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

COMPTE-RENDU DE REUNION

REF : SPP/22.0405 DATE : 24/02/2022 – Lieu : virtuel

Objet : Réunion du GT Métadonnées – Relance des travaux

Ordre du jour :

Présentation de la démarche au MTE qui vise notamment à construire un portail des données des politiques du pôle ministériel (Benoit David)

Présentation de l'investigation menée par Jailbreak (Johan Richer)

Discussions sur les travaux qui seraient à mener dans le cadre du GT CNIG Métadonnées

Principales conclusions :

Cf. Ci-dessous

Prochaine réunion :

29 mars de 14h à 17h

Retex Géo2France au sujet de DCAT (Benjamin Chartier)

Retex idéobfc sur les différents moissonnages entrepris (Jerome Boutet)

Présentation autour des évolutions du Géocatalogue (Gregory Delobelle)

La réunion suivante sera à fixer.

Liste de diffusion :

Participants – Organisme / Service		Personnes à informer – Organisme / Service	
Alix Marc	Ville de Montpellier	Adeline Souf	Shom
Arnauld Gallais	CEREMA	Amandine Thomas	Ifremer
Benoit David	CGDD/SRI/ecolab	Dimitri Meunier	OIEAU
Carlo Mouzannar	Isogéo	Berenice Lequesne	Shom
Clement Jaquemet	CGDD/SRI/ecolab	Fanny Lecuy	SHOM
Benjamin Chartier	Bchartier	David Viglietti	OIEAU
Grégory Delobelle	BRGM	Erwann Quimbert	Ifremer
Guillaume Ryckelynck	Région Grand Est	Julien Meillon	Ifremer
Jerome Boutet	IDéO BFC	Guillaume Grech	MNHN
Léo Darengosse	Isogéo	Laurène Debray	OIEAU
Leslie Lemaire	SG/SNUM/UNI/DRC	Thierry Vilmus	BRGM
Sébastien Dias	Gipatgeri	Maël Reboux	Rennes Métropole
Noël Cellarier	DGALN	Maria Tomanov	Région Auvergne Rhone Alpes
Robert Rivière	Service du Numérique du MTE	Magali Pessidou	DGFIP
		Mathieu Becker	Isogéo
		Vincent Gomand	DGFIP
		Sébastien Léger	DGFIP

Modèle TN-02/020 - 1.4

Yvan Le Bras	MNHN
Marion Lacroix	IGN
Marie Lambois	IGN
Johan Richer	Jailbreak

Date	Visa	Nom	Service
25/02/2022		Marie Lambois	IGN

Décisions et minutes

Présentation de la démarche au MTE qui vise notamment à construire un portail des données des politiques du pôle ministériel (Benoit David)

Présentation

L'objectif principal de la construction de ce portail est de faciliter l'accès aux données sur les politiques du Ministère. Le périmètre prévu pour ce portail couvre d'abord l'administration centrale, les services déconcentrés du pôle ministériel et les DDT(M). Il pourra ensuite être étendu aux opérateurs.

On peut utiliser le terme de « guichet » pour le décrire car il consiste en une agrégation des catalogues existant en un point unique. On peut utiliser le terme de portail d'accès aux données plus que catalogue (car il n'y a pas de saisie de métadonnées directement sur l'outil).

L'outil développé sera basé sur CKAN.

Un des aspects connexes de ce projet est l'amélioration progressive de la qualité des métadonnées, cela passera par l'animation d'un réseau d'acteurs.

Un thésaurus est utilisé, l' « arborescence COVADIS » qui recense les politiques publiques. Ce thésaurus pourrait être maintenu dans le cadre du GT métadonnées.

Discussions/Questions

Questions de G. Ryckelynck et réponses

- Est-ce un catalogue de plus ? - *B. David : On ne crée pas un nouveau catalogue (saisie de données) mais un portail (moissonnage et réponse à la recherche de données uniquement). C'est dans ce contexte qu'il est important d'éviter la confusion entre portail et catalogue. Il n'existe pas tellement de portails : data.gouv.fr, etc. et il y a un besoin avéré d'être en capacité de mieux trouver les métadonnées.*

- Y a-t-il eu un benchmark des outils ? Serait-il diffusable ? Il se dit surpris du choix de CKAN et pas de Géonetwork. CKAN, qu'ils utilisent, leur paraît trop compliqué, ils espèrent pouvoir l'abandonner prochainement – *Aucun benchmark n'a été formalisé. Le portail se veut différent et plus large que le périmètre de l'information géographique. CKAN permet d'aller vers un portail générique plus englobant. Il y a actuellement beaucoup d'investissement sur CKAN, en particulier au niveau du portail européen, même si l'outil reste complexe à utiliser.*

Questions de B. Chartier et réponses :

- Il considère que CKAN est plutôt en disgrâce au niveau des plateformes de données géographiques. Il est mentionné dans la présentation un besoin de structuration des métadonnées dans CKAN. Est-ce que l'idée est de construire un nouveau standard ? - *CKAN est assez souple en termes de structuration de métadonnées. Dans la mesure où plusieurs acteurs utilisent CKAN, cela a du sens d'avoir une structuration partagée*

- B. Chartier confirme que CKAN est très souple en termes de modélisation ce qui peut entraîner des défauts d'interopérabilité car chaque instance de CKAN peut implémenter son propre modèle de métadonnées (B. David tout à fait d'accord). Il est nécessaire de standardiser une structuration CKAN des métadonnées.

- La présentation mentionne un catalogue de métadonnées statistiques, Dido. L'objectif est-il de mettre en œuvre stat-DCAT-AP ou SDMX de façon plus spécifique ? - *Stat-DACT-AP n'a pas été particulièrement analysé. Il n'y a à priori pas de besoin particulier pour exploiter stat-DCAT-AP.* [en complément : l'unique catalogue de données statistiques inclus dans le périmètre à ce stade utilise DCAT-AP et non StatDCAT-AP, donc la question ne s'est simplement pas posée. Idem pour SDMX.]

- Le moissonnage de CSW, ISO implique-t-il la mise au point de règles de correspondances ? – *oui, le projet est en train de faire une mise en correspondance INSPIRE/DCAT-AP.*

- Cela ne devrait-il pas être fait dans le cadre du GT CNIG MD ? - *Oui, sera fait en coordination avec le CNIG.*

Questions et remarques de S. Dias :

- Quels sont les liens avec la Géoplateforme ? - *Le lien effectivement n'est pas établi pour l'instant. De toute façon le guichet sera un acquis exploitable par la Géoplateforme.*

- Retours concernant les difficultés avec CKAN :

1. Problème de moissonnage CSW
2. Problème de maintenance du cœur de CKAN (longtemps resté sur la version 2 de Python - obsolète)

   → *La question de la version 2 a été réglée pour le cœur de CKAN (la version de CKAN utilisée pour ce projet est la version 2.9). Toutes les extensions n'ont pas été portées en Python 3, mais les extensions du dépôt officiel oui (ckanext-spatial, ckanext-dcat...) et les fonctionnalités nécessaires pour le projet de guichet sont bien disponibles aujourd'hui.*
3. Ne semble pas adapté à la diffusion avec droits (données non ouvertes)

   → *Sur ce point, ce n'est pas gênant pour le guichet car il y a la volonté d'avoir des métadonnées ouvertes même pour des données non ouvertes.*

J.Boutet suggère de réinventer le Thésaurus thématique. Un travail collectif à ce sujet serait nécessaire.

M. Lambois confirme que c'est un chantier dont le GT MD (pour les thèmes) et le GT QDG (pour les usages de l'IG) peuvent se saisir.

B. David précise que le Ministère utilise déjà pas mal le "Thésaurus Covadis", les travaux devront donc être menés sur cette base.

G. Ryckelynck met en garde sur le fait que les thèmes pour saisir les MD (l'expert) ne sont pas nécessairement ceux recherchés par l'utilisateur (lambda). B. David précise que la démarche d'UX Design qui sera mise en œuvre pour le guichet pourra fournir des éléments sur les besoins utilisateurs.

J. Boutet se montre intéressé pour réutiliser les travaux sur CKAN menés dans le cadre du guichet.

Présentation de l'investigation menée par Jailbreak (Johan Richer)

Liens utiles

Synthèse de l'investigation (octobre 2021) : https://jailbreak.gitlab.io/investigation-catalogue/synthese.html

Réflexion sur un "schéma commun" : https://github.com/etalab/schema-catalogue-donnees

Développement démarré en décembre 2021 : https://github.com/etalab/catalogage-donnees

Backlog (MVP) : https://github.com/etalab/catalogage-donnees/projects/1

Enregistrement d'une discussion autour du "catalogage de données" à l'initiative notamment de Géobretagne (28/01/2022) : https://peertube.virtual-assembly.org/videos/watch/cc586e34-57ac-475d-b13e-b15a62909fb5

Présentation

La présentation porte sur l'étude menée en amont de catalogue.data.gouv.fr. Cet outil permettrait de venir en support de data.gouv.fr, permettant la saisie des métadonnées, sans que les données soient nécessairement accessibles.

La définition donnée de catalogue versus portail diffère ici. Un catalogue se définit comme hébergeant juste les métadonnées (pas la donnée en tant que telle), tandis qu'un portail peut également héberger la donnée.

La plupart des Ministères ont actuellement des portails qui ne sont pas conçus pour le catalogage (pour lequel on crée des outils spécifiques, dédiés et déconnectés). Cet outil vient donc combler ce manque.

Le schéma du catalogue utilisé est très proche de DCAT-AP.

La traçabilité de la donnée (qui contacter ? Comment avoir accès à la donnée) est un élément primordial. Les investigations montrent un manque de traçabilité de la donnée dans les informations actuellement sur data.gouv.fr (= quelles sont les données "vérité" ?).

La maintenabilité est également problématique. Les métadonnées sont initialisées mais trop rarement maintenues.

In fine, le critère pour l'utilisateur (qui implique néanmoins que les autres critères soient satisfaits) est la découvrabilité : trouver quelles données existent, notamment via les moteurs de recherche (généralistes ou de portails).

Discussions/Questions

S. Dias partage la vision sur la distinction entre portail et catalogue. Le catalogue est un outil. Il y a eu des évolutions sur les termes, il serait intéressant de partager/harmoniser cette vision au niveau national. La liaison avec les moteurs de recherches est évidente et actuelle.

Y. Le Bras mentionne les entrepôts de données qui sont aussi à intégrer dans un triptyque avec portail et catalogue. N'est-ce pas simplement qu'une question/confusion entre backoffice et frontoffice ? – *Johan Richer : Un entrepôt intègre la notion de stockage / d'infrastructure, par exemple Dataverse. Un portail = entrepôt + catalogue. La notion de stockage est peu importante sur l'aspect catalogue. Côté backoffice il faut savoir adapter l'entrepôt. Mais du côté frontoffice c'est neutre.*

Les infrastructures pour les données ET pour les MD doivent être adaptées aux différents métiers. Ex : différence entre l'information géographique et données tabulaires : la géomatique et la recherche sont des métiers à part entière avec des besoins de portails et d'entrepôts différents. Bref, les MD doivent être adaptées au métier et à l'usage considéré.

Un portail unique, un dénominateur commun, c'est faire rentrer des ronds dans des carrés, la centralisation entraîne de la perte. Dans data.gouv on voit que le moissonnage pose problème. Dans le cas des données économiques, sur 80 types de données, 80% sont hétérogènes, donc il faut faire du spécifique. Le problème de l'hétérogénéité est résolu au prix de coûts énormes, beaucoup de complexité à essayer de centraliser et harmoniser ce qui n'a pas forcément vocation à l'être.

C. Jaquemet demande : on veille à ce que les MD soient bien "intégrables" par les moteurs de recherche. Comment enrichir les automatismes des moteurs de recherche ? Est-il nécessaire d'avoir un noyau/cœur des MD garantissant le passage d'un modèle à un autre (cf. les travaux des GT CNIG). Mais quid des cartes, dataviz, traitements ? Tout cela mérite réflexion au sein du GT MD.

B. David précise qu'un moteur de recherche généraliste comme Google est adapté au grand public mais moins adapté à des utilisations professionnelles capables de spécialiser leur recherche.

J. Richer est d'accord avec cette analyse. Il ajoute qu'il souscrit aux propos de B. David sur CKAN en tant que meilleure voie pour un portail générique (cela lui semble être ce

qui se fait de mieux même si GeoNetwork fonctionne très bien aussi) avec un gros travail sur le fait de pouvoir être exploité par les moteurs de recherche.

La migration vers Python3 a été évoquée, mais il reste dans CKAN à mieux différencier back et front office. Il y a une évolution vers le "Data Management" comme nouvelle logique de l'open data.

J. Richer a travaillé avec ODF sur schema.data.gouv.fr dans une logique de MD lisibles par les humains.

J. Boutet mentionne la maturité des organisations (entre autre la présence d'administrateur de données) et demande si cet aspect organisationnel est pris en compte dans le projet de la DINUM.

J. Richer approuve et remercie pour cette approche, ainsi que salue le travail technique côté Info Géo et IDG Régionales.

La traçabilité doit être la plus granulaire possible (nom du personne contact, contact mail, etc.), en tout cas mieux que ce que l'on trouve actuellement sur data.gouv.fr. Ce serait possible sur catalogue.data.gouv.fr avec des catalogues non téléchargeables. Mais J. Boutet met en garde sur le fait que la donnée doit être publique dans une fiche de métadonnées elle aussi publique. Raison pour laquelle il est compliqué d'intégrer des informations de contact, qui en aval amèneront du travail donc des moyens supplémentaires. L. Lemaire indique que c'est également une question de culture d'entreprise à faire évoluer. Ce n'était pas dans les habitudes des administrations de rompre avec le côté boîte noire d'une manière générale, et ça a mis du temps aussi à évoluer dans la sphère géomatique.

J. Richer souligne l'unicité dans la géomatique entre la personne qui produit les données, et celle qui saisit les métadonnées, ce qui représente un cas assez particulier. Pour lui c'est un problème à résoudre dans data.gouv.fr : les rubriques existent et sont claires mais les agents n'arrivent pas à saisir tous les champs ! Questionnement sur l'utilisateur des adresses sur data.gouv.fr, l'intérêt étant le flux basé sur le schéma commun base adresse local (BAL), qui est réutilisé par Google. Ce processus est invisible et c'est tant mieux. Mais pour cela il faut mettre en place un certains nombres de systèmes, schémas communs, cadre légal etc. et sur certains sujets il faut partir de rien.

B. David signale la démarche de concertation et de rapprochement entre la standardisation CNIG et schema.data.gouv. Cette démarche est à l'ordre du jour de la prochaine Commission Règles et Qualité du CNIG, du 31 mars.

J. Richier signale également ouverture.data.gouv.fr qui présente une feuille de route sur la publication des données.

L. Darengosse demande quelle est la stratégie pour l'indexation des données et s'il y aura des connecteurs pour automatiser la création des fiches ? - *Certains ministères ont déjà des outils (très basiques), on ne va pas imposer aux utilisateurs un nouvel outil de catalogage que s'il leur convient. Il imagine un bouton "importer mon jeu de données" dans catalogue.data.gouv.fr, mais il y aura aussi des API permettant d'importer d'autres outils de catalogage présents dans les ministères. Mais le MTE n'aura pas besoin de l'outil. Le cas d'usage de niv 0 : ce sont les ministères qui n'ont pas d'outil, pas d'automatisation, qui utilise des outils tabulaires basique Airtable, excel, qui ne sont pas branché dans un flux au service de la politique de la donnée. Le nouvel outil leur permettra d'être plus efficaces par rapport à ce cas d'usage niveau 0. Mais dans le cas du MTE cela va s'insérer dans une chaîne déjà bien construite. Il faut juste éviter les pertes existantes actuellement car un moissonnage signifie souvent une perte.*

Les ministères pourraient avoir besoin d'extensions à DCAT-AP (exemple de deux champs "personne contact", "généalogie du jeu de données = d'où vient le jeu de données ? ")

Cas d'usage identifié dans l'investigation : https://jailbreak.gitlab.io/investigation-catalogue/synthese.html#/19/5

M. Lambois demande si des Thésaurus inter-ministériels ont été développés. - *Le parti pris est d'utiliser des tags plutôt que des thésaurus, plutôt une logique de recherche sémantique pour réduire la saisie de métadonnées. Le besoin d'ontologie est très variable (et peut être très spécifique) suivant les ministères.*

L. Lemaire indique les principaux vocabulaires contrôlés utilisés par DCAT-AP : https://op.europa.eu/en/web/eu-vocabularies/authority-tables

Discussions sur les travaux qui seraient à mener dans le cadre du GT CNIG Métadonnées

Pistes de points à traiter (nouveaux ou précisés) par le GT Métadonnées

- Mapping DCAT / Inspire : voir les points durs et limiter les pertes lors des moissonnages

- Besoin de travailler sur les thésaurus. Faire évoluer ceux de la COVADIS, et adapter/traduire ceux de GéoDCAT-AP

- Définir une structure socle commun CKAN, qui serait également un méta-modèle

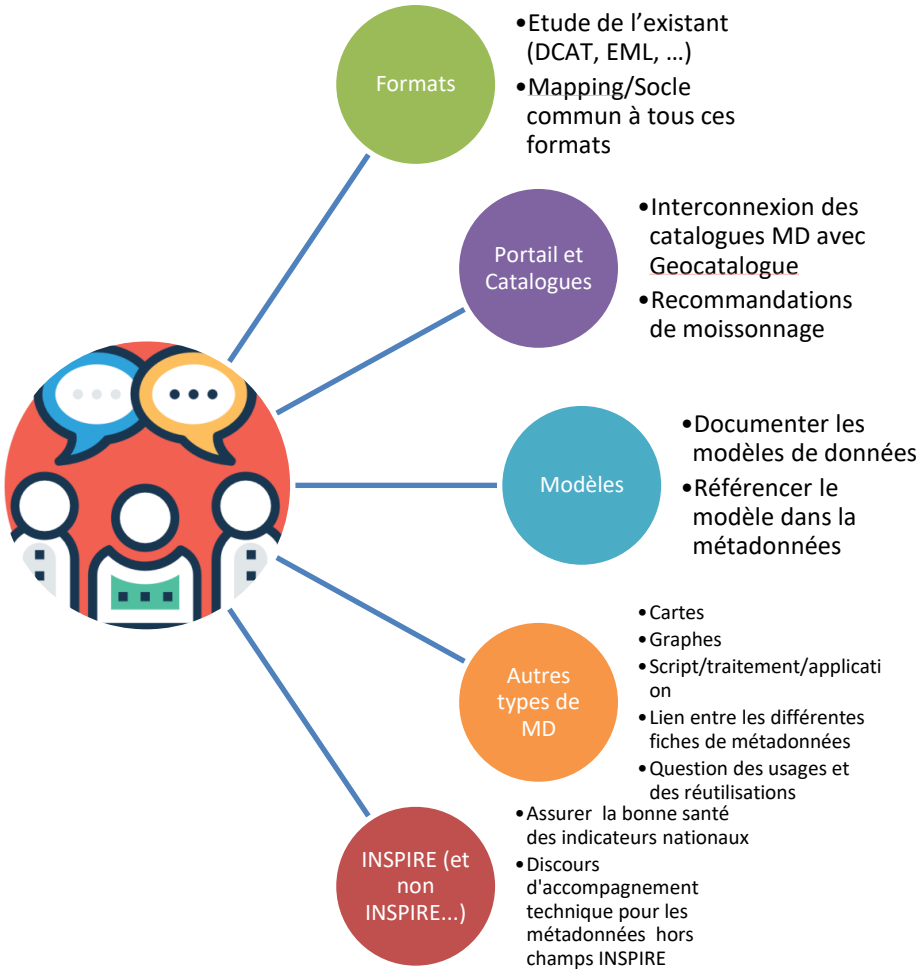
- Faciliter l'indexation / le référencement par les moteurs de recherche (rejoint les points précédents)

- Continuer le retour et partage d'expérience

- Rapprochement entre les standards CNIG et schema.data.gouv (ce sujet sera évoqué lors de la commission Règles et qualité du 31 mars) : c'est un besoin transverse qui concerne a minima les GT Métadonnées et Qualité.

- Le rapportage Inspire 2021 sera également évoqué à la commission Règles et qualité du 31 mars, qui déterminera si des actions de la part du GT Métadonnées sont nécessaires.

Ces points viendront s'inscrire dans le programme de travail qui a déjà commencé à être élaboré :



Toutes ces actions sont ambitieuses et la prochaine réunion devrait permettre de déterminer ce qui sera fait dans l'année.

B. Chartier soulève également le besoin d'intégrer davantage d'éditeurs de logiciels / des utilisateurs /etc., par exemple de diversifier vers ESRI, Alkante, Business Geografic, Opendatasoft, etc. Le besoin pourra si nécessaire être remonté le besoin au conseil plénier du CNIG.