Cerema

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

QuaDoGéo

Groupe de Travail Qualité des Données Géographiques
Commission Règles et Qualité du CNIG

Compte-rendu de réunion du 30 juin 2022

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
Mickaël BORNE	IGN	x	
Jérôme BOUTET	Idéo Ternum Bourgogne-Franche-Comté		x
Benjamin CHARTIER	Consultant - anim. GT Open Data Afigéo		x
Adeline COUPÉ	IGN / Géoplateforme	x	
Laurène DEBRAY	Office international de l'eau / Sandre		x
Olivier DISSARD	MTE/CGDD Admin données algos codes	x	
Arnauld GALLAIS	Cerema Ouest	x	
Thomas GRATIER	Etalab - data.gouv.fr	x	
Guillaume GRECH	UMS Patrinat, OFB - MNHN		x
Jean-Marie FAVREAU Jeremy KALSRON	Université CA / LIMOS	x	
Marie LAMBOIS	IGN / Cellule normalisation	x	
Sebastien LAUNAY	IGN / Dépt Contrôles Qualités		x
Mathieu Le Moal	Axes Conseil		x
Marie MAHIER	OFB / SIG		x
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
Johan RICHER	multi / OpenDataFrance / Validata	x	
Stéphane ROLLE Cédric LEPICIER	CRIGE PACA		x
Benoît SEGALA	Consultant - Cabinet d'études	x	
Pierre VERGEZ	IGN / Mission CNIG		x
David VIGLIETTI	Office international de l'eau / Sandre	x	

Ordre du jour :

- Validation du précédent compte-rendu. Points d'info et d'actu.
- Qualité des données dans OpenStreetMap *(Jean-Marie Favreau)*
- Actions dans le cadre de la feuille de route de la donnée *(O. Dissard)*
- Classification des usages de l'information géographique *(J-M. Arsac et S. Rolle)*
- Validateur(s) de données pour la Géoplateforme *(A. Coupé)*

Prochaine réunion : **13 octobre 2022 à 9h30**

Les documents relatifs à cette réunion sont disponibles ici.

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

Revue des actions du précédent compte-rendu :

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

Actions présentées en séance :

- Consolider la nomenclature des usages (J-M. Arsac, S. Rolle, J. Boutet)

Actions à lancer ou poursuivre :

- Rencontre bilatérale Sandre - IGN pour approfondir les questions relatives aux processus de contrôles qualité et leur restitution aux producteurs et aux utilisateurs (N.Py)
- Partager des modèles de rapport de contrôle qualité (tous)
- suivi du dispositif GUF Géonetwork v4 (M. Lambois, C. Rabevolo)

Point d'infos & actus :

- Présentation du contrôle qualité de la production de l'OCSGE au prochain GT QuaDoGéo de septembre par Arnaud Braun (IGN)
- Etalab a lancé un questionnaire sur l'expression (d'un score) de la qualité des données géographiques. L'utilisation d'un code couleur serait bienvenu.

- La formation “Qualité des données géographiques” organisée par le CRIGE (S. Rolle) et le CEREMA (M. Rajerison) s'est déroulée du 2 au 13 mai dans une version renouvelée et toujours gratuite. Plus de 60 géomaticiens ont suivi les 5 sessions de formation. *(source : Géonews - mai 2022)*
- Datactivist a lancé un projet de recherche-action sur la standardisation. Il *"repose sur une*

enquête sociologique, principalement qualitative, qui se déroulera sur la période 2022–2023 avec des entretiens semi-directifs avec des concepteurs, diffuseurs et réutilisateurs de standards ; des observations, y compris participantes, dans des espaces de conception collaborative de nouveaux standards ; un rapport sur la diversité des démarches passées de standardisations". A. Gallais a été interviewé dans ce cadre par E. Ho-Pun-Cheung.

- data.gouv.fr a publié une série d'articles sur la réutilisation des données ouvertes.

- O. Dissard signale la publication par Etalab (via #Teamopendata) d'un tableau de cas d'usage du service public de la donnée (SPD) fournissant des exemples concrets de l'utilité de l'ouverture des données.

- N. Py a mentionné la publication Land Use Cover Datasets and Validation Tools traitant du contrôle qualité de données d'occupation du sol et incluant des cas pratique sous QGIS.

2. Qualité des données dans OpenStreetMap

Par Jeremy Kalsron. La présentation est disponible ici.

J. Kalsron fait une présentation générale d'OSM, carte collaborative créée il y a 18 ans et intéressant 8 millions d'utilisateurs. La donnée, sous licence ODbl, est éditée par ses contributeurs grâce à un système de tags. Contrairement au fonctionnement classique d'un SIG, OSM ne superpose pas des couches de données, mais des données identifiées par leurs tags. Les trois types d'objets sont les nœuds, polylignes et relations. OSM offre différents outils et différents modes de contribution collaborative.

En termes de qualité, la précision de position des objets découle souvent des référentiels de données sous-jacents et elle est souvent équivalente à celle du référentiel cadastral, ou d'une orthophotographie de 20cm de résolution, ou de traces GPS post-traitées manuellement.

Concernant la cohérence logique, les choix de modélisation peuvent s'avérer hétérogènes suivant les territoires, à l'instar de la modélisation des trottoirs qui peut revêtir différents modes.

Questions / débats

Le critère d'exhaustivité n'est pas systématiquement évalué, des études ont traité du sujet mais elles datent de 2015. Des outils tels que Streetcomplete permettent de l'améliorer. J.Kalsron s'appuie sur l'exemple des trottoirs pour indiquer une forte croissance de la collecte des données pour OSM à partir de 2013.

La précision sémantique est variable en fonction des zones géographiques et de la motivation des contributeurs, mais elle s'améliore sensiblement lorsque des collectivités, à l'instar de Montpellier, développent des outils pour contrôler et valider les données OSM avant et afin de les intégrer au SIG de ses services techniques.

A ce sujet C. Rabevolo demande si OSM intègre en retour des tags de type "validé par la collectivité locale". J.Kalsron indique que OSM n'a pas cette vocation dans la mesure où la donnée peut être modifiée dans l'heure qui suit par un autre contributeur.

3. Actions dans le cadre de la feuille de route de la donnée

Par O. Dissard.

O. Dissard fait un point sur les deux premières actions menées dans le cadre de la feuille de route de la donnée et des codes sources, dans le cadre desquelles s'est monté un groupe de travail avec des membres de la communauté de la donnée.

Dans le cadre de la politique publique de la donnée, une circulaire du Premier ministre du 27 avril 2022 dispose que chaque ministère nomme un Administrateur ministériel des

données, des algorithmes et des codes sources (AMDAC ou AMD), et une feuille de route de la donnée, des algorithmes et des codes sources a été rédigée pour le ministère de la transition écologique et le ministère de la mer.

https://www.ecologie.gouv.fr/feuille-route-donnee-des-algorithmes-et-des-codes-sources

La qualité des données y est décrite comme une action phare et un ensemble d'actions y sont proposées. Celles-ci se décomposent en :

- des actions de prise de conscience
- des actions d'anticipation
- des actions d'amélioration

Dans un 1er temps, nous nous attachons à la prise de conscience. En effet durant les interviews qui ont précédé la rédaction de la feuille de route (entamée au moment de la mission Bothorel), personne ne s'est focalisé sur ce sujet pourtant prégnant. Aussi a-t-il été décidé de travailler sur des **principes généraux** de la qualité des données pour le pôle ministériel destiné à sensibiliser les agents à tous niveaux, étant entendu que maintenant tout le monde ou presque travaille sur des données ou sur le fondement de données au quotidien, y compris au plus haut niveau de décision : les décisions s'appuient sur des données et leur pertinence dépend en partie de la qualité des informations qui les appuient.

Pour cette phase, il a été décidé de rédiger un « recto-verso » simple d'approche, lisible rapidement et s'appuyant sur un cahier d'illustrations commentées composées de data-visualisations et de dessins.Nous bénéficions pour ce travail de l'appui d'un dessinateur professionnel (Etienne Appert) fourni par Sopra-Steria.

Un groupe de travail constitué de membres de la communauté de la donnée du pôle ministériel participe à ce travail qui est complété par un outil d'autodiagnostic « de 1er niveau » à l'intention des entités qui produisent et utilisent de la donnée.

Une fois ces travaux terminés, ils seront bien sûr présentés au GT QuaDoGéo.

4. Classification des usages de l'info géographique

Par J-M. Arsac. La présentation est disponible ici*.*

Le groupe de travail remodèle la nomenclature des usages établies en 2020 par J-M. Arsac dans le cadre de la méthodologie pratique pour qualifier des données.

Cette nomenclature constitue l'élément indispensable à la qualification de la donnée par rapport à ses utilisations potentielles.

Un document collaboratif (framapad) a été mis en ligne que la communauté des utilisateurs peut abonder avec les différents usages identifiés.

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

Des nomenclatures existent mais, par exemple, le catalogue INSPIRE apparaît trop détaillé. La même remarque peut être faite pour la nomenclature des usages exposée par l'IGN sur son site Géoservices, qui ont le mérite de renvoyer à des usages trop précis ou trop thématiques, pour apparaître comme génériques.

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

Le GT QuaDoGéo s'est donc prononcé pour la sélection d'une courte liste d'actions (un verbe à l'infinitif) correspondants à une dizaine d'usages génériques.

J-M Arsac rappelle que le critère d'utilisablité ou d'usage des données se fonde sur les trois piliers : 1/ l'action (pour quoi faire) ; 2/ l'échelle / l'étendue du territoire ; 3/ la thématique (le domaine, le métier).

Il indique par ailleurs que l'utilisabilité (terme qui finalement remplace "usage") diffère selon les deux points de vue du producteur (utilisation prévue) et de l'utilisateur de la donnée (utilisation expérimentée ou effective), concept que l'on retrouve dans les publications d'Etalab.

Dans une nomenclature hiérarchique, un temps envisagée, le second niveau découle du premier. Or, dans la proposition qui combine action et thématique, il s'agirait plutôt d'une nomenclature de type matriciel (2D). De ce fait, l'idée de structurer la nomenclature de façon hiérarchique à deux niveaux a finalement été abandonnée au profit de l'approche matricielle usage / thème.

Cette nomenclature est avant tout destinée à des humains, plutôt qu'uniquement à des machines. De ce fait elle a été simplifiée : la notion d'échelle de représentation n'est plus retenue, le nombre d'actions et de thèmes a été réduit à l'essentiel, et il conviendra de bien documenter à la fois les actions (bien définir ce que sous-entend chaque verbe car les termes sont polysémiques) et les thématiques en s'appuyant sur des illustrations concrètes de chaque croisement thème / usage.

A ce stade, la matrice comprend :

- une vingtaine de thème identifiés par l'IGN (cf. diapo 7) : agriculture ; aménagement du territoire ; biodiversité ; chasse ; climat ; culture ; défense ; eau ; éducation ; énergie ; espace ; europe ; fiscalité locale ; forêt ; innovation et numérique ; médico-social ; mer et littoral ; prévention des risques ; santé ; sécurité ; social ; télécoms ; tourisme ; transports et mobilité.

  => Ils peuvent être complétés, par exemple : accessibilité.
- une dizaine d'usages génériques : [Visualiser / cartographier] ; [Inventorier / recenser] ; [Localiser] ; [Naviguer / acheminer] ; [Gérer] ; [Suivre / observer] ; [Analyser] ; [Planifier] ; [servir de donnée référentielle et/ou de donnée pivot]. Il reste à identifier en tant que telles, ces données dont l'usage est également de constituer une donnée pivot (adresse, code INSEE, identifiant universel, etc.)

Ce tableau collaboratif constitue le support de la réflexion du sous-groupe sur la nomenclature des usages, en présentant la double entrée thématique / usage.

Questions / débats

Le tableau sera publié prochainement et complété de façon collaborative.

S. Rolle exprime la difficulté d'identifier le bon couplage thématique / usage, et que certains usages aujourd'hui référencés paraissent trop spécifiques. Un stagiaire au CRIGE travaille sur la question de la qualification des données par l'usage, et/ou comment définir un usage de manière générique en privilégiant l'entrée par l'usage (localisation, observation, données pivot, etc.) dans les deux visions producteurs / utilisateur.

La discussion porte ensuite sur la compréhension et l'utilisation concrète tableau, qui actuellement offre plutôt une entrée par la thématique que par l'usage (colonne B). Les colonnes C-D-E portant la réflexion collective, sont amenées à disparaitre.

Décision / Actions :

- Edition collaborative du tableau thème/usage (tous)

- Publication de la première version de la nomenclature d'ici le prochain GT QuaDoGéo avec la documentation nécessaire à l'utilisateur (J-M Arsac, S. Rolle, J. Boutet)

5. Validateur(s) de données pour la Géoplateforme

Par A. Coupé et M. Borne. La présentation est disponible ici.

> *L'IGN a présenté en mai le projet de généralisation du validateur du GPU lors de la présentation de la cartographie fonctionnelle de la Géoplateforme. Pour mémoire, le rapprochement Etalab et IGN au sujet de la validation des données nous avait été présenté par J. Marc et T. Gratier au GT QuaDoGéo du 20 janvier 2022.*

La présentation au GT QuaDoGéo vise à éclairer les travaux de l'IGN quant à la question de la validation des données de différentes thématiques, et la généralisation de la question de la généralisation dans le contexte du déploiement de la Géoplateforme. Les thématiques Urbanisme, TRI, PCRS font aujourd'hui l'objet de validateurs dédiés mais il est question d'évoluer à l'avenir vers l'utilisation d'une API qui servirait de "passe-plat" et rendrait ainsi transparent et plus performant l'appel à tel processus de validation de structure de données.

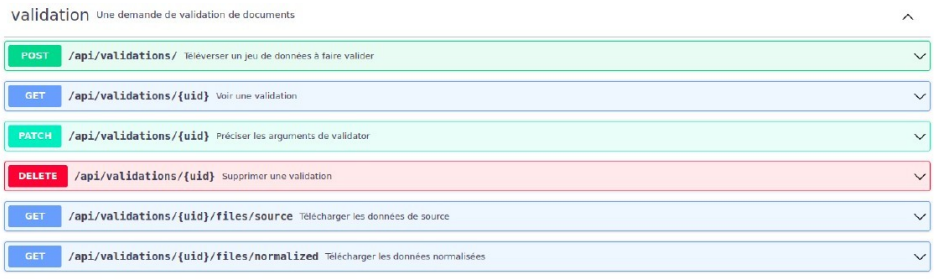
Le validateur du GPU est le précurseur en la matière. Développé à partir de 2014, il s'appuie sur un méta-modèle conforme aux standards CNIG de dématérialisation des documents d'urbanisme. Les deux (standard et méta-modèle) ont conjointement évolué au fil du temps. Un effort particulier a été réalisé pour clarifier les rapports de validation et fournir l'aide nécessaire, notamment l'explicitation des codes d'erreurs. Le méta-modèle opère à la fois sur la modélisation des fichiers et de leur arborescence, et sur chaque table de données prise individuellement. Les cas complexes (validation inter-attributs, entre plusieurs tables, ou portant sur les métadonnées) ont été reportés dans un plugin dédié pour éviter d'alourdir le cœur du dispositif.

Le validateur des territoires à risque important d'inondation (ou validateur TRI) est issu du validateur GPU mais a permis de lui appliquer des améliorations quant aux contrôles d'unicité et de référence. Il fait l'objet du même mode d'intégration que le validateur GPU avec comme différences : l'export GeoJson des rapports de validation et une documentation pdf des codes d'erreurs au lieu de l'aide en ligne.

Le validateur PCRS se différencie des deux premiers en se présentant comme un validateur de référence pour un modèle GML complexe. Les données GML du PCRS sont en effet modélisées avec un XSD par le standard CNIG PCRS. Ce cas d'utilisation a amené à ajouter le concept de fichiers "multi-tables" au validateur pour les données GML contenant plusieurs collections qui pourra être réutilisé au besoin pour valider des données GeoPackage.

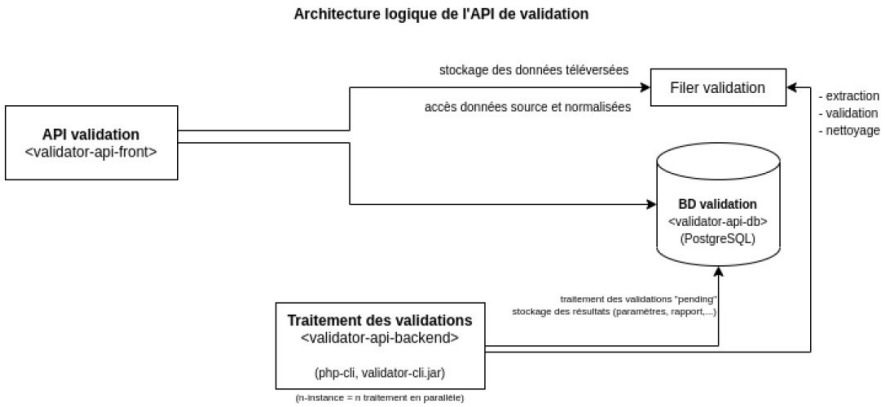
Le déploiement de ces trois outils de validation amène naturellement l'IGN à la volonté d'éviter de créer autant de validateurs que de standards de données, et à l'ambition de développer une API web de validation pour l'espace collaboratif. Cette vision cadre tout à fait avec la réorientation générale des développements à l'IGN vers des architectures API et micro-services dans le contexte du déploiement de la Géoplateforme.

Un démonstrateur générique est déployé, reposant sur une API simple inspirée de validata, concrétisant le rapprochement IGN / Etalab.



Les modèles de table sont au maximum alignés sur le "modèle de modèle" Tableschema mais avec des ajustements nécessaires ainsi que la nécessité de valider des types géométriques et la volonté de conserver des syntaxes simples de type SQL pour la définition des contraintes pour faciliter la maintenance des standards

L'ensemble repose sur une architecture assez simple... qui fait aujourd'hui l'objet d'un démonstrateur.



L'ensemble reste à industrialiser et devra dépasser certaines limites inhérentes à cette nouvelle approche "API".

Questions / débats

Finalement tous les validateurs peuvent-ils être regroupés en un seul ? Cela semble encore illusoire à court terme, du moins tant qu'il n'existe pas de "modèle-de-modèle" de référence et générique qui couvre tous les cas de modélisation, ce qui ne semble pas simple car chaque "modèle de modélisation" (tableschema, plusieurs tables, à plat, relationnel, avec une arborescence de fichiers, etc.) est différente avec à la fois ses forces et ses faiblesses. M. Borne considère qu'il demeure incontournable de travailler à façon en fonction de ce qui est à valider. Finalement, une validation modulaire constitue une bonne approche pour s'adapter à différents standards / thématiques / validations à traiter. Il indique que la gestion et l'implémentation des méta-modèles / standards s'avère en fait plus chronophage que le développement des validateurs.

A. Gallais suggère que le CNIG, et autres organismes de standardisation, pourraient réfléchir au thème de la généralisation des concepts des méta-modèles dans le cadre de la commission Règles et usages.

J. Richer se montre très intéressé par le sujet et établit le lien avec Tableschema. Il indique que suite aux travaux sur le validateur Validata depuis plusieurs années, multi développe pour OpenDataFrance, et en collaboration étroite avec Etalab et transport.data.gouv.fr, un nouveau validateur pour données JSON sur conformité schema.data.gouv.fr. Le double objectif est d'aider les producteurs de données à comprendre les erreurs qui leur sont remontées et d'améliorer la validation déjà prévue par transport.data.gouv.fr. La prochaine réunion du GT QuaDoGéo sera pour lui l'occasion de présenter ce validateur sur des données GéoJson qui ne sont plus essentiellement tabulaires.

Il remarque que Tableschema est vu comme un sous-groupe de Jsonschema mais qu'il n'y a pas encore de validateur de données géographiques dans le contexte des données ouvertes, mais cela va sans doute émerger sous l'impulsion d'Etalab. Le validateur GeoJson n'est pour le moment pas considéré comme générique mais utilisable pour le cas

d'utilisation aménagements cyclables, sans garantie encore qu'il puisse être compatible et utilisable pour d'autres modèles de données.

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 ;
- Les webinaires de formation à la qualité par le CRIGE PACA et CEREMA
- 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) 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"