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 20 mai 2021

Participants :

Nom Prénom	Organisme	Présent	Excusé
Bernard ALLOUCHE	Cerema / DG / supervision des données		x
Jean-Marie ARSAC	AZIMUT	x	
Mathieu BECKER	ISOGEO	x	
Jean-Marie BOURGOGNE	OpenData France		x
Jérôme BOUTET	Idéo Ternum Bourgogne-Franche-Comté	x	
Benjamin CHARTIER	Consultant	x	
Olivier DISSARD	CGDD/DRI/MIG		x
Bénédicte DURAND	CARENE service Topo-Cartographie		x
Arnauld GALLAIS	Cerema Ouest	x	
Stéphane GARCIA	IGN / Cellule normalisation		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
Suzanne NICEY	Idéo Ternum Bourgogne-Franche-Comté		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.
- Recherches sur la classification des usages de l'info géographique *(M. Lambois)*
- Principes généraux de qualification des données du SIMM *(C. Rabevolo)*
- Proposition de révision de deux fiches méthodologiques *(B. Segala)*
- Outil de restitution de la qualification des données *(M. Rajerison)*
- Point divers

Prochaine réunion : **7 octobre 2021 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 nouveau président du CNIG est Bertrand Monthubert. Il ouvrira le 10 juin la Commission Données du CNIG jumelée avec la Commission Animation territoriale.

La norme ISO 19157 sur la qualité des données géographiques est en cours de révision en version : ISO 19157-1:2022, dont la publication est attendue en 2022.
M. Lambois indique d'une par la disparition de la partie "useability" et d'autre par le report des mesures sur la qualité dans une partie 3 en cours de rédaction, ce qui offrira une meilleure lisibilité.

M. Rajerison signale un nouveau standard en cours de constitution concernant les arrêtés de circulation, élaboré par le Cerema et OpenData France dans le cadre de la fabrique de la logistique et hébergé sur le Github de Etalab. Ce modèle de données fait partie du Socle Commun des Données Locales.

La formation « Data dans les territoires » s'est bien déroulée avec ses huit modules. Les supports et vidéos sont disponibles ici sur le Gitbook d'OpenData France. Une deuxième session est envisagée au deuxième semestre.

S. Rolle et M. Rajerison annoncent la reprise, avec un lifting, de la formation à distance QuaDoGéo. Le CRIGE PACA organise une séance de promotion le 13 juillet. La formation se déroulera au deuxième semestre 2021 sous la forme de modules courts, techniques et axés sur la pratique.

Revue des actions :

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

Actions réalisées :

- A. Gallais a présenté les travaux du GT QuaDoGéo à la Commission Données du CNIG du 23 mars. L'intégration de spécifications de qualité dans les géostandards se fera au fur et à mesure des nouveaux géostandards ou de leurs mises à jour.
- Suite à l'appel à commentaires, le Registre français des métadonnées relatives à la qualité des données géographiques a été actualisé sur le Géocatalogue (maintenu par le BRGM) et le document a été publié sur le site du CNIG.

Actions en cours traitées en séance :

- proposition de révision de fiches méthodologiques (B. Segala, cf. §4)
- poursuite du développement de l'outil de génération du résultat graphique conforme à la maquette et adaptation le script au flux XML de la norme ISO 19157 (M.Rajerison, cf. §5)

Actions à lancer ou poursuivre :

- étudier le GUF de GeoNetwork, et tester la restitution sur une plateforme GeoNetwork ;
  Les premiers retours de M. Lambois révèlent que l'outil s'avère a priori assez compliqué à installer (une recompilation est nécessaire) et à mettre en œuvre en l'intégrant si possible à GeoNetwork 4 pour développer une démonstration. Cependant, l'institut cartographique espagnol l'utilise avec succès dans l'un de ses géocatalogues et M. Lambois a demandé un compte de démonstration. Elle ne lâche pas l'affaire, à suivre…
- définir un modèle de géostandard incluant la qualification des données (cf. ISO 19131) à l'instar de ce que proposaient les géostandards COVADIS.

2. Classification des usages de l'info géographique

Par M. Lambois. La présentation est disponible ici.

M.Lambois a recherché une classification des usages de l'information géographique dans la documentation et différents ouvrages normatifs : le thesaurus GEMET (General Multilingual Environmental Thesaurus), le portail européen des données, les cas d'usages du Geospatial User Feedback (GUF), la norme de spécification de données ISO 19131 elle-même inspirée de la modélisation proposée dans ISO 19119 pour les services...
Elle constate que l'on retombe toujours sur la nomenclature GEMET ou sur des catégories thématiques ou, dans le cas du GUF, sur la notion d'activité autour de la donnée (donnée pour la recherche, donnée pour la production, etc.) mais hélas jamais sur la notion d'usage de l'information géographique.

Débats :
M. Le Moal s'interroge sur le fait de typer les usages : est-ce le bon angle ? Ne devrait-on pas plutôt lister en fonction des actions opérées sur l'information géographique, par exemple : analyse spatiale, statistique, différents niveaux de traitement ?
Le groupe de travail décide de construire / d'abonder l'ébauche de 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 pour abonder un framadoc ou un googledoc ou une carte mentale en ligne. Le besoin de contribution pourra notamment être relayé par le CNIG et par Géorezo.
Cette nomenclature constituera un entrant de la maquette de restitution graphique (cf. §5)

Décision / Actions :
- Lancer une démarche collaborative pour abonder la nomenclature des usages de l'information géographique (J-M. Arsac)

3. Principes généraux de qualification des données du SIMM

Par C. Rabevolo. La présentation est disponible ici.

Le Système d'Information Milieu Marin (SIMM) a pour objectif de centraliser les données françaises sur le milieu marin.
Le portail milieumarinfrance permet l'accès aux données du SIMM via un catalogue (intégré dans le portail par une API) hébergé par l'IDG de l'Ifremer Sextant, qui repose sur Geonetwork.
Le SIMM doit mettre en place des processus qualité afin de permettre un accès à une information fiable. Pour cela, le Service d'Administration des Référentiels (SAR) du SIMM a publié une note proposant des recommandations : Les principes généraux de qualification des données du SIMM. Ces recommandations concernent notamment :

- le renseignement des fiches de métadonnées, qui se veut le plus exhaustif possible ;
- l'implémentation du standard GUF sur le retour utilisateur ;
- l'intégration de certains des critères de qualification de la norme ISO 19157 via le registre dans les fiches et les standards du SIMM pour les jeux de données géographiques homogènes.

Il faudra effectuer un important travail d'accompagnement des producteurs de données afin que les indicateurs de qualité soient renseignés correctement. Les retours d'expérience d'autres IDG sur le GUF et l'utilisation de la norme ISO 19157 permettront d'accélérer la

démarche. La mise à disposition d'une première version d'un outil de retour utilisateur sur le catalogue du SIMM est attendue pour l'été 2021. Il va être envisagé d'ajouter un champ sur l'usage des données.

Débats :

En marge de sa présentation C. Rabevolo signale un excellent poster publié à l'IMDIS d'avril 2021 au sujet de l'utilisation des métadonnées sur la qualité pour rechercher et filtrer les lots de données dans les (géo)portails, les bonnes pratiques et l'utilisation d'un label à quatre niveaux (qualité contrôlée et bonne, qualité contrôlée et à confirmer, qualité contrôlée mais inconnue, et qualité non contrôlé), qui constitue un levier très motivant pour la publication de lots de données de qualité reconnue et affichée.

B. Chartier signale que le remplissage du champ SIMM sur l'usage des données pourra venir alimenter la typologie des cas d'usages initiée par le GT QuaDoGéo (cf § 2)

Décision / Actions :

- [hors réunion] présentation du GT QuaDoGéo à la journée des utilisateurs Sextant le 22 juin (A.Gallais)

4. Proposition de révision des fiches méthodologiques "précision de position" et "précision thématique"

Par B. Segala. La présentation est disponible ici, avec deux contributions : l'une sur le critère de précision de position et l'autre sur le critère de précision thématique.

Précision de position

Il s'agit de disposer d'un indicateur pour qualifier le positionnement d'une couverture polygonale. Lorsque cela ne relève pas de travaux topographiques, les deux indicateurs retenus dans la fiche méthodologique n'apparaissent pas appropriés. Il est proposé de recommander l'utilisation de méthodes de type « dénombrement des écarts à la norme », ce qui allège grandement le travail de contrôle : avec un seuil unique (mesures 30 et 31 de la norme 19157) ou mieux, avec deux seuils pour cadrer la dispersion des incertitudes.

Précision thématique

La question est de choisir les meilleurs indicateurs permettant de caractériser la justesse d'un classement. Il est tout d'abord suggéré de mettre en concordance les indicateurs "retenus" (sélectionnés et recommandés) dans la fiche méthodologique avec le registre des mesures à intégrer dans les métadonnées. Concernant les autres remarques, il est proposé :

- de prioriser l'utilisation du « taux global de bon classement » (ou « taux d'accord global ») ;
- de reconnaître la « matrice de classement erroné » (MCM, ou "matrice de confusion") comme mesure de base, au lieu (actuellement) d'une des deux matrices relatives de classement erroné (RMCM) ;
- et d'alerter sur les défauts du « coefficient kappa ».

Débats :

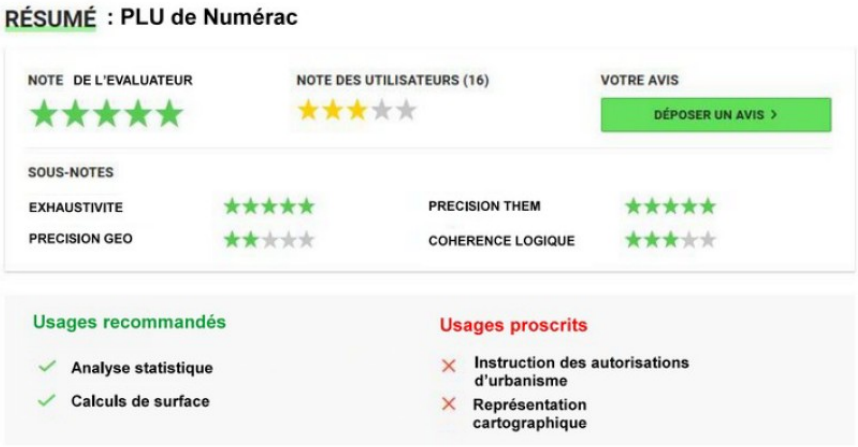
- N. Py et B. Segala appuient la recommandation de la matrice brute MCM pour la raison que l'on peut en déduire tous les autres indicateurs. A. Gallais rappelle le choix initial de recommander la matrice relative RMCM avec des pourcentages plus adaptés à une compréhension immédiate, mais c'était sans compter sur le fait qu'il existe deux RMCM.

Décision / Actions :
- les indicateurs à recommander dans la fiche sont : l'indicateur de bon accord global, puis la matrice de classement erroné brute. Le coefficient Kappa, jugé peu fiable, restera mentionné dans la fiche mais sans être recommandé.
- la révision sera suivie d'une nouvelle publication des deux fiches méthodologiques (B. Segala et A. Gallais)
- le registre des métadonnées relatives à la qualité des données géographiques sera mis en cohérence avec la révision des choix méthodologiques exposés dans les fiches (A. Gallais)

5. Outil de restitution de la qualification des données

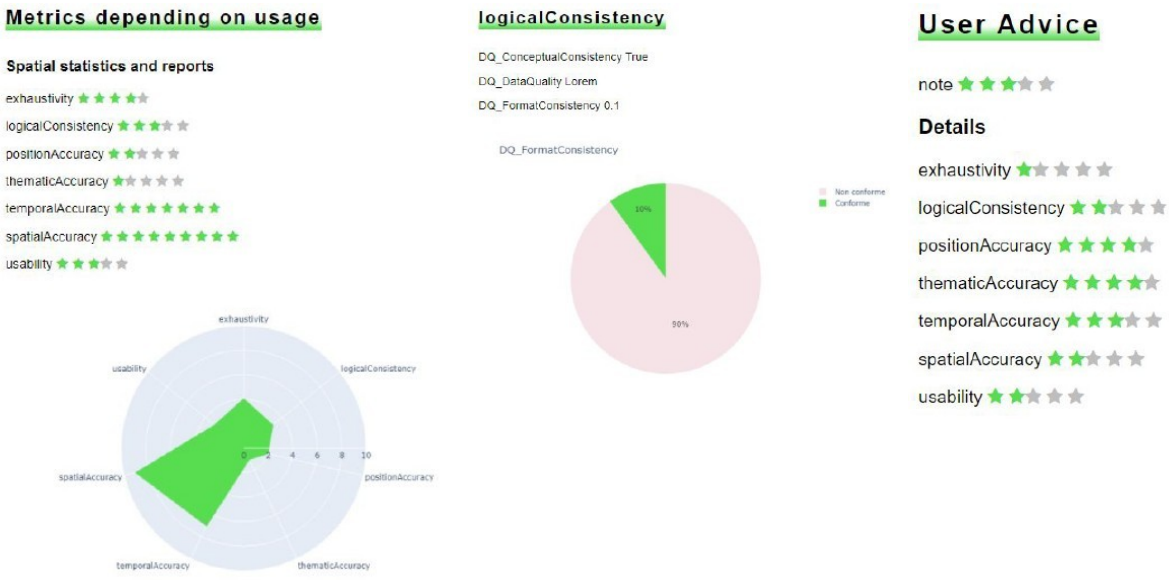
Par M. Rajerison. La présentation est disponible ici.

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 ».

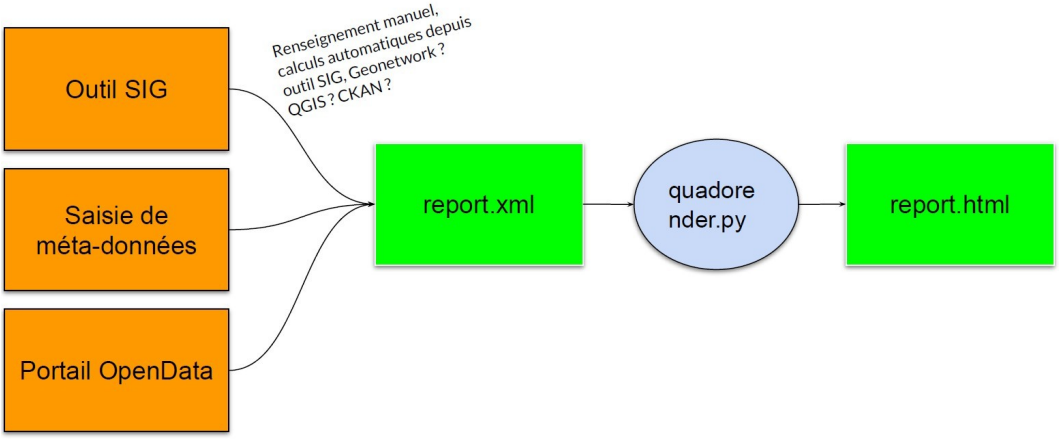


M. Rajerison a poursuivi le développement de l'outil de génération du résultat graphique.
L'outil est développé en python. Son code est libre et disponible sur le Github dédié. Il permet de restituer graphiquement les mesures correspondant à la qualité interne du lot de données, des notes de synthèse de la qualité ainsi que des notes de qualité selon les usages.

Suite aux recommandations de M. Lambois et N.Py en réunion précédente de réutiliser les flux XML de la norme ISO 19157 et du Geospatial User Feedback (GUF), la nouvelle version de l'outil (nom de code : "Quadorender 19157"), s'appuie désormais à la fois sur la norme 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 graphiquement restitués sous différentes formes graphiques via une feuille de style ré-adaptée au rendu de la maquette graphique.



M. Rajerison rappelle d'une part la difficulté de définir des notes de qualité à partir des valeurs des mesures brutes (la fonction n'est pas linéaire…), et d'autre part que l'outil nécessite un référentiel des usages de l'information géographique en entrée (cf. §2). Il souhaite qu'un lien soit établi entre les mesures de la norme apparaissant dans le .xsd et celles retenues dans le registre des mesures de qualité des données géographiques (en ce qui concerne celles issues de la norme, car certaines proviennent de ISO 19115, ou directement des fiches méthodologiques).
Pour tester son outil M. Rajerison souhaiterait pouvoir disposer de jeux-tests de données de qualification au format XML et compatibles à la norme.
Enfin, il préconise que la solution soit ultérieurement testée au sein d'un portail OpenData.



Débats :
Sur le schéma du processus, 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, l'action prévue se focalise 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, francisée et davantage abordable par le grand

public avec les trois parties présentes 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.
M. Lambois informe qu'elle dispose de jeux de données conformes à ISO 19157 pour les tests, et elle suggère de présenter le prototype à des conférences internationales.
N. Py indique que, si d'un point de vue informatique la maquette démontre la faisabilité de cette restitution, il convient cependant de construire l'échelle de valeur des indicateurs.
Il est par exemple courant *(à tort, cf* cette *critique* *§7.5)* de cibler une valeur de 85% de bon classement d'une occupation du sol comme critère permettant de juger cette donnée comme bonne. Il s'avère en effet d'autant plus difficile de dépasser les 95% de bon classement que le nombre de classes est élevé. Par conséquent, le taux de bon classement se voit donc généralement réduit à une échelle de jugement réduite à la plage]85-95[, ce qui ne laisse que peu de place à la nuance. Corollairement, une structure investissant faiblement dans la constitution d'une occupation du sol atteindra aisément un tel seuil de bon classement jugeant la donnée acceptable, tandis qu'une structure visant l'excellence sera dans l'obligation d'injecter des moyens bien supérieurs. Un travail bibliographique pourrait être mené pour recenser dans une matrice (indicateur x produit) les seuils de jugement de valeurs communément admis.
Le groupe s'accorde sur le fait que l'outil devra être capable d'exploiter les métadonnées.
M. Becker recommande de rester indépendant de tout outil SIG, et d'utiliser si besoin des outils de conversion (outils de types extract translate load (ETL).

Décision / Actions :
- Etablir la correspondance entre les mesures de la norme (cf. diapo 7) et le registre des mesures de qualité des données géographiques (M. Rajerison et A. Gallais)
- Cibler la maquette graphique, se rapprocher de son usage grand public (M. Rajerison)
- La partie "usages recommandés / à éviter" de l'outil s'appuie sur la nomenclature des usages, qui reste à développer (cf. §2).
- Fourniture de jeux-tests de données de qualification ISO 19157 au format XML (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 de la 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