nominal, c'est à dire "vu suivant le filtre des spécifications" ;
- de prévoir une rubrique dédiée à la gestion des identifiants, présentant la description de leur format et/ou construction, recommandant leur unicité, leur stabilité et pérennité, avec des règles de filiation entre objets, et des consignes pour l'interopérabilité ;
- de préciser que le fichier de métadonnées est à constituer au format xml ;
- de revoir les titres des § 4.3.8 et 4.3.9 dans la mesure ou le premier concerne les métadonnées de qualité minimales ISO 19115, et le deuxième des métadonnées supplémentaires directement liées aux mesures préconisées pour la thématique concernée.

Décision / Actions :
- *Relire le document jusqu'au 31 octobre. Silence vaudra acceptation (tous)*
- *Fournir les éléments de la rubrique dédiée à la gestion des identifiants (N. Py)*
- *Prendre en compte des remarques de N. Py et de tous (A. Gallais)*
- *Confronter la partie métadonnées à l'avis du GT Métadonnées (M. Lambois)*

6. Outil de restitution de la qualification des données

Par M. Rajerison. La présentation est disponible ici.

Le groupe de travail s'est orienté vers une maquette graphique comprenant :
- *deux notes globales : la note du testeur (vert), et celle des utilisateurs (orange) avec une rubrique « déposer un avis » ;*
- *des sous-notes suivant les critères (facilité d'obtention, facilité d'usage, fraîcheur, etc.) ;*
- *une rubrique « usages recommandés / usages à éviter ».*



L'outil est développé en python. Son code est libre et disponible sur le Gitlab dédié. Il

permet de restituer graphiquement les mesures correspondant à la qualité interne du lot de données, des notes de synthèse de la qualité ainsi que des notes de qualité selon les usages. Il s'appuie à la fois sur les mesures de la norme ISO 19157 via la structure définie dans dataQualityElements.xsd, et sur le registre des mesures de qualité de l'information géographique.

Les paramètres de qualification sont générés et restitués sous différentes formes graphiques via une feuille de style. L'outil est indépendant de tout logiciel SIG.

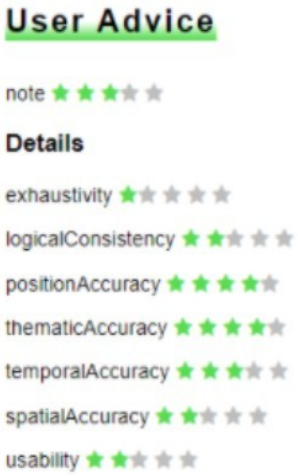
En comparaison à la précédente version de la preuve de concept présentée lors de la dernière réunion, l'objectif était pour M. Rajerison de s'appuyer sur un véritable fichier XML conforme à 19157. Il a donc poursuivi le développement de l'outil en s'appuyant sur un fichier XML ISO-19157 fourni par M. Lambois : **fake-dng2.0-iso.xml** comprenant 1553 nœuds dont 979 effectivement renseignés.

Certains indicateurs, liés à la balise "DataQualityInfo" fournissent une indication synthétique sous forme de lettre, c'est notamment le cas de la qualité de précision planimétrique et altimétrique.

M. Rajerison s'est donc interrogé sur l'existence d'un dictionnaire ou des clés d'interprétation du XML, et sur celle d'un registre de valeurs pour certaines mesures.

Le rendu graphique s'est attaché à correspondre au maximum aux champs de ce fichier :

L'outil permet de représenter graphiquement des champs pouvant correspondre soit à une saisie manuelle, soit à un calcul automatisé.



Exemple de champs dont les valeurs peuvent être saisies manuellement.

Pour la suite des travaux, M. Rajerison prévoit d'enrichir le rendu avec de nouvelles informations et de développer un traducteur vers la maquette.

Il souhaiterait également pouvoir disposer d'autres fichiers xml à titre d'exemple.

Questions / débats

A. Gallais distingue deux phases : la partie amont consistant à obtenir les indicateurs de qualification à partir des résultats de mesures (=> report.xml), et la partie aval consistant à restituer ces résultats sous forme d'indicateurs graphiques (quadorender et report.html).

La partie amont étant assez complexe car toutes les valeurs n'étant pas exploitables (à commencer du fait de leur nombre) et tous les calculs n'étant pas nécessairement automatisables comme l'a montré M. Rajerison, il recommande de se focaliser préalablement sur la partie aval en visant une restitution graphique conforme à la maquette initiale en s'appuyant sur des simulations de mesures.
L'objectif visé est un rendu graphique moins technique qu'actuellement, en français et directement compréhensible par le grand public avec les parties présentées dans la maquette : la notation globale et la possibilité de retour des utilisateurs, la qualification synthétique des critères de la norme, et la recommandation suivant les usages en lien avec la nomenclature des usages en cours de construction (cf §3).
M. Lambois et N. Py recommandent l'exploitation préalable du xml et sa traduction graphique détaillée, et proposent que la représentation synthétique soit traitée dans un deuxième temps.
Dans les deux cas, suggestion est faite que l'utilisateur puisse accéder aux informations d'évaluation synthétique de la qualité, pour ensuite pouvoir cliquer et "déplier" l'évaluation pour - in fine - pouvoir consulter l'information d'évaluation détaillée.

Décision / Actions :
- Enrichir le rendu avec de nouvelles informations graphiques, développer un traducteur vers la maquette, établir le lien avec la nomenclature des usages (M. Rajerison)
- Fournir d'autres fichiers xml conformes à ISO 19157 (M. Lambois)

A ce stade, la production effective ou en cours du GT CNIG QuaDoGéo comprend :

- Les fiches méthodologiques du Cerema : sur le site du Cerema ou en accès individuels : Introduction - généralités - contexte du contrôle qualité - éléments statistiques - méthodes d'échantillonnage - mode de représentation - cohérence logique - exhaustivité - précision thématique - précision de position - qualité temporelle ;
- Le registre des mesures de qualité des données géographiques et le registre en ligne sur le Géocatalogue ;
- La carte mentale au sujet de la qualification des données suivant la norme ISO 19157 ;
- Deux pistes exploratoires de qualification de données (synopsis, logigramme) sur des thèmes particuliers ;
- La méthodologie de qualification de données *(à consolider…),* comprenant :
  - une ébauche de nomenclature des usages génériques de l'information géographique ;
  - une maquette de restitution de la qualité des données ;
- Une présentation du GUF, outil de retour utilisateur quant à la qualité des données.
- *(en cours)* Le prototype d'un outil de restitution graphique de la qualification des données géographiques suivant la maquette.
- *(en cours)* révision des fiches "Précision de position" et "Précision thématique"
- *(en cours)* la Nomenclature des usages génériques de l'information géographique
- *(en cours) Modèle de parties qualité et métadonnées des géostandards*
- Le fil de discussion Géorezo : "Qualité des données géographiques"