PIGMA Plateforme d'échange de données en Nouvelle-Aquitaine

ATELIER DATAVIZ
COMPTE RENDU

DATE DE REUNION	**08/06/2023**
Lieu de réunion	Salle 145 (IJBA)

Affaire suivie par : Guillaume BLANCHARD, Héloïse DESCHAMPS et Emeric PROUTEAU

1. Ordre du jour

- Travaux contrôle qualité des données de la plateforme PIGMA, étape intermédiaire
- Travaux sur la feuille de route dataviz de la plateforme PIGMA

2. Compte rendu

2.1. Introduction

L'équipe PIGMA remercie les participants venus nombreux à cet atelier, organisé dans le cadre des Rencontres régionales PIGMA 2023, avant de présenter l'ordre du jour issu du dernier GT OPEN DATA PIGMA tenu en novembre 2022.

PIGMA rappelle que ces travaux font suite à un constat d'un nombre conséquent de données ouvertes diffusées dans la plateforme notamment suite à l'ouverture des données avec la loi Lemaire (octobre 2016) et qui sont peu réutilisées. PIGMA a soulevé la nécessité de monter en qualité pour valoriser ces données, pour leur meilleure réutilisation et diffusion.
Pour obtenir un indicateur de qualité, PIGMA s'est basé sur des références nationales en retenant des critères qui puissent faire l'objet de traitement automatique, être communs à toutes les données, soient appréciables par des non thématiciens de la donnée et présents dans le modèle du jeu de donnée PIGMA.

L'indicateur de qualité concerne à la fois la métadonnée (fiche d'information sur la donnée) et la donnée elle-même. Une méthode commune avait été définie, testée sur 5 premiers jeux de données et présentée au dernier GT OPEN DATA PIGMA.

- L'indicateur de qualité de la métadonnée s'appuie sur le score de qualité mis en place par datagouv, portail national des données open data porté par l'Etat : 5 sur 6 critères ont été retenus de par leur présence dans le modèle du jeu de donnée PIGMA.

- L'indicateur de qualité de la donnée s'appuie quant à elle sur les fiches méthodologiques produites par le CEREMA concernant la norme ISO 19 157 (2013) qui mesure la qualité de la donnée : 3 sur 5 critères ont été retenus, certains sous-critères ont été simplifiés ou écartés.

Ce sont les résultats de mesure de la qualité des métadonnées et des données testées qui ont servi de base pour le calcul de la note finale. Plutôt que de proposer une note de type complexe (mesures brutes) ou simplifiée (schéma radar) ou schématique (smiley, note, étoile), le GT OPEN DATA PIGMA 3 a acté un système de label dont la présentation et l'intitulé restaient à définir. Ce label s'appliquerait seulement à partir d'une certaine note pour valoriser les jeux de données dont la qualification est notable et pour faire améliorer la qualité des autres jeux de données.

Le GT OPEN DATA PIGMA avait identifié ces actions pour l'année 2023 :

- le choix des indicateurs de qualité (notation)
- le choix des données prioritaires à analyser (élargir le nombre de jeux de données à qualifier)
- le recrutement d'un stagiaire pour poursuivre les travaux déjà engagés par PIGMA : stage qui a commencé début avril et finit fin août

2.2. Les travaux de contrôle de la qualité des données de la plateforme PIGMA, étape intermédiaire

L'objectif de la présentation est d'exposer la méthode et les premiers résultats des travaux de qualification des données du catalogue PIGMA réalisés dans le cadre du stage de Guillaume Blanchard au sein du GIP ATGeRi.

Le besoin de qualification des données s'inscrit dans une démarche plus large de l'open data et de l'augmentation du nombre et du volume des données, notamment sur la plateforme PIGMA, qui recense plus de 11 500 jeux de données. Les enjeux de la qualification sont triples :

- Alerter l'utilisateur sur la qualité des données consultées
- Favoriser la réutilisation des données
- Sensibiliser les producteurs et les publicateurs à la qualité de leurs données

Ces travaux ne sont pas nouveaux et s'inscrivent dans la continuité des travaux de synthèse de la norme ISO 19157, menés par le Cerema, et du GT QuaDoGéo du CNIG. Ils s'appuient également sur deux travaux récents dans ce domaine :

- Le stage de Cédric LÉPICIER au CRIGE PACA sur l'établissement de chaîne de contrôle qualité de la donnée autonome, sur une donnée
- Les travaux de qualification menés par PIGMA, visant à fixer des critères de qualité des métadonnées et de la donnée, à proposer une méthode de travail et à opter pour une représentation sous forme d'un label qualité, plus valorisant

La qualité des données s'articule autour de deux axes : la qualité des métadonnées, avec pour référence de travail le Score de métadonnées de DataGouv et la qualité de la donnée géographique, avec pour référence les fiches techniques du Cerema. En complément de ces travaux, PIGMA a arrêté une stratégie consistant à définir deux niveaux de qualité (en premier la qualité des métadonnées, en second la qualité de la donnée) puis à prioriser les jeux de données à qualifier. Au total, 530 jeux de données de couverture régionale sont concernés par le champ d'étude du stage.

À partir de ces enjeux, ces travaux et ces références de travail, 3 objectifs concourent à **valoriser la qualité** sur le catalogue PIGMA :

- Définir des critères de qualité généraux, non-spécifiques
- Automatiser le contrôle qualité pour un grand nombre de jeux de données
- Établir une représentation sous forme de Label Qualité

Depuis le GT OpenData 3, différentes actions ont été réalisées :

- Identification des champs du modèle PIGMA pour la qualité
- Définition des contours du Label Qualité
- Automatisation du contrôle qualité et premiers résultats

2.2.1. Identification des champs du modèle PIGMA pour la qualité

À partir des 5 critères Qualité des métadonnées proposés par DataGouv et retenus par PIGMA, les champs correspondants au modèle PIGMA ont été identifiés :

- Description des données : Champs Nom, Description, Type de données, Mots-clés et Thématiques
- Mise à jour : Champs Fréquence de Mise à Jour, Date de Publication et Date de Mise à Jour (des Métadonnées)
- Licence : Champs Type de licence, Préconisation d'usage et Contraintes d'utilisation
- Métadonnées des ressources : Nom, Format, Type de ressource, Date de Mise à Jour (de la ressource)
- Couverture spatiale : Champs Granularité et Emprise spatiale

Parmi ces champs, plusieurs sont obligatoires lors du renseignement d'une fiche de métadonnée. Également, certaines informations sont pertinentes pour la qualité (comme la présence d'un contact) mais ne sont pas rattachables à une catégorie de DataGouv.

2.2.2. Définition des contours du Label Qualité

À partir de ces informations, deux niveaux pour le Label Qualité ont été définis :

- Un Label Minimal, centré sur la qualité des Métadonnées, qui tient compte de 3 critères :
    - La présence de champs de base indispensables
    - L'accès à une ressource principale dans un format réutilisable
    - Une donnée qui soit fraîche ou millésimée
- Un Label Optimal, analysant la qualité des Métadonnées de façon plus poussée et la qualité de la donnée, avec :
    - Les critères du Label Minimal
    - Des informations de métadonnées détaillées et concordantes
    - Une donnée exempte des défauts topologiques retenus

Les champs indispensables sont les suivants :

- Noms (Métadonnées et Ressource)
- Description
- Mots-clés
- Licence
- Date de Publication
- Date de MAJ des Métadonnées
- Fréquence de Mise à Jour
- Type de données

- Organisation rattachée
- Contact au sein de l'organisation
- Thématiques
- Granularité
- Type de ressource

À ces champs de base s'ajoute également le Millésime, qui est obligatoire dans certains cas (cf Donnée fraîche et millésimée). Si un seul de ces champs est absent, le jeu de données ne peut obtenir le Label Qualité.

La ressource doit être disponible dans des formats réutilisables : GeoJSON,Shapefile ZIP, GTFS,NeTEx, WFS, CSV, JSON, XML, Jpeg2000, etc.

Enfin, la donnée doit être fraîche ou millésimée. 3 sous-critères sont pris en considération :

- La fréquence de mise à jour : elle doit être régulière (annuelle, quotidienne, mensuelle, etc.) et respectée
- Les dates de mise à jour : elles doivent être inférieures à 2 ans
- Si la fréquence de mise à jour n'est pas régulière ou si les dates de mise à jour ont plus de 2 ans, alors le millésime doit être renseigné

La borne temporelle pour les dates de mise à jour a été fixée en première approche à 2 ans et pourra être abaissée à un an. Cette notion de fraîcheur de la donnée permet de garantir à l'utilisateur que les données associées au jeu de données seront vraisemblablement utilisables et pertinentes, au regard des utilisations qu'il souhaite en faire.

La présentation détaillée du Label Optimal n'était pas l'objet de cette présentation, l'application concrète des critères retenus n'ayant pas encore été menée concernant ce Label. De plus, les critères de qualité de la donnée à retenir sont encore à définir. Ainsi, les résultats en termes de qualité concernent uniquement le Label Minimal de Qualité. Cependant, des premiers critères ont déjà été définis et seront à améliorer.

2.2.3. Automatisation du contrôle qualité et premiers résultats

Ces critères ont été transposés sous forme d'un script FME, récupérant les jeux de données étudiés et analysant chacun des critères et sous-critères définis précédemment. L'obtention du Label Minimal de qualité ne dépend pas d'une notation globale, avec un seuil au-dessus duquel le label serait attribué, comme cela avait été défini en amont du stage. À la place, le Label Minimal est attribué à tous les jeux de données qui remplissent l'ensemble des critères et sous-critères, sans exception. Sur 530 jeux de données étudiés, seuls 71 obtiennent le Label Minimal, ce qui représente 13% du total. Ainsi, 459 jeux de données sont rejetés par le script de qualification.

Plusieurs enseignements peuvent être tirés de l'analyse détaillée des résultats. Tout d'abord les champs indispensables sont globalement bien renseignés avec peu de champs manquants. Les premiers résultats mettent en lumière l'absence de valeurs dans certains champs, dépendant de la récupération des données. Ces absences ne sont donc pas dues aux producteurs ou publicateurs des données et ont vocation à être résolues prochainement.

Ensuite, on remarque que certains critères sont déterminants pour la qualité : l'obligation d'un millésime, la présence d'un format réutilisable, les dates de mise à jour inférieures à deux ans et le respect de la fréquence renseignée éliminent respectivement 312, 301, 121 et 103 jeux de données. Ainsi, le processus de qualification fonctionne et applique correctement les critères définis et permet de mettre en lumière les améliorations à apporter aux jeux de données.

En effet, le faible nombre de jeux de données labellisés pousse à mettre en œuvre des méthodes de montée en qualité des données, soit par modification directe des valeurs renseignées par les administrateurs de PIGMA, soit par un retour fait au producteur sur la qualité des jeux de données concernés.

Des pistes d'intégration ont également été proposées. Le Label pourrait s'insérer dans le portail de recherche, à proximité de la vignette du jeu de données et afficher, au survol de la souris, un encart rappelant les critères remplis ayant permis l'obtention du Label (champs de base, format réutilisable et fraîcheur de la donnée). Comme cet encart ne serait visible que si le jeu de données est labellisé, tous les critères apparaîtraient nécessairement remplis sur l'encart. Ainsi, le label permettrait bien de mettre en valeur la qualité d'un jeu de données, sans déprécier les autres jeux de données de la plateforme.

Proposition non-développée à ce jour, un rapport HTML pourrait également être généré ou récupéré, détaillant les champs remplis, les valeurs prises en compte pour ce jeu de données, offrant un résumé détaillé de la fiche de métadonnées de ce jeu de données et de sa qualité.

2.2.4. Suite du travail de qualification

À l'issue des premiers résultats de qualification présentés lors de cet atelier, le travail se poursuivra jusqu'au mois d'août, de la façon suivante :

- Amélioration de la qualité des jeux de données
- Qualification du Label Optimal : Précision des critères de qualité des métadonnées, définition des critères de qualité de la donnée à retenir (cohérence topologique, cohérence de format, etc.)
- Bilan du processus de qualification : Perspectives d'intégration du Label et de son fonctionnement dans la plateforme, réflexions autour de l'animation autour de la qualité à mener auprès des acteurs et des partenaires

2.3. Les travaux sur la feuille de route datavisualisation de la plateforme PIGMA

2.3.1. Retours sur l'atelier dataviz 25/04/23

Pour rappel, le dernier GT OPEN DATA PIGMA proposait d'intégrer un 1er niveau de service dataviz dans la plateforme PIGMA. La proposition du socle minimal à intégrer concernait l'intégration d'un outil dataviz dans le catalogue PIGMA.
Il a été décidé également lors de ce GT que les dataviz ne se feraient pas à la volée et ne pourraient être réalisées que par les administrateurs de la plateforme et éditeurs des organisations pour des raisons techniques et de connaissance des données.
A l'issue de ce GT il a été proposé de co-construire ce 1er niveau de service au sein d'un sous GT dataviz pour définir les spécifications et développements de cet outil dataviz.

Ce sous GT s'est tenu fin avril 2023 et a réuni 24 participants. Lors de sous GT, les participants ont été interrogés sur 3 points :

- **Le choix de l'intégration de l'outil dans la plateforme** : il s'est porté sur l'implémentation dans un nouvel onglet du jeu de donnée PIGMA (intitulé qui reste à déterminer). L'onglet dédié à la dataviz serait proposé en dernier onglet.

- **Le choix des graphiques à intégrer en priorité et par ordre de préférence** : anneau de répartition, chiffres clés/ zone à texte avec champ calculé, histogramme, courbe et carte statistique (corrélation indicateur avec un niveau géographique)
- **Le recueil des besoins et hiérarchisation des cas d'usages/ fonctionnalités** : ce recueil sous forme d'atelier collaboratif a fait l'objet d'une feuille de route

2.3.2. Feuille de route dataviz PIGMA

Suite aux recueils des besoins et hiérarchisation des cas d'usages/fonctionnalités une feuille de route a été définie. Cette feuille de route dresse les fonctionnalités du premier niveau de service ainsi qu'une hiérarchisation des fonctionnalités ressortie lors de l'atelier Dataviz d'avril 2023.

Fonctionnalités du premier niveau de service :

- Outil de création de dataviz accessible aux administrateurs et éditeurs des organisations de la plateforme
- Mise à disposition d'un applicatif dédié permettant la création de dataviz
- Possibilité de création de 1 à n dataviz par jeu de données affiché(es) dans un onglet dédié
- Mise en place d'un onglet dédié sur le catalogue PIGMA
- Possibilité de personnalisation des datavisualisations
- En cours d'étude : Création de datavisualisation en lien avec la gestion des droits de la plateforme

Hiérarchisation des fonctionnalités retenues dans le cadre de l'atelier participatif :

1. Rapports et portraits de territoire : Fonctionnalité donnant la possibilité de réaliser des datavisualisations permettant la mise en place de rapport et portraits à des échelles de territoires différentes, pouvoir les comparer et ceci sur plusieurs thématiques.

2. Partage de datavisualisation : fonctionnalité permettant de partager les datavisualisations sous différents formats : PDF, Image ou encore IFRAME pour une intégration sur des sites internet tiers.

3. Export de données : une datavisualisation nécessite souvent un filtrage, une sélection, … cette fonctionnalité permettrait un export de la donnée retravaillée utilisée dans une datavisualisation.

4. Croisement de données et analyses : fonctionnalité permettant de réaliser un croisement entre plusieurs données (multithématiques, référentiels, …) afin de répondre à un besoin spécifique de datavisualisation.

5. Outil de création de dataviz à la volée : fonctionnalité permettant une découverte à la volée d'un jeu de données directement depuis sa fiche d'informations

2.3.3. Présentation outil dataviz du SMEAG

Le Syndicat mixte d'Etudes et d'Aménagement de la Garonne (SMEAG), partenaire PIGMA et utilisateur également de la solution OneGeo Suite (sur laquelle repose la plateforme PIGMA) fait une brève présentation de son organisation avant d'aborder le projet de refonte de l'observatoire Garonne crée en 2015.

Le SMEAG a lancé un audit en 2022 pour sonder ses partenaires sur l'existant et pour recueillir les besoins en terme d'outil. Le déploiement de l'observatoire par le prestataire NEOGEO est prévu pour octobre/ novembre 2023.

L'observatoire a pour objectif d'avoir un parcours utilisateur le plus optimisé possible (professionnel comme utilisateur), il est structuré par thématique ou projet. Les indicateurs sont mis à jour annuellement et maîtrisés en interne.

Le SMEAG fait part d'une forte demande de portraits territoires avec des indicateurs à visualiser à partir d'un territoire.

Dans le cadre de cet atelier, le SMEAG fait une démonstration des outils dataviz de leur observatoire :

- **Tableaux de bord** : développé par ONEGEO, où il est possible de faire un choix par thématique et année + sous-thématique. Il y a différents graphiques disponibles : chiffre clé/ texte, image, cartographie directement intégrée et graphe. Les données brutes s'affichent au survol. La création de tableaux de bord est également simple (15/20 mn).

- **Cartographie :** il est possible faire un lien entre l'analyse de la donnée et sa représentation géographique. En cliquant sur un point un pop-up informations s'affiche avec l'implémentation dataviz présente dans le tableau de bord.

- **Portraits de territoire :** il n'y a pas de visuel disponible car est en cours de développement. Les utilisateurs pourront choisir le périmètre défini (syndicats de rivière, EPCI …) et comparer 2 territoires entre eux. Le SMEAG alerte sur le fait que c'est un travail redondant et chronophage pour les administrateurs de données. Sur ce point, le SMEAG souligne la nécessité de maîtriser des données notamment pour la découpe par territoire.

Le SMEAG conclut son intervention en indiquant que la dataviz donne un réel apport visuel pour les utilisateurs par rapport à l'ancien observatoire. La gestion de contenu est facilitée avec les logiciels FME/ Metabase/ OneGeo Suite. Il soulève également l'intérêt à récupérer les données dataviz au format image ou CSV.

3. Conclusion

L'équipe PIGMA conclut cet atelier sur le besoin prégnant de la qualité de la donnée, au cœur de tous les métiers et thématiques, sans laquelle il n'est pas possible de bien la réutiliser/ valoriser. Il est donc indispensable de travailler sur la qualité des données (mise à jour, structuration..) en amont.

La dataviz tient un rôle important dans la communication des données mais nécessite une qualité des indicateurs pour qu'elle soit cohérente et claire, un travail sur les données pour pouvoir les faire parler.

L'équipe PIGMA rappelle son rôle d'accompagnement et d'animation pour aider les producteurs de données à monter en qualité leurs jeux de données.

Enfin, en perspectives, PIGMA ouvre la question sur la combinaison des travaux de qualification en cours avec les nouvelles technologies qui arrivent (IA, script automatique) afin de répondre au mieux aux besoins de la plateforme et des utilisateurs.