La standardisation des données ouvertes : favoriser l'interopérabilité, accroître l'impact de l'open data

Samuel Goëta, Elise Ho-Pun-Cheung

02

LES CAHIERS DE L'OBSERVATOIRE

LA STANDARDISATION DES DONNÉES OUVERTES : FAVORISER L'INTEROPÉRABILITÉ, ACCROÎTRE L'IMPACT DE I'OPEN DATA

DataPublica

Ce cahier a été préparé par l'Observatoire Data Publica et a été publié en novembre 2022. Ses auteurs principaux sont Samuel Goëta & Elise Ho-Pun-Cheung avec l'appui éditorial de Aurélie Legrand, Jacques Priol, Joël Gombin, Mathieu Morey, Justine Banuls, Anne-Laure Donzel, Clément Feyt, Diane Thierry & Sylvain Lapoix.

Conception & réalisation graphique : fact260.fr

La publication des cahiers de l'Observatoire Data Publica bénéficie du soutien de la Banque des territoires.

BANQUE des TERRITOIRES | Caisse des Dépôts GROUPE

LA STANDARDISATION DES DONNÉES OUVERTES : FAVORISER L'INTEROPÉRABILITÉ, ACCROÎTRE L'IMPACT DE L'OPEN DATA

INTRODUCTION

UNE ÉTUDE DES STANDARDS DE DONNÉES OUVERTES

Depuis le 7 octobre 2016, l'ouverture des données est devenue la règle pour l'ensemble des acteurs investis d'une mission de service public disposant de plus de 50 agents et pour les territoires de plus de 3500 habitants. L'Observatoire Open Data des Territoires de l'association OpenDataFrance montre de fortes disparités entre les territoires dans la mise en œuvre de ce principe. Par exemple, en 2021, 59,2 % des communes et EPCI de plus de 100 000 habitants ont enclenché une démarche d'open data. Ce chiffre tombe à 8,7 % pour les EPCI de moins de 100 000 habitants et à 8,4 % pour les communes entre 3500 habitants et 100 000 habitants[1].
Ces différences entre les territoires se cumulent à une hétérogénéité dans les données publiées. La loi impose un principe d'ouverture généralisée dans le cadre prévu par le code des relations entre le public et l'administration mais chaque territoire publie les données selon ses compétences, son patrimoine de données et ses pratiques. Ainsi, les données ne sont pas nommées de la même manière selon les territoires. Sur la cyclabilité par exemple, il faudra alternativement chercher « aménagements cyclables » ou « pistes cyclables » et on retrouvera rarement le mot clef « vélo » dans les descriptions des jeux de données alors que le terme vient spontanément sur le sujet. En plus des différences de terminologie entre collectivités, il existe plus généralement un décalage (un *vocabulary mismatch*) entre les producteurs, qui publient des documents avec leur propre vocabulaire, et des utilisateurs formulant leur besoin avec un autre. De la même manière que le terme « vélo » induit rarement à une recherche fructueuse, alors que l'utilisateur cherchera plus volontiers le terme « rond-point » c'est pourtant sous l'appellation « giratoire » qu'il trouvera les jeux de données les plus pertinents.

1. Résultats de l'enquête 2021 de l'observatoire open data des territoires.

DES DONNÉES OUVERTES DIFFÉRENTES D'UN TERRITOIRE À L'AUTRE

À cet enjeu de découvrabilité s'ajoute un défi lié à la normalisation des données ouvertes. D'un producteur à l'autre, les fichiers ne contiennent pas nécessairement les mêmes champs ou ne donnent pas le même niveau de détail. Les valeurs dans les champs eux-mêmes ne sont pas normalisées. Prenons l'exemple des menus de cantine. Certains territoires publient des jeux de données descendant jusqu'aux ingrédients des recettes, là où d'autres n'indiquent que l'intitulé des plats sans informations sur la provenance ou le caractère végétarien ou non du plat. Si nous voulions comparer l'offre de cantine de ces collectivités, un important travail d'harmonisation des données serait nécessaire. C'est pour répondre à ce manque d'interopérabilité des données que l'association OpenDataFrance a proposé de décrire les menus proposés par les collectivités locales ou les syndicats mixtes de restauration par le biais d'un standard.

UNE LIMITATION DE L'IMPACT DE L'OPEN DATA

Concrètement, ces enjeux de découvrabilité et de normalisation des données limitent l'impact de l'open data. Sans harmonisation des pratiques, il est très compliqué de construire des services ou des usages qui dépassent un seul territoire. Par exemple, l'application Handimap, qui propose des itinéraires accessibles aux personnes à mobilité réduite en tenant compte des trottoirs surbaissés, a ainsi été entravée dans son développement par l'absence de normalisation des données sur l'accessibilité de la voirie. Chaque nouvelle instance locale de l'application nécessitait un développement conséquent pour s'adapter aux données du territoire. Alors que l'utilité de cette application est indéniable pour les personnes à mobilité réduite, ce service n'a pas pu se développer au-delà d'un nombre restreint de territoires (Rennes, Montpellier, Nice, La Rochelle, Lorient)[2].

La standardisation des données réduit donc les frictions en facilitant la découverte de données similaires ouvertes dans différents territoires et en permettant de consolider les données produites localement dans une base nationale exploitable facilement. Les standards, en tant qu'ensemble de recommandations préconisées par un groupe d'utilisateurs, permettent à différents outils numériques de communiquer pour construire un tout cohérent au service d'un ou plusieurs objectifs.

Ce cahier a pour objectif de présenter un état des lieux de la standardisation des données ouvertes, en portant une attention particulière aux défis liés à leur conception et à leur réutilisation, notamment par les collectivités productrices de données.
À travers ce cahier nous cherchons notamment à montrer que les standards de données ouvertes, en tant qu'outil d'échange, peuvent être au service d'un projet collectif : développer la transparence de l'action publique, permettre une mobilité plus inclusive, ou encore orienter plus efficacement les publics visés par des dispositifs d'insertion.

2. Le standard sur l'accessibilité du cheminement en voirie, publié fin 2021 suite à de nombreuses réunions d'un groupe de travail du Conseil national de l'information géographique (CNIG), devrait faciliter le déploiement national de ces usages.

L'IMPORTANCE CROISSANTE DE LA STANDARDISATION

UN MONDE DE STANDARDS

LES standards ne sont pas propres au monde de l'open data, ni même à celui des données de manière générale. Normes socio-techniques destinées à harmoniser les pratiques et usages d'un ou plusieurs secteurs d'activité, à produire une uniformité par l'application de règles, ils sont omniprésents au quotidien[3]. Dans les premières lignes d'un ouvrage arguant l'intérêt de prendre ces standards comme objet d'étude, le sociologue Lawrence Busch dresse ainsi le constat que les standards peuplent notre quotidien sans qu'on ne les voie :

> *« […] On peut lire le journal presque tous les jours et trouver des articles sur les standards — des standards pour les personnes, pour l'environnement, pour les produits de consommation, pour le bien-être animal, pour la comptabilité des finances publiques, pour la pression acceptable sur les ponts routiers, pour les soins de santé, pour l'éducation, pour à peu près tout »*[4]

Car ils permettent d'unifier, de distinguer, de hiérarchiser des catégories ou encore de formaliser et de distinguer des pratiques[5], les standards ont un impact concret sur la vie sociale. Ils forment un ensemble composite allant de normes « rigides » attestant par exemple de l'implémentation de conditions de travail sécurisées (par exemple, norme ISO45000 sur la santé et la sécurité au travail) jusqu'à des pratiques codifiées mais adaptables encadrant la conduite des politiques publiques (les standards de « bon gouvernement »). De nombreux travaux académiques ont étudié un ou des standards dans un domaine bien déterminé : la formation en ligne[6], la finance[7], l'information

3. Voir par exemple cet article, dont nous empruntons le titre : Timmermans S., Epstein S., 2010, « A world of standards but not a standard world: Toward a sociology of standards and standardization », *Annual Review of Sociology*, *36*, p. 69-89.
4. Busch L., 2011, *Standards: Recipes for Reality*, Cambridge, MA, MIT Press (Infrastructures), p. ix (nous traduisons).
5. Les sociologues américains Geoffrey Bowker et Susan Star estiment ainsi que les standards sont centraux dans le processus de classification du monde. Voir notamment : Bowker G.C., Star S.L., 1999, *Sorting Things Out: Classification and Its Consequences*, Cambridge, MA, MIT Press (Inside Technology), 392p.
6. Grandbastien M., 2004, « Premiers pas dans le monde des standards pour la formation en ligne », *Distances et savoirs*, *2*, 4, p. 395-408.
7. Berner R., Judge K., 2019, « The Data Standardization Challenge », SSRN Scholarly Paper, Rochester, NY, Social Science Research Network.

géographique[8], ou encore l'agroalimentaire[9], pour n'en citer que quelques-uns. Ces recherches en sciences sociales, tout en montrant la particularité de chaque cas, tendent vers quelques constats communs :

- Les standards ne vont pas de soi, ils sont socialement construits et font l'objet de négociations entre des parties prenantes parfois nombreuses ;
- Leur adoption n'est pas acquise d'avance ;
- Ils sont inégalement adaptés aux individus, en fonction notamment des compétences de ces derniers — un standard peut être adapté à une personne tout en étant le « cauchemar » d'une autre[10] ;
- Une fois adoptés, les faire évoluer nécessite souvent un travail conséquent ;
- En dévier peut mener à marginaliser un individu au sein d'une communauté/d'un monde social.

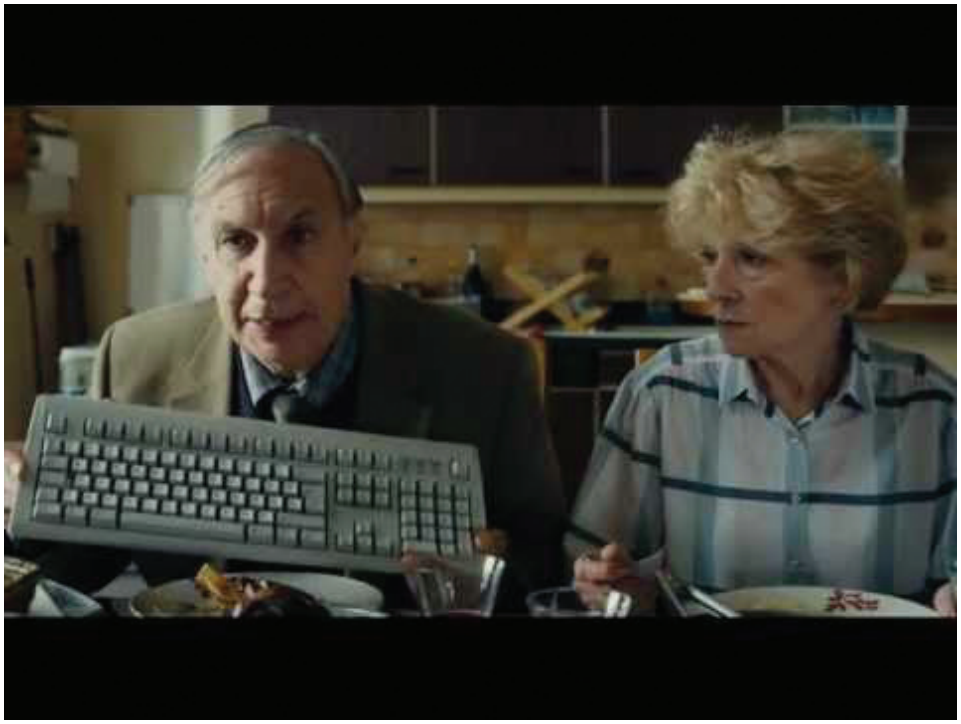
La sociologue Susan Leigh Star résume ainsi la difficulté de faire évoluer une norme et de s'en distancer :

> « *[…] une fois que les arrangements deviennent standards dans une communauté, créer des standards alternatifs peut être coûteux ou impossible, à moins qu'une communauté alternative se développe. […] Il est coûteux de travailler dans un monde [social] et d'exercer en dehors de cet ensemble de standards* ; [11]

La norme française faisant du point-virgule le séparateur de colonnes dans un fichier CSV plutôt que la virgule (celle-ci étant le séparateur décimal) illustre bien le coût de l'écart à un standard. Cette alternative à une règle de formatage internationalement acceptable n'est en effet pas prise en compte par un grand nombre d'outils d'analyse et de visualisation de données.

---

8. Auguste Atemezing G., Troncy R., 2012, « Vers une meilleure interopérabilité des données géographiques françaises sur le Web de données », *23es Journées Francophones d'Ingénierie des Connaissances (IC 2012)*, p. 369-384.
9. Fouilleux È., 2010, « Standards volontaires. Entre internationalisation et privatisation des politiques agricoles », dans *Les mondes agricoles en politique*, Paris, Presses de Sciences Po (Académique), p. 371-396.
10. Sur cet aspect, voir le début de l'introduction de l'ouvrage collectif coordonné par Susan Star et Martha Lampland : Star S.L., Lampland M., 2009, « Reckoning with standards », « *Classifying and Formalizing Practices Shape Everyday Life* », p. 2-25.
11. Star S.L., 1990, « Power, Technology and the Phenomenology of Conventions: On being Allergic to Onions », *The Sociological Review*, 38, 1, p. 41 (nous traduisons).



Dans le film Le Nom des gens, *les Martin regrettent la persistance du clavier AZERTY, norme incontestée aujourd'hui, malgré la plus grande ergonomie du clavier BEPO.*

standardiser les catégories statistiques

Le domaine des statistiques est fortement traversé par des enjeux liés à la standardisation puisque la représentation des populations et des territoires repose sur des catégories et des indicateurs devant être standardisés afin d'être comparables.

Dans le domaine de la statistique urbaine, les chercheurs s'interrogent de longue date sur les enjeux et sur le processus de création de nouveaux indicateurs. Dès 1970, Ira Rosenwaike proposait par exemple un retour critique sur les Standard Metropolitain Statistical Areas (SMSAs), standards statistiques permettant en théorie de rendre compte de la « vie quotidienne économique et sociale » des territoires mais n'étant en pratique pas systématiquement utilisés par l'administration fédérale en charge de l'allocation des budgets[12]. Il s'agit donc d'un exemple où l'utilisation d'un standard dévie dans la pratique des usages imaginés par les concepteurs.

Dans la deuxième moitié du vingtième siècle, de nombreuses catégories statistiques ont été créées sous l'impulsion d'une dynamique de standardisation européenne, voire internationale. Soustraire les spécificités des contextes nationaux au profit de critères relativement simples devait en effet permettre la comparaison internationale. L'affirmation du périurbain comme catégorie statistique en France s'inscrit par exemple au croisement des injonctions de comparaison internationale croissantes dans les années 1960-1970 et d'un mouvement d'objectivation des aires géographiques — autrement dit de définition de ce qui constitue une ville (mouvement amorcé avec la naissance de l'INSEE en 1948).

Cette tentative d'objectivation statistique n'est pas propre aux questions urbaines. L'outil « état 4001 », qui est aujourd'hui le principal indicateur de la délinquance en France, est un autre exemple de catégorisation à l'impact très concret : développé en 1972 pour mesurer l'efficacité des services de police, il prend en compte de plus en plus d'infractions à partir des années 1990. L'évolution de la catégorie induit une hausse des statistiques de la délinquance et participe au développement du « sentiment d'insécurité ». Rendant-compte du plus grand nombre de délits entrant dans l'indicateur et non pas d'une croissance du nombre des délinquants, cet « état 4001 » montre bien l'impact que peut avoir la construction d'un référentiel sur la perception du réel[13].

12. Rosenwaike I., 1970, « A Critical Examination of the Designation of Standard Metropolitan Statistical Areas », *Social Forces*, 48, 3, p. 322-333.
13. Voir l'épisode « Insécurité : peur sur les chiffres » de la chaîne Youtube Data Gueule : https://www.youtube.com/watch?v=bMZ4YVrIY_s

La recherche de standardisation s'observe également dans le domaine du travail. Dès sa création, en 1919, le Bureau international du travail (BIT) se donnait pour objectif de standardiser les normes de travail et la statistique internationale, par exemple en promouvant une définition consensuelle du chômage[14]. En 1934, 1943 et 1959 le BIT publie ainsi trois synthèses concernant son travail de standardisation internationale des statistiques relatives au travail[15].
L'homogénéisation des critères de classement de catégories (géographiques, sociales…) et de calcul d'indicateurs est donc un enjeu identifié depuis longtemps. Récemment, la coexistence de différentes méthodes de mesure du taux de mortalité, dans un contexte de pandémie de COVID-19, a rappelé l'importance des conventions partagées à l'international dans le recueil de données statistiques, ici de santé, incomparables en l'absence de standards[16].

Les standards du web : 3 questions à Sylvain Le Bon

Dans le domaine de l'informatique, un ensemble de standards, notamment promus par l'organisme de standardisation à but non lucratif W3C, a été essentiel au développement du world wide web dans les années 1990.
Sylvain Le Bon est co-fondateur de la coopérative Startin'blox, qui a développé une technologie open source basée sur des standards du W3C, permettant la création d'applications pour les data spaces. Il nous apporte un éclairage sur les origines et les évolutions des standards du web.

Pouvez-vous revenir sur les origines du web et de ses standards ?

On est à la fin des années 1980 et Tim Berners-Lee est ingénieur au Cern, à un moment où tous les chercheurs utilisent l'informatique mais où l'échange de données est assez complexe. C'est un spécialiste de la donnée qui a bien conscience qu'il y a des données partout mais qu'elles sont très difficiles d'accès. Il a alors une idée assez floue, celle d'un world wide web, une toile mondiale d'information. Il l'évoque dans une demande de financement, à la marge de laquelle son chef, Mike Sendall, indique « *vague but exciting* ».

14. Sauthier I., 2009, « Histoire de la définition du chômage », *Courrier des statistiques*, 127, p. 5-12.
15. Kévonian D., 2008, « La légitimation par l'expertise : le Bureau international du travail et la statistique internationale », *Les cahiers Irice*, 2, 2, p. 81-106.
16. Leon D.A., Shkolnikov V.M., Smeeth L., Magnus P., Pechholdová M., Jarvis C.I., 2020, « COVID-19: a need for real-time monitoring of weekly excess deaths », *Lancet*, 395, 10234, p. e81.

Ce qu'invente Tim Berners-Lee c'est techniquement peu extraordinaire : il fait un langage de balises un peu simplifié pour qu'il soit facile à adopter et un protocole assez basique pour qu'un logiciel puisse accéder à des documents. Mais ce qu'il fait surtout, c'est formaliser d'une part l'URL, donc le fait qu'un document donné ait une adresse universelle, et d'autre part l'hypertexte, le fait que les textes peuvent contenir un lien vers d'autres documents. C'est ça qui crée le world wide web : c'est le fait que les documents puissent se référencer entre eux, en se reposant sur des standards ouverts. Cette ouverture des standards a permis au web de s'imposer au début des années 1990 par rapport à d'autres alternatives, désormais complètement oubliées. Le world wide web a donc vite pris dans le monde des chercheurs, pour qui c'était un espace documentaire commun.
Et assez rapidement le W3C (World Wide Web Consortium) a été créé, qui a porté la responsabilité de ces nouveaux standards. Il a notamment standardisé les langages informatiques HTML et CSS.

On entend parfois parler de web sémantique, ou web des données, qui serait la suite du world wide web des années 1990. De quoi s'agit-il ?

Tim Berners-Lee a continué à travailler à la standardisation des données et ce travail a donné naissance au web sémantique. Il repose sur une standardisation de la représentation des données. Toutes les données y sont décrites sous forme de triplet : sujet, prédicat, objet. Ces sujets, prédicats et certains objets ont la particularité d'être référencés par un identifiant qui est une URL universelle. Ça permet de représenter de façon unique et universelle une information. On a donc un espace de données où la donnée a une valeur universelle, indépendamment de celui qui la stocke. Le web devient donc une immense base de données.
Le standard clef est RDF (Resource Description Framework). Sorti en 2004, il permet de décrire toutes les données web. Mais ça ne suffit pas d'utiliser un identifiant universel. Ça ne règle pas la question de comment s'authentifier sur un serveur, comment structurer des données au sein d'un serveur ou comment faire des requêtes. Il y a donc des dizaines de standards qui ont été publiés depuis 2004 par le W3C. Toute cette famille de standards est regroupée sous l'appellation SOLID, qui permet de mettre en œuvre un web décentralisé, où les utilisateurs ont le contrôle sur leurs données, et où les applications peuvent accéder aux données où qu'elles soient pour en tirer toute la valeur.

La technologie de Startin'blox repose notamment sur l'ensemble de standards SOLID. Pouvez-vous nous en dire quelques mots ?

Le travail du W3C a abouti en 2016 à SOLID. À partir du moment où on se met d'accord pour structurer des données dans le respect de ces standards, on voit le web différemment : on a d'un côté des espaces de stockage des données, de l'autre des applications qui peuvent y accéder. N'importe quelle application web peut accéder à n'importe quelle donnée sur n'importe quel serveur. SOLID concerne à la fois la structure des données, leur stockage, leur authentification pour y accéder… Sa magie, c'est de considérer tous ces éléments pour rendre accessible les données par n'importe quelle application.
Nous avons donc développé une technologie qui permet de construire très facilement des applications qui respectent ces standards. Une application peut donc se connecter aux données de tous les partenaires qui utilisent la même représentation de données, accéder à toutes les informations d'un espace de données (ou *data space*). Cela favorise la collaboration entre partenaires, sans nécessiter un intermédiaire « plateforme » qui centraliserait l'écosystème.

Le développement de standards dans tous les secteurs d'activité est par ailleurs fortement encouragé par des institutions nationales et internationales. Les organisations internationales sont ainsi demandeuses de normes et de « bonnes pratiques »[17] et soutiennent l'intense travail normatif d'organismes de standardisation. Le plus connu d'entre eux est sans doute l'International Standard Organisation (ISO), dont les normes sont censées garantir la qualité de produits, de services ou encore de processus.

17. Voir par exemple : Klein A., Laporte C., Saiget M., 2015, *Les bonnes pratiques des organisations internationales*, Paris, Presses de Sciences Po, 241 p.

DE NOMBREUSES INITIATIVES EN FAVEUR DE LA STANDARDISATION DES DONNÉES OUVERTES

Le monde de l'open data n'échappe pas à la question de la standardisation. Au contraire, les schémas de données encadrant la production et la réutilisation de données, notamment par les collectivités territoriales, sont de plus en plus nombreux.

standardiser des données grâce à des schémas : l'exemple des aménagements cyclables

La standardisation des données ouvertes s'articule autour de schémas. Ces derniers sont des standards lisibles par des machines, des conventions qui décrivent les champs et les valeurs admises dans un jeu de données conforme à ses préconisations. C'est donc en s'y conformant que nous produisons des jeux de données standardisés. Compréhensibles par les machines, les schémas sont réexploités dans des formulaires et des interfaces à destination des humains.

Prenons un exemple. La loi d'orientation des mobilités de 2019 (LOM) exige entre autres des collectivités qu'elles mettent à disposition des données concernant les caractéristiques de leur réseau cyclable sur la plateforme transport.data.gouv.fr, qui est le point d'accès national des données de mobilité. Ces données doivent être normalisées sur l'ensemble du territoire, donc produites localement selon le schéma de données d'aménagements cyclables : le producteur décrit donc ses aménagements selon les champs définis par le schéma (types d'aménagement sur chaque voie, largeur des aménagements, sens de circulation…). Ce schéma définit également la manière de représenter les valeurs des champs. Certains champs n'admettent que les valeurs d'une liste arrêtée (15 valeurs sont ainsi possibles pour les types d'aménagements), d'autres n'acceptent qu'un certain format (le code INSEE de la commune doit nécessairement comporter 5 chiffres), quelques valeurs peuvent être représentées sans contrainte (texte libre).
La saisie de ces données standardisées s'effectue sur des logiciels métiers internes, sur OpenStreetMap[18], ou grâce aux outils développés par l'association Vélo & Territoires.

18. Projet cartographique collaboratif, qui associe des clefs et des valeurs pour ajouter des informations aux objets géographiques. L'attribut « motor_vehicle=no » permet par exemple de préciser qu'une voie est interdite aux véhicules moteurs.

```
},
"ame_g": {
    "type": "string",
    "description": "Type d'aménagement présent sur la voie de gauche",
    "examples": [
        "BANDE CYCLABLE"
    ],
    "enum": [
        "PISTE CYCLABLE",
        "BANDE CYCLABLE",
        "DOUBLE SENS CYCLABLE PISTE",
        "DOUBLE SENS CYCLABLE BANDE",
        "DOUBLE SENS CYCLABLE NON MATERIALISE",
        "VOIE VERTE",
        "VELO RUE",
        "COULOIR BUS+VELO",
        "RAMPE",
        "GOULOTTE",
        "AMENAGEMENT MIXTE PIETON VELO HORS VOIE VERTE",
        "CHAUSSEE A VOIE CENTRALE BANALISEE",
        "ACCOTEMENT REVETU HORS CVCB",
        "AUCUN",
        "AUTRE"
    ]
},
"regime_g": {
    "type": "string",
    "description": "Régime présent sur la voie de gauche",
    "examples": [
        "AIRE PIETONNE"
    ],
    "enum": [
        "ZONE 30",
        "AIRE PIETONNE",
        "ZONE DE RENCONTRE",
        "EN AGGLOMERATION",
        "HORS AGGLOMERATION",
        "AUTRE"
    ]
},
"sens_g": {
    "type": "string",
    "description": "Sens de circulation pour les cyclistes sur la voie de gauche",
    "examples": [
        "UNIDIRECTIONNEL"
    ],
    "enum": [
        "UNIDIRECTIONNEL",
        "BIDIRECTIONNEL"
    ]
},
```

Extrait du schéma de données des aménagements cyclables, dans lequel sont spécifiées les valeurs possibles pour les champs « type d'aménagement présent sur la voie de gauche », « régime présent sur la voie de gauche » et « sens de circulation pour les cyclistes sur la voie de gauche ».

La standardisation des données ouvertes s'inscrit dans la lignée d'une série de démarches relativement anciennes. Dans la lignée du monde de la statistique, dont la comparabilité des indicateurs repose sur leur standardisation, celui de l'information géographique et environnementale a engagé des travaux avancés sur la standardisation depuis 2007. La directive INSPIRE proposait ainsi des dispositions relatives à l'interopérabilité des données géographiques et environnementales passant par la standardisation des métadonnées mais aussi des données elles-mêmes. En France, une commission interministérielle (la COVADIS) avait été mise en place en 2008 par les ministères en charge de l'écologie, du logement et de l'agriculture pour établir des « géostandards » selon la méthodologie de la directive INSPIRE. La COVADIS avait donc pour mission de standardiser leurs données géographiques les plus fréquemment utilisées dans les métiers de ces ministères. Progressivement, depuis 2013, la gouvernance de ces standards a été internalisée au sein du CNIG. En s'appuyant sur des groupes de travail ouverts, il pilote certains standards, comme GraceTHD pour les réseaux très haut débit, RAEPA pour l'eau potable, ou encore Star-DT pour les travaux publics.

Le CNIG, acteur central de la standardisation des données géographiques

Le conseil national de l'information géolocalisée est le principal concepteur de standards de données géographiques en France (généralement des standards réglementaires). En 2022, l'ancienne commission « données » est d'ailleurs rebaptisée commission « standards », attestant de la place importante tenue par la conception de géostandards dans les activités du CNIG.
La conception de nouveaux standards s'effectue au sein de groupes de travail (GT) réunissant régulièrement et sur une durée d'environ un an minimum producteurs et réutilisateurs de données. Les comptes-rendus de ces GT sont systématiquement mis en ligne. Le travail des GT est encadré par un mandat, dans lequel sont notamment (mais non exclusivement) indiqués : les raisons d'être du futur standard (à quels problèmes répondra-t-il ?), le contexte réglementaire (directives européennes, lois françaises…), le fonctionnement du GT (principaux acteurs, méthodologie, durée…). Sont par exemple précisés les organismes chargés du pilotage et de l'animation du GT — le CEREMA assure régulièrement l'animation et le secrétariat technique.
Un animateur de plusieurs d'entre eux nous explique les enjeux de la composition de ces groupes de travail, au sein duquel doivent être représentés des experts « données » et « métier » :

> *« [les membres du groupe de travail] sont les collectivités qui souhaitent participer, qui se manifestent. On se base vraiment sur le volontariat, il n'y a aucune gratification. Si une collectivité trouve un intérêt particulier, les groupes de travail CNIG sont ouverts, tout simplement. […] Ce qu'on recherche ce sont des profils mixtes, métier et technique. Dans certains cas, il y a des personnes qui ont les deux profils et c'est l'idéal, mais ce sont des moutons à cinq pattes. Alors ce qu'on recherche en général ce sont des binômes. Pour le sujet accessibilité, un excellent binôme serait un ou une géomaticienne et une personne dans le métier de l'accessibilité depuis des années. »*

Au-delà des données géographiques et des initiatives internationales, la France connaît depuis peu un regain d'intérêt et un foisonnement d'initiatives sur la question de la standardisation. Depuis 2018, l'association OpenDataFrance, qui fédère les collectivités engagées dans une démarche d'ouverture des données, développe le Socle Commun des Données Locales (SCDL) pour homogénéiser la publication en open data de données essentielles produites par des acteurs territoriaux, aider les producteurs à améliorer la qualité des données qu'ils publient et faciliter l'exploitation des données publiées par les réutilisateurs. Huit jeux de données préalablement sélectionnés comme prioritaires[19] ont fait l'objet d'une démarche de standardisation et ce socle est en train de s'étendre. Le SCDL a aussi impulsé une dynamique dans l'administration d'État avec le lancement en juin 2019 de schema.data.gouv.fr, une initiative d'Etalab (département de la direction interministérielle du numérique — DINUM) qui référence les standards français qui ont été adoptés par voie réglementaire ou conçus par la communauté des producteurs et réutilisateurs de données. Le site référence également des schémas en cours d'investigation et de construction.

19. Ces schémas sont : Budget des collectivités et établissements publics locaux ; Catalogue des jeux de données publiés en open data par une collectivité ; Composition des plats proposés par les collectivités locales ou les syndicats mixtes de restauration ; Délibérations adoptées par une collectivité ; Équipements collectifs publics d'une collectivité ; Composition des menus proposés par les collectivités locales ou les syndicats mixtes de restauration ; Prénoms des nouveaux-nés déclarés à l'état-civil ; Subventions attribuées par une collectivité.

schema.data.gouv.fr : produire collaborativement des schémas pour homogénéiser les données.

Etalab, département de la direction interministérielle du numérique, coordonne la conception et la mise en œuvre de la stratégie de l'État dans le domaine de la donnée. En 2019, il met en place l'outil de référencement des schémas de données publiques français schema.data.gouv.fr.

Geoffrey Aldebert, data engineer à Etalab, nous explique ses origines : *« Suite à la publication de nombreux jeux de données en open data est née une réflexion autour de la qualité de ces données et de la mise en place de schémas qui permettrait d'homogénéiser la façon dont elles sont publiées. Ça a été porté initialement par OpenDataFrance, via le Socle Commun des Données Locales, à travers quelques schémas de données exemplaires. Devant ces initiatives qu'on [Etalab] jugeait fructueuses, on a lancé schema.data.gouv.fr, qui vise à être un site internet et une communauté qui rassemble les différents producteurs de schémas de données et producteurs de données structurées pour accéder à de la documentation partagée sur un même type de données. »*

À l'été 2022, près de cinquante schémas de données sont référencés, dont une dizaine en investigation. La documentation proposée sur le site permet aux producteurs de données de s'approprier ces schémas et donc de produire des jeux de données les respectant.

Geoffrey Aldebert insiste sur l'importance d'une communauté pour faire vivre schema.data.gouv, qui référence des schémas produits par une multitude d'acteurs, que l'équipe d'Etalab peut accompagner : *« On avait déjà axé data.gouv.fr sur un aspect communautaire : les administrations peuvent publier des données, les associations peuvent publier des données, à partir du moment où ces données sont d'intérêt public. schema.data.gouv.fr c'est la même chose. Ce n'est pas Etalab qui référence les schémas de données, ce sont des acteurs qui font des propositions. Et Etalab est là pour les accompagner, à la fois techniquement et méthodologiquement sur la construction de ce schéma de données. »*

Afin de faciliter la production de jeux de données respectant ces schémas, Etalab a également développé l'outil publier.etalab.studio. Celui-ci permet de valider et de corriger un jeu de données standardisé en vue de sa publication sur data.gouv.fr.

Ces initiatives s'inscrivent dans le contexte d'une impulsion politique favorable à la standardisation. Le rapport sur la politique de la donnée remis par le député Bothorel au Premier ministre le 23 décembre 2020, porte une recommandation (n° 24) sur la définition et la mise en œuvre d'une politique interministérielle d'interopérabilité et de qualité de la donnée, insistant sur l'importance des démarches de standardisation. Suite à ce rapport, 15 feuilles de route ministérielles ont été publiées le 27 septembre 2021 pour assurer la mise en œuvre de ces préconisations. La feuille de route du ministère de la Cohésion des territoires fixe dans son action 15 l'objectif suivant :

> *« encourager l'ouverture de données selon des référentiels partagés est un gage de qualité qui, à terme, facilitera l'interopérabilité, voire l'émergence de solutions ouvertes. En collaboration avec les associations de collectivités, des territoires pionniers à différentes échelles, ainsi que des éditeurs de solutions numériques équipant les collectivités, il s'agit de converger et de promouvoir les meilleures pratiques de normalisation. »*

LES STANDARDS : UNE CONDITION DE L'ÉCHANGE DES DONNÉES, UN SUJET ET OUTIL DE COLLABORATION ENTRE PRODUCTEURS ET RÉUTILISATEURS

CET encouragement institutionnel à la standardisation des données ouvertes intervient dans un contexte où les acteurs publics publient des données très hétérogènes, en l'état difficilement agrégeables au niveau national. La standardisation apparaît ainsi comme une solution pour favoriser l'échange des données et donc multiplier leurs réutilisations. En d'autres termes, les standards sont des outils permettant de représenter de manière homogène des données collectées de façon disparate, participant ainsi à décupler les effets de l'open data. Le sociologue Tim Davies estime ainsi que, dans un cadre où la collecte des données est disparate, les standards permettent d'uniformiser les données et de répondre mieux aux attentes des réutilisateurs :

> *« Les données doivent être collectées de la manière qui a le plus de sens au niveau local (sous réserve de la capacité à s'adapter ensuite à un standard) — et doivent être présentées de la manière qui répond aux besoins des utilisateurs. Mais lorsque les données proviennent de nombreuses sources et vont vers de nombreuses destinations, le standard est un outil clef de collaboration. »*[20]

À la fois outil et produit de la collaboration, un « bon » standard résulte d'une entente entre plusieurs producteurs de données, ayant chacun leurs propres contraintes, et des réutilisateurs aux attentes parfois non-anticipées par ces producteurs. Conciliant ainsi les astreintes des uns et les souhaits des autres, il est le résultat d'une démarche collaborative ayant mené à des compromis. Le processus de conception est donc susceptible de créer un dialogue utile, sinon nécessaire, pour s'assurer de la réutilisation optimale des données ouvertes[21].

20. Davies T., 2021, « A data portal deep dive », timdavies.org.uk (nous traduisons).
21. La partie « Les enjeux de la concertation », p. 44, développe plus en détails la nécessité de faire dialoguer des acteurs aux profils diversifiés.

QUID DES DONNÉES NON STANDARDISÉES ?

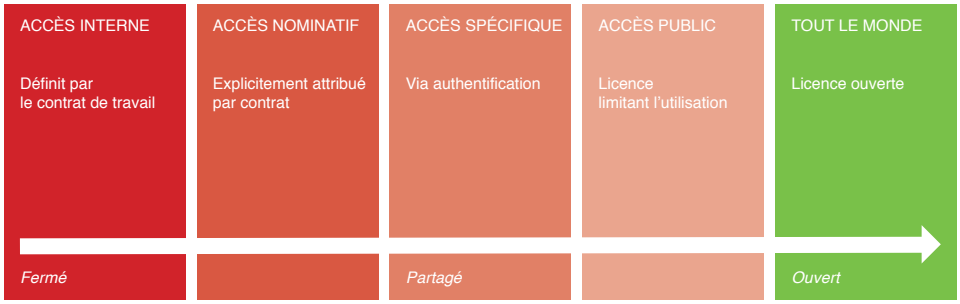
LA standardisation prend une importance croissante dans le monde de l'open data mais cela ne remet pas en question la valeur des données non standardisées. La publication de jeux de données dans un format ouvert (CSV plutôt que PDF par exemple) constitue déjà une étape importante dans l'application du principe de transparence. Ayant encouragé le législateur à rendre l'ouverture des données publiques obligatoire « par défaut », celui-ci est en effet encore loin d'être appliqué systématiquement. Pour rappel, cette obligation est largement respectée par les régions, les départements et les communes de plus de 100 000 habitants, mais les collectivités de moins de 100 000 habitants ne sont environ que 10 % à la suivre[22]. Rendre accessible des données, sous une licence ouverte permettant sa libre réutilisation, répond déjà à un double impératif juridique et démocratique.

La standardisation de données déjà publiées peut tout de même constituer une étape supplémentaire dans la stratégie open data d'une collectivité. Rendre conforme un jeu de données à un schéma est en effet possible post-publication. De la même manière que les données ne sont pas disponibles et prêtes à l'emploi[23], retravailler ces données afin de les normaliser demande un travail conséquent. Dans certains cas, il peut tout de même être facilité par des progiciels. Nous reviendrons plus en détail sur ce point dans la dernière partie du cahier.

Tout comme il y a parfois de « bonnes raisons » entravant l'ouverture des données publiques[24], les collectivités peuvent également avoir des raisons objectives, autres que le surplus de travail, de ne pas se lancer dans une démarche de standardisation.

---

22. Résultats de l'enquête 2021 de l'observatoire open data des territoires. Voir également le premier cahier Data Publica, « Donnée, intérêt général & territoires : la construction d'un cadre de confiance », qui propose un état des lieux de l'ouverture des données publiques.
23. Sur ce travail de la donnée, voir notamment : Denis J., 2018, *Le travail invisible des données : Éléments pour une sociologie des infrastructures scripturales*, Paris, Presses des Mines (Sciences sociales), 206 p.
24. Samuel Goëta traite de cette question dans sa thèse de sociologie, en particulier dans le chapitre 4 : « Les frictions de l'identification : quelques "bonnes raisons organisationnelles» de ne pas ouvrir des données ». Goeta, S., 2016, *Instaurer des données, instaurer des publics : une enquête sociologique dans les coulisses de l'open data*, thèse de doctorat en sociologie, Télécom ParisTech.



Le spectre d'ouverture des données (ODI, traduit par Datactivist)

Tout d'abord, les schémas de données ne sont pas infinis. Les données publiées variant encore largement d'un territoire à l'autre, il y en a pour lesquelles il n'existe tout simplement pas de standard. Cela peut être le cas des thématiques très spécialisées. Il n'est cependant pas exclu qu'un schéma soit conçu dans les prochains mois ou années.
Par ailleurs, un schéma est parfois limitant. Il peut ne pas prévoir des champs ou valeurs qu'un producteur est pourtant en mesure de renseigner et qui auraient une forte valeur pour les usagers. En d'autres termes, il est susceptible d'« appauvrir » un jeu de données. Pour répondre à cette problématique, des extensions sont parfois créées. Cela permet au producteur de données de répondre à des besoins utilisateurs particuliers. C'est par exemple le cas pour l'Open Contracting Data Standard (OCDS), qui encadre l'ouverture des données liées à toutes les étapes de la vie d'un marché public[25]. Tout producteur voulant renseigner des données supplémentaires peut créer une extension. Elle doit alors être documentée de manière à être réutilisable et intégrable à une nouvelle version du standard. Il en existe actuellement une soixantaine. L'extension « *shareholder* » (actionnaire) permet par exemple de fournir des détails sur la structure de propriété des entreprises impliquées dans un contrat.

25. Voir la section « Les standards au service de la transparence de l'action publique » (p. 30) pour plus de détails sur ce cas.

QUELQUES STANDARDS SECTORIAUX : LES CAS DE LA TRANSPARENCE ET DE LA MOBILITÉ

NOUS avons présenté les principales raisons théoriques justifiant la standardisation des données ouvertes (meilleure interopérabilité, plus de réutilisations possibles…). En permettant l'agrégation et le croisement de données produites par un grand ensemble d'acteurs, les standards peuvent être des outils concrets au service de l'action publique et de la mise en œuvre de grands principes démocratiques. Plusieurs standards, aux niveaux français, européen et international, permettent ainsi aux acteurs publics de s'emparer du sujet de la transparence de l'action publique, particulièrement budgétaire, en publiant des données relatives à la commande publique ou encore aux subventions. Dans le domaine de la mobilité, où sont produites de nombreuses données d'intérêt public, la standardisation permet de développer des applicatifs à destination des usagers (calculateur d'itinéraire par exemple) mais également de mettre en lumière les améliorations possibles en matière d'inclusivité des services.

LES STANDARDS AU SERVICE DE LA TRANSPARENCE DE L'ACTION PUBLIQUE

Jean-Marc Sauvé, alors vice-président du Conseil d'État, estimait dans un discours lors de l'Assemblée générale de l'inspection générale de l'administration, que la transparence est « *perçue comme la condition de la participation des citoyens à l'élaboration et au contrôle de l'action publique. Elle est nimbée d'une aura de modernité, de respectabilité, voire de rectitude, et elle tend à s'imposer comme une obligation incontournable de l'administration*. » Il ajoute qu'elle favorise une meilleure gouvernance publique et efficacité de l'administration, en permettant notamment des achats publics et des recrutements « plus pertinents »[26].

Ce principe essentiel de la vie démocratique s'applique particulièrement au domaine de la commande publique. Dès 2006, les acheteurs publics ont l'obligation de publier la liste des marchés conclus l'année précédente. Considérant la nécessité de standardiser ces données en vue de faciliter leur exploitation, le code de la commande publique[27] impose depuis 2019 une publication normalisée des données des marchés publics dont la valeur est égale ou supérieure à 40 000 € HT et de tous les contrats de concessions sur les profils d'acheteurs. Elles renseignent sur la procédure de passation, le contenu du contrat et son exécution. L'arrêté du 22 mars 2019 relatif aux données essentielles de la commande publique, remplaçant celui du 14 avril 2017, impose un schéma pour la publication de ces dernières. Celui-ci est pris en compte par les progiciels des acheteurs afin de permettre l'export de jeux de données correctement formatés.

26. Intervention de Jean-Marc Sauvé lors de l'Assemblée générale de l'inspection générale de l'administration le 3 juillet 2017.
27. Il est entré en vigueur le 1er avril 2019, suite à l'abrogation du code des marchés publics.

contrôler le recours aux cabinets de conseil par les acteurs publics : le rôle des DECP

En janvier 2021, le média *Politico* révélait le rôle du cabinet McKinsey dans la campagne de vaccination française[28]. Un an plus tard, en mars 2022, un rapport sénatorial dénonçait le poids croissant des cabinets de conseil sur les politiques publiques[29]. Les sénateurs de la commission d'enquête y regrettaient notamment l'absence de base de données centralisée permettant de dénombrer exactement les missions de conseil pour les acteurs publics et leur contenu. Ils estimaient alors leur nombre à plus de 900. Des enquêtes journalistiques ont permis de mettre à jour ces chiffres. En utilisant notamment les DECP, qui ont constitué une des sources pour quantifier les dépenses liées au recours au conseil, *Le Monde* a ainsi produit une liste de 1600 missions pour l'État, représentant un montant de près de 2 milliards d'euros.
En plus de permettre une « exploration » des données sur le site du journal, l'équipe de journalistes a ouvert l'accès à la base de données créée à partir du « BOAMP » (Bulletin officiel des annonces de marchés publics), où sont recensés les appels d'offres au-dessus de 40 000 euros, de son équivalent européen « TED » (« Tenders Electronic Daily ») et des DECP. Dans une majorité des cas, il est cependant impossible de préciser la nature des prestations, voire le montant du contrat et le nom de l'acheteur. Les Décodeurs du *Monde* précisent ainsi : *« Dans les sources publiques comme le « BOAMP » et les DECP, il est difficile de tracer le périmètre des missions de conseil, qui ne sont pas désignées par une nomenclature spécifique »*[30].

L'absence de données d'exécution dans le schéma des DECP est une de ses principales limites : cela restreint la fiabilité de l'information et rend nécessaire un véritable travail d'enquête afin de préciser les montants réellement engagés et la nature des prestations, notamment dans le cadre d'un accord-cadre. L'économiste Pierre-Henri Morand pointe en particulier le manque de données entourant ces accords-cadres, dont au moins 15 ont été conclus par l'État pour des prestations de conseil. À travers un exemple impliquant trois cabinets, il montre que les données ouvertes ne permettent pas de savoir ce qui se passe réellement à l'intérieur : *« Nous évoquions un lot à 20 millions d'euros pour lequel Roland Berger, Boston Consulting Group et McKinsey avaient été sélectionnés. Mais en vérité on ne sait pas, dans les données ouvertes, qui des trois groupes a réalisé le plus de prestations. Est-ce équilibré […] ou une des trois entreprises s'est-elle taillé la part du lion ? Quel service, pour quelle mission et pour quel type de conseil, a utilisé cet accord cadre ? Là encore, les données ouvertes ne nous permettent pas de le savoir. »*[31]

28. Braun E., de Villepin P., 8 février 2021, « Comment les cabinets de conseil comme McKinsey ont conquis la France », *Politico*.
29. Assassi, E., mars 2022, "Un phénomène tentaculaire : l'influence croissante des cabinets de conseil sur les politiques publiques », rapport fait au nom de la commission d'enquête "Cabinets de conseil » du Sénat.
30. Romain M., Martinon L., Vaudano M., Les Décodeurs, 17 mars 2022, « Explorez les 1 600 missions des cabinets de conseil pour l'Etat recensées par "Le Monde" », Le Monde.
31. Morand, P-H., 2022, *Où va l'argent public ? La commande publique au défi des données ouvertes, Avignon*, Éditions universitaires d'Avignon, p. 97-98.

En plus des DECP, les collectivités, soumises depuis la loi pour une République numérique de 2016 au principe d'ouverture par défaut des données publiques et d'intérêt général, ont d'autres obligations en matière de transparence de l'action publique. Celles de plus de 3 500 habitants ou de 50 agents doivent par exemple également rendre accessibles les données relatives à l'attribution des subventions de plus de 23 000 €. Pour ce faire, le Socle Commun des Données Locales (SCDL) met à disposition un schéma. Le SCDL contient également un schéma du budget des collectivités et établissements publics locaux ainsi qu'un schéma des délibérations, qui, sans être réglementaires, participent à l'amélioration de la transparence des décisions publiques.

En plus de dépasser les obligations réglementaires en s'appuyant sur des schémas conçus en France, il est également possible pour les acteurs publics de recourir à des standards internationaux. Couvrant des sujets certes proches, ils répondent en pratique à des besoins différents, ce qui justifie leur coexistence.

La première version de l'Open Contracting Data Standard (OCDS) a été publiée à l'été 2015, notamment grâce au soutien de la Banque mondiale. Ce standard permet de suivre la passation de marchés publics, depuis l'appel d'offre jusqu'à la mise en œuvre, en passant par l'attribution et la contractualisation, grâce à un identifiant unique. Selon l'Open Contracting Partnership, l'ONG en charge de coordonner la gouvernance du standard : « *le standard fournit des conseils sur ce qu'il faut publier et la manière de le publier, il est possible de comparer et d'analyser plus facilement les données normalisées. L'utilisation de ces données peut aider les gouvernements à améliorer l'efficience financière [* « deliver better value for money »*], à prévenir et à découvrir la corruption, à accroître la concurrence et à mieux suivre les prestations de services.* »[32] Ce standard est actuellement utilisé dans une trentaine de pays. Il a par exemple participé à la lutte contre la corruption en Ukraine. Suite à la révolution de 2014, la publication de données selon le standard OCDS a été généralisée. Cela a participé à l'ouverture à la concurrence des marchés et à l'augmentation de l'efficacité du processus de passation des marchés. D'après le projet Dozorro, géré par Transparency International, en décembre 2020 plus de 33 000 appels d'offres frauduleux équivalant à 4,5 milliards de dollars avaient été mis en plein jour et des économies de plus de 6 milliards de dollars avaient été réalisées.

32. Site de l'Open Contracting Partnership : https://www.open-contracting.org/ (nous traduisons).

Visuel créé par l'Open contracting partnership.

L'International Aid Transparency Initiative (IATI) est une initiative lancée en 2008, dont le standard de données du même nom permet aux donateurs d'aide de partager des informations concernant leurs dépenses et les projets financés. Développé pour répondre aux besoins des donateurs et des bénéficiaires d'une aide institutionnelle, le standard est à présent utilisé par plus de 400 organisations, dont un nombre croissant d'ONG[33]. Ce standard a notamment permis à des acteurs publics de pays en développement de mieux connaître les projets menés sur leur propre territoire et ainsi de repenser la coordination de l'aide, en assurant une meilleure allocation des ressources complémentaires à l'aide internationale. C'est par exemple grâce à l'IATI que le gouvernement de Madagascar a découvert 830 millions de dollars US d'engagements de dépenses dont il n'avait jusqu'à lors pas connaissance[34].

33. Pour plus de détails sur ce cas, voir Goëta S., Davies T., 2016, « *The Daily Shaping of State Transparency: Standards, Machine-Readability and the Configuration of Open Government Data Policies* », Science, Technology and Innovation Studies.
34. L'IATI a publié une étude de cas (en français) de Madagascar sur son site internet : iatistandard.org

En proposant des normes mondiales pour la publication de données, l'IATI et l'OCDS favorisent donc la circulation d'une information fiable et durable concernant les flux financiers publics.
La coexistence de ces schémas se justifie par leur réponse à des besoins différents mais nécessite cependant d'envisager les modalités de leur compatibilité. Rédiger des règles de conversion entre ces schémas est nécessaire, afin notamment d'éviter les saisies multiples de données. Mais cela nécessite un travail conséquent, d'autant plus s'il n'est pas entamé très en amont du processus de conception d'un nouveau schéma de données. Lors d'un entretien, un acteur des DECP explique l'intérêt de la coexistence d'un standard national et international et insiste sur l'importance de leur compatibilité :

> *« [DECP et OCDS] ne répondent pas au même besoin. On pourrait se dire qu'ils sont concurrents, se demander pourquoi il n'existe pas qu'un schéma, mais ils ne sont pas là pour les mêmes besoins. Du coup ce n'est pas grave qu'il y ait plusieurs schémas, c'est même une bonne chose parce que ça permet d'avoir des schémas plus adaptés à ce pour quoi ils sont conçus. Par contre, effectivement, quand on développe un nouveau schéma il faut se renseigner sur ce qui existe déjà pour qu'il y ait une compatibilité, pour pouvoir transformer facilement les données d'un schéma à l'autre. »*

LA NÉCESSAIRE STANDARDISATION DES DONNÉES DE MOBILITÉ

LES collectivités territoriales, en tant qu'autorités organisatrices de mobilité (AOM), ont plusieurs obligations de publication de données standardisées liées à la mobilité, spécifiquement à l'information voyageur. Celles-ci concernent notamment les réseaux, les arrêts, les horaires et l'accessibilité des transports en commun. Elles doivent également publier les données relatives à l'offre de véhicules en partage (autopartage, vélos en partage), de transport longue distance, de stationnement, aux réseaux cyclables, ou encore à la localisation et aux caractéristiques des Infrastructures de Recharge pour Véhicules Électriques (IRVE)[35]. Des schémas européens ou français assurent l'homogénéité de ces données.

35. Pour plus de détails, voir le site de transport.data.gouv.fr, qui centralise tous les schémas de données réglementaires en matière de mobilité.

transport.data.gouv.fr : le point d'accès national aux données de transport.

transport.data.gouv.fr rassemble toutes les données liées à l'offre de mobilité en France. Les producteurs de données sont invités (dans la majorité des cas à la suite d'une obligation réglementaire) à y publier leurs jeux de données dans le respect de standards.

Lorsque les schémas de données ne pas définis à l'échelle européenne (NeTEx — *Network Exchange* — est par exemple un format de référence pour échanger des données d'offre théorique du transport collectif au niveau européen) ou préexistants aux obligations légales, transport.data.gouv.fr en impulse la conception, à travers l'animation de groupes de travail. Ces derniers réunissent, sur la base du volontariat, des producteurs de données représentatifs de la diversité territoriale (grandes métropoles, petites communes urbaines et rurales…).
Miryad Ali, responsable de l'ouverture des données chez transport.data.gouv.fr, nous explique ainsi : *« Pour d'autres données, par exemple les aménagements cyclables ou les aires de covoiturage, on demande une ouverture mais il n'y a pas forcément de format qui a été défini. Donc notre travail est de définir un schéma pour pouvoir demander aux collectivités d'ouvrir leurs données — puisqu'elles ont cette obligation, mais on ne peut pas aller les voir alors que rien n'est fait. »* Elle ajoute que l'équipe transport.data.gouv.fr joue un rôle actif dans la promotion des schémas de données liés à la mobilité et, pour en faciliter l'adoption, propose une aide à la production de données standardisées : *« On est à la fois concepteur et promoteur. On va d'abord concevoir le schéma, en s'appuyant sur des groupes de travail, et une fois ce schéma consolidé on va contacter les collectivités pour leur demander d'ouvrir leurs données avant une certaine échéance, tout en leur expliquant qu'elles peuvent revenir vers nous en cas de questions ou de difficultés. »*

Capture d'écran du site transport.data.gouv.fr — le 29/07/2022.

En plus de ces standards réglementaires, il existe des standards d'usage non réglementaire internationaux. Certains d'entre eux ont été documentés dans un benchmark réalisé par La Fabrique des Mobilités, portant spécifiquement sur les standards MaaS (*mobility as a service*)[36]. Le rapport présente notamment le standard GTFS (General Transit Feed Specification), initialement développé à Portland en partenariat avec Google, qui s'est imposé comme référence en matière d'horaires. Ce standard cumule une composante « horaire » (*schedule*), contenant des données sur les horaires théoriques, les tarifs et les transports en commun géographiques, et une composante « temps réel » (*realtime*) qui informe sur les prévisions d'arrivée. Il permet notamment l'import de données dans des calculateurs d'itinéraires.
Au niveau de l'Union Européenne, le format ayant vocation à s'imposer pour l'échange des données théoriques du transport collectif est NeTEx. Plus complexe que le GFTS, il est encore assez peu utilisé. Il permet cependant d'échanger un plus grand nombre d'informations utiles pour les voyageurs. NeTEx est par exemple beaucoup plus précis que le GTFS dans la description de l'accessibilité. Les champs « wheelchair_accessible » et « wheelchair_boarding » du GTFS étant facultatifs, les personnes en fauteuil n'ont pas nécessairement d'informations concernant l'accessibilité des véhicules et d'une station/d'un quai. Et les valeurs possibles ne renseignent pas sur les modalités concrètes d'accès aux quais. Ce standard réduit par ailleurs ce que regroupe l'accessibilité, considérant seulement une forme de handicap physique et en laissant de côté de nombreuses autres. La prise en compte de la malvoyance supposerait par exemple de renseigner l'existence ou non de marquages au sol. Les choix opérés par les concepteurs ont ici des conséquences très concrètes : l'absence de certains champs et valeurs rend invisibles des aménagements nécessaires à certains publics.

36. Grandjean M., Delabie G., janvier 2022, « Standards MaaS. Gouvernance & performance », La Fabrique des Mobilités.

NeTEx : un ensemble commun et des profils complémentaires

NeTEx couvre trois grandes catégories de données de mobilité collective : typologie des réseaux, horaires théoriques, informations tarifaires. Il répond à une nécessité d'homogénéisation des pratiques d'échange de données, que transport.data.gouv.fr juge essentielle pour :

- l'usager final : pour accéder à une offre complète ;
- les autorités organisatrices : pour compiler des informations homogènes venant des différents opérateurs et envisager des systèmes d'information multimodaux ;
- les opérateurs : pour enrichir leurs systèmes de planification, d'exploitation, de billettiques et d'information voyageur ;
- les industriels : pour pérenniser les investissements sur les formats d'échanges.[37]

NeTEx se compose d'un profil général d'échange, c'est-à-dire d'une base européenne commune, auquel s'ajoutent d'autres profils plus spécifiques, adoptés par un pays membre. Ces profils nationaux sont éventuellement généralisés à la suite d'un travail collectif lors duquel sont représentés tous les États membres. C'est le cas pour le profil accessibilité français, qui fait actuellement l'objet de discussions pour être étendu, tout en étant modifié à la marge, au niveau européen.
Contrairement à certains standards relativement « simples », qui peuvent conduire à la production manuelle de jeux de données, NeTEx est un ensemble volumineux de normes auquel le producteur de données n'est jamais directement confronté. Un des concepteurs du profil NeTEx accessibilité précise que les producteurs de données n'ont pas besoin de s'intéresser aux éléments techniques :

> *« Tout ça c'est outillé. On ne regarde jamais le contenu d'un fichier word. Vous tapez votre texte et ça vous fait un truc avec un format complètement illisible à l'intérieur, pire que NeTEx. On ne regarde jamais non plus le code HTML derrière votre page internet. Donc le travail de normalisation qu'on fait permet d'alimenter des outils. À aucun moment l'utilisateur ne devra être confronté au format NeTEX. Des applications leur donnent une vue purement fonctionnelle de ce qu'ils devront faire, puis ils exportent les données automatiquement. »*

37. transport.data.gouv.fr, janvier 2022, « Éléments communs aux profils d'échange pour les informations planifiées du transport en commun », https://normes.transport.data.gouv.fr/posts/elements_communs/

Les choix en matière de publication des données liées à l'information voyageur ne sont pas neutres. Collecter et rendre public de nouvelles données peut avoir un impact majeur sur le quotidien des personnes pour qui l'accès à la mobilité collective est entravée par le manque d'informations sur l'accès aux arrêts, aux moyens de transport, ainsi qu'aux établissements recevant du public (ERP) vers lequel elles se déplacent généralement. Publier des données en suivant un standard conçu pour répondre aux besoins de ces publics, aujourd'hui encore largement exclus des services de mobilité, participe à rendre ces derniers plus inclusifs. Cela est également essentiel pour faire un état des lieux précis de l'accessibilité et préciser ou ajuster les stratégies territoriales.

3 standards pour décrire l'accessibilité dans le monde de la mobilité

La LOM impose aux gestionnaires de voirie et aux autorités organisatrices de transports de créer des bases de données décrivant l'accessibilité à l'horizon décembre 2023[38]. L'objectif de cette obligation est de permettre aux personnes handicapées ou à mobilité réduite de se déplacer et d'accéder à tous les espaces urbains de la même manière qu'une personne valide. Car le manque de données, ou des données erronées, a des conséquences concrètes, comme le rappelle un expert de la normalisation des données de transport :

« Les exemples de mauvaises données qui sont donnés par les associations sont vraiment casse-pied. C'est quelqu'un en fauteuil roulant qui prend le métro à un endroit et qui ne se rend compte que dans la station où il voulait aller, ça ne marche finalement pas : il ne peut pas sortir. Mais il ne peut pas non plus faire demi-tour car il ne peut pas aller sur le quai d'en face. Il ne peut que continuer. Donc le problème généré pour les gens est réel, lourd de conséquences. Plus encore que l'absence de données, une mauvaise donnée a des conséquences très lourdes. »

38. Le site du ministère de la Transition écologique et des solidarités résume les obligations et les chantiers en cours concernant les données d'accessibilité : https://www.ecologie.gouv.fr/donnees-daccessibilite

Beaucoup de collectivités n'ont pas encore à disposition l'intégralité des données d'accessibilité voirie et transport de leur territoire. Un important travail de collecte est donc nécessaire. **Le standard d'échange de données d'accessibilité voirie**, validé par le CNIG fin 2021, permet d'harmoniser cette collecte d'informations géographiques de description de l'accessibilité des cheminements de la voirie, de l'espace public, et des établissements recevant du public (ERP).
Le **profil accessibilité France de NeTEx** permet quant à lui d'échanger des informations liées à l'accessibilité du transport public (d'une gare, d'un quai…). Des spécifications communes et des règles de conversion doivent assurer l'interopérabilité des données entre ces deux standards.
La direction ministérielle à l'accessibilité encourage par ailleurs les collectivités à aller au-delà des seules obligations légales en collectant et publiant également **des données sur l'accessibilité des établissements recevant du public en utilisant un troisième standard**, également compatible avec les deux autres. Une membre de l'équipe acceslibre.beta.gouv.fr, plateforme nationale de référencement de l'accessibilité des ERP explique ainsi :
« L'objectif c'est que les communes, qui ont cette obligation de collecter et publier de la donnée transport et voirie, aillent au bout de la démarche en collectant également les ERP qui sont autour de ces points d'arrêt. La loi impose de collecter les informations d'accessibilité des points d'arrêt et la voirie dans les 200 mètres. Ce qui serait intéressant c'est d'aller au bout de la démarche. Parce que c'est bien de dire qu'on peut prendre les transports en commun, qu'on peut descendre à tel arrêt. Mais généralement quand on se déplace ce n'est pas pour tenir les murs et rester sur le bitume, c'est pour aller quelque part. On va d'un point A à un point B et généralement un point B c'est un ERP. »

Ces trois standards répondent à des besoins complémentaires et ont été conçus sous l'impulsion de la direction ministérielle à l'accessibilité, à travers des groupes de travail réunissant des profils légèrement différents (des spécialistes de la description de la voirie ou des transports selon le sujet). Ils se sont attachés à traduire dans des schémas de données les besoins exprimés par les associations de personnes handicapées. Avec l'exemple du sens de l'escalier, information cruciale pour les personnes malvoyantes et aveugles, la même interviewée nous explique l'importance de se tourner vers les usagers afin de prendre en compte toutes les demandes. Cette demande n'a pas pu être intégrée dans le modèle de données du CNIG pour des raisons techniques, mais le sens de l'escalier (montant ou descendant depuis la porte d'entrée d'un ERP) peut être décrit dans acceslibre, ce qui permet de minimiser un vrai danger :
« C'est une illustration du fossé qu'il y a entre les personnes valides et les personnes handi. Pour la personne handi un escalier qui descend c'est un danger. L'escalier qui descend, si on le rate ça peut être le saut de l'ange. C'est pour ça qu'elles insistaient pour l'avoir. C'est extrêmement important, lorsqu'elles se déplacent, de savoir comment ça se passe à l'entrée, pour savoir s'il y a danger ou pas. Pas de danger c'est de plain-pied. Danger modéré des marches qui montent. Danger très fort des marches qui descendent. »

LES DÉFIS DE LA CONCEPTION DE STANDARDS

LES trois principales étapes de la standardisation demandent du travail, posent des défis et chacune a des conséquences sur la suivante.

Concrètement, les champs et les valeurs choisis au moment de la conception conditionnent ce que doit renseigner un producteur. Un standard comportant plusieurs dizaines de champs supposera un travail conséquent (et potentiellement décourageant) du producteur au moment de la publication de ses données, surtout en l'absence de logiciel permettant l'export automatique des données. À cela s'ajoute la collecte de données éventuellement manquantes. L'existence d'un standard, même simple, implique par ailleurs que des jeux de données déjà publiés seront jugés non conformes. La reprise de ces données peut être source de frustration et également de découragement pour leur producteur.

Afin de faciliter le parcours de production de la donnée il est donc nécessaire que les concepteurs d'un standard concertent les futurs utilisateurs, afin de s'assurer de l'adéquation de la norme aux usages. Ils doivent également choisir le format de données adéquat, ce qui n'est pas qu'une question technique : en plus de permettre d'exprimer les données, il faut qu'il convienne aux producteurs. Enfin, afin de s'assurer de l'utilisation du standard un important travail de promotion et d'accompagnement doit être effectué après sa publication. Le processus de conception est d'autant plus important qu'une fois un standard implémenté dans une multitude d'outils et de services il devient très difficile de revenir en arrière.

LES ENJEUX DE LA CONCERTATION

POURQUOI la concertation est-elle une étape importante, sinon essentielle, du processus de conception d'un schéma de données ?
Selon la chercheuse Louise Bezuidenhout, qui s'intéresse aux concepteurs de standards FAIR[39], la standardisation n'est pas qu'une stricte affaire d'expertise. Elle estime qu'il est nécessaire de réunir des expériences variées — et pas seulement des experts techniquement compétents — pour concevoir un « bon » standard de données. Ils diffèrent en effet en fonction de la composition des groupes qui les conçoivent.

> *« les approches constructivistes des données et des normes préconiseraient que la composition du groupe est d'une importance cruciale pour les standards qui émergent. L'hétérogénéité des expériences, des contextes de recherche et des contextes socio-juridiques influencent les normes qui sont fixées. »*[40]

Ici appliquée à un contexte où les groupes de travail sont essentiellement composés de chercheurs, l'argument selon lequel le standard n'est pas qu'un objet d'expertise mais le produit d'un travail collectif, et donc influencé par les expériences des acteurs impliqués dans le processus, est généralisable. Il nous apparaît en effet nécessaire de constituer avec systématisme des instances regroupant une diversité de parcours, d'expériences et de compétences, ce qui peut nécessiter de solliciter directement des personnes qui ne participeraient pas spontanément à ces groupes de travail (y compris faute de connaissance de leur existence). L'absence de représentation de certains groupes d'acteurs, particulièrement s'agissant de schémas de données pouvant influencer directement leurs expériences quotidiennes, poserait par ailleurs des questions éthiques : serait-il par exemple acceptable de laisser la conception des standards liés à l'accessibilité à de seuls « techniciens », sans impliquer à un ou plusieurs moments du processus des représentants de personnes en situation de handicap ? Une actrice des standards de données

39. Les données FAIR sont des données qui répondent aux principes de découvrabilité, d'accessibilité, d'interopérabilité et de réutilisation (findability, accessibility, interoperability, reusability). Bezuidenhout s'est en particulier intéressée aux évènements internationaux regroupant plusieurs réseaux d'acteurs s'intéressant aux principes FAIR, auxquels de nombreux chercheurs participent mais au sein desquels les acteurs venant des pays dit du Nord sont surreprésentés.
40. Bezuidenhout L., 2020, « Being Fair about the Design of FAIR Data Standards », *Digital Government: Research and Practice*, 1, 3, p. 18:1-18:7 (nous traduisons).

liés à l'accessibilité insiste sur l'importance d'impliquer les associations des personnes pour qui ils sont conçus :

> *« On a invité les associations pour faire un recueil de besoin, qu'on ne pouvait pas deviner. […] De là on a retenu les informations essentielles. C'est-à-dire que sans ces infos, tout ou partie en fonction de ses besoins, la personne ne sort pas, pour ne pas prendre de risque, ou alors elle trouve un moyen détourné d'avoir cette information. Et le moyen c'est de prendre son téléphone. […] On s'est rendu compte, avec nos interviews utilisateurs, que les personnes avaient une routine de questions. Le cas extrême, c'était une personne qui en posait quarante. acceslibre répond à trente d'entre elles. On a considéré que c'était un indicateur très intéressant pour nous, puisque ça signifiait qu'on répondait à une grande partie des attentes. »*

Au-delà de l'aspect éthique, l'implication de nombreux acteurs dans la phase de conception permet :

- De faire émerger des cas d'usage, c'est-à-dire de déterminer ce que rendra possible le standard (par exemple : mieux connaître une offre de service public sur un territoire et orienter plus efficacement les usagers, aider à l'allocation des ressources publiques…) ;
- De faire un bilan des données existantes. Trouver un « plus petit dénominateur commun », autrement dit les données que les producteurs ont en commun, permet de s'assurer que tous puissent publier relativement facilement et rapidement des données standardisées, sans avoir à en collecter de nouvelles ;
- De favoriser l'appropriation du standard après sa publication, en s'appuyant sur des acteurs susceptibles de l'adopter rapidement et de s'en faire les promoteurs au sein de la communauté.

En pratique, la conception d'un schéma mixe souvent une approche par les cas d'usage (conception selon réutilisations futures) et par les données existantes (conception selon les contraintes immédiates), tout en étant plus marquée par l'une ou l'autre.
Dans le cas du schéma des lieux de médiation numérique, les cas d'usage ont permis d'arbitrer entre des champs et des valeurs contraints par ce que les producteurs de données seraient en mesure de renseigner. OCDS est quant à lui un schéma très ambitieux, fondé sur l'idée que les données publiées doivent correspondre aux besoins des usagers plutôt qu'à la réalité des données collectées par les producteurs. Cela conduit à réorienter les

processus de collecte de données — cela a par exemple été le cas en Ukraine, où la refonte du système d'information marchés a eu des effets très concrets, mentionnés plus haut.
La concertation permet également de prendre en compte les attentes des acteurs métiers. L'échange autour des pratiques des producteurs et utilisateurs de données permet d'envisager le futur parcours de la donnée et donc de faciliter à terme l'adoption du schéma. Des ateliers participatifs dans le cadre de la conception de schéma de données des lieux de médiation numérique ont ainsi permis de prendre la mesure de l'hétérogénéité des territoires : alors que certains sont fortement structurés par des réseaux locaux de médiation numérique, d'autres ne bénéficient pas d'une structure intermédiaire permettant de rassembler les acteurs au niveau local. Cela a fait émerger la nécessité de penser plusieurs parcours de la donnée : au niveau du lieu (il publie seul ses données, dans un fichier au format CSV), du territoire (un acteur intermédiaire local assure la collecte, l'agrégation et la publication de données standardisées), et à l'échelle nationale (un formulaire national permet aux lieux de renseigner leurs données et d'apparaître sur une cartographie nationale).
La concertation permet en somme de produire un schéma qui ne répond pas « que » à des exigences techniques mais qui prend également en compte les attentes et les habitudes des acteurs métiers et des usagers/réutilisateurs finaux.

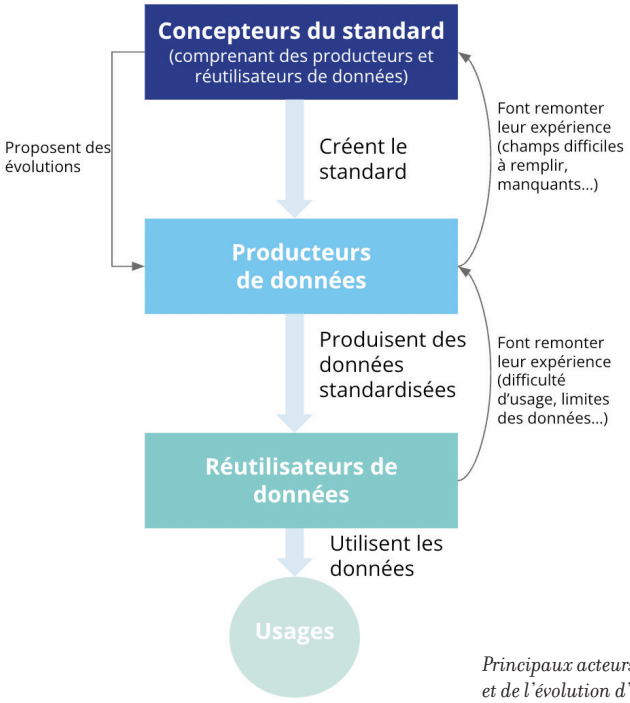
La concertation peut prendre de nombreuses formes (cumulables), parmi lesquelles :

- Des groupes de travail fermés/comités de pilotage (en présentiel et/ou en ligne) : ils regroupent une quinzaine de personnes pour la représentativité de leurs compétences et de leur expérience métier. Elles sont directement contactées par les concepteurs du standard. Nous recommandons de compléter ces groupes de travail par des ateliers ouverts ;
- Des ateliers ouverts (en présentiel et/ou en ligne) : ces groupes sont libres d'accès. Afin d'éviter un entre-soi entre personnes du réseau des concepteurs du schéma, il est important de communiquer largement et en avance sur les dates et le contenu des ateliers. Il est nécessaire d'insister sur la légitimité de chacun à participer — un risque est en effet l'auto censure des participants, qui en raison d'un faible bagage technique ne se sentiraient pas aptes à prendre part à ces ateliers, alors que leur expérience est nécessaire ;

Numérique en communs 2021 — CC-BY-SA Marion Bornaz.

- Un appel à commentaires (en ligne) : Avant publication officielle du schéma, il est fortement recommandé de permettre aux membres de la communauté d'exprimer leur avis sur un standard prototype. D'une durée allant d'environ 3 semaines à 2 mois, cet appel à commentaire permet de s'assurer une dernière fois de l'adéquation du standard aux attentes de la communauté de producteurs et de réutilisateurs de données. Cette relecture collective permet également de corriger d'éventuelles erreurs de formes (coquilles, tournures peu claires…) dans la formulation des champs et valeurs ainsi que dans la documentation. Il s'agit d'une bonne pratique de nombreuses institutions conceptrices de standards, en France le CNIG et transport.data.gouv.fr par exemple ;
- Des entretiens (en présentiel et/ou en ligne) : il s'agit d'une forme de concertation plus indirecte. Afin de faire émerger les usages et les besoins, les concepteurs peuvent mener des entretiens individuels et/ou collectifs des producteurs et réutilisateurs de données. Cette modalité ne permet cependant pas, contrairement aux ateliers et groupes de travail, la délibération collective.

L'implication de la communauté ne s'arrête idéalement pas au moment de la publication du standard. Les retours constants (par exemple via une adresse mail dédiée ou un forum en ligne), traités régulièrement par des instances de gouvernance ouverte, permettent d'éventuelles mises à jour du standard.



Principaux acteurs et étapes de la conception et de l'évolution d'un standard de données ouvertes

La concertation, particulièrement sur le temps long et un mode collaboratif (ateliers) plutôt que déclaratif (entretiens), suppose cependant des compétences particulières. Ces savoir-faire peuvent tout à fait s'acquérir par la pratique mais leur absence préalable demandera d'autant plus de travail de préparation : construire les trames de sessions, créer des supports pour s'assurer de l'implication des participants, en somme se faire expert de la participation s'apprend mais ne s'improvise pas.

Il semble difficile de s'opposer aux principes participatifs, d'autant plus que les modalités de leur réalisation sont nombreuses[41]. En pratique, mener à bien une concertation est pourtant loin d'être simple, en raison notamment d'obstacles logistiques et des frustrations que le processus en lui-même peut générer chez les participants. La première difficulté consiste à réunir un panel représentatif de producteurs et de réutilisateurs de données du domaine du standard. Il n'est d'ailleurs pas rare que la composition de ces temps de discussion et de délibération ne corresponde pas à l'idéal sur les concepteurs se représentaient, pour plusieurs raisons : la méconnaissance du

41. Et elles sont bien documentées tant par la littérature pratique que scientifique. Voir par exemple la revue de sciences sociales *Participations*.

sujet de la standardisation, menant à un sentiment d'illégitimité à prendre part aux instances participatives, les limites de la communication en ligne[42], ou encore la difficulté très concrète à réunir un nombre important de personnes pendant plusieurs heures, en particulier si on s'attache à assurer la représentation d'une diversité de profils. Par ailleurs, la conception d'un standard implique de faire des choix qui vont parfois à l'encontre de l'avis de certains acteurs. Il est donc générateur de frustrations. Un membre du comité de pilotage du schéma des lieux de médiation numérique estime même qu'un léger mécontentement général est un des signes du succès de la démarche participative :

« C'est bien que tout le monde soit un peu déçu. De toute façon tout le monde ne sera pas content du résultat. Et il ne faudrait pas que quelques-uns soient très contents, ça voudrait dire que tous les autres seraient très déçus. Il vaut mieux que tout le monde soit un peu déçu, ça veut dire que vous aurez trouvé un compromis et un niveau médian. »

Les choix opérés nécessitent cependant d'être justifiés afin d'être acceptés. Il est de la responsabilité des concepteurs de faire preuve de transparence sur le processus décisionnel en rédigeant par exemple des comptes-rendus et en les faisant largement circuler, en se rendant disponibles pour d'éventuelles questions ou remarques, en prévoyant des temps de restitutions suite à des arbitrages faits en petits comités, pendant lesquels ils répondront et traiteront les objections faites à l'égard de leurs décisions.

Dans l'ensemble, la concertation est donc un mode de fonctionnement exigeant, parfois frustrant, mais nécessaire à la conception d'un schéma effectivement utilisé par les producteurs de données à qui il s'adresse et permettant ensuite des réutilisations. Il n'existe pas de recette magique pour une concertation bien menée, les modes de celle-ci pouvant prendre de nombreuses formes en fonction du sujet et des acteurs impliqués. Elle doit cependant répondre à des principes de représentativité et de transparence.

42. Pour éviter la diffusion aux seules personnes du cercle de connaissances des concepteurs il est important de pouvoir compter sur des relais de communication, par exemple des acteurs ayant connaissances des listes de diffusion actives dans un secteur particulier, ou des membres actifs d'un réseau professionnel.

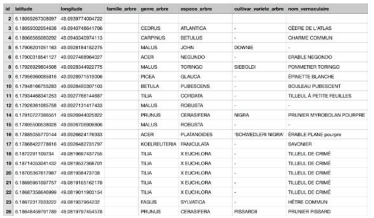
recommandations pour la phase de concertation :

- Constituer un groupe aux profils et expériences variées ;
- S'assurer de la représentation des producteurs et des réutilisateurs de données (et si possible des éditeurs de logiciels métiers) ;
- Communiquer et solliciter directement des participants potentiels (et en cas d'impossibilité de participation, les interroger sur leurs usages et besoins), en s'appuyant par exemple sur des réseaux d'acteurs déjà constitués dans le domaine du standard (associations, groupes d'usagers, réseaux professionnels…) ;
- Clarifier, en accord avec les producteurs et les réutilisateurs, les usages attendus des données ;
- Faire un bilan des données existantes et des contraintes des producteurs ;
- Penser le parcours de la donnée avec les producteurs ;
- Communiquer les arbitrages de la concertation au fur et à mesure de celle-ci, justifier tous les choix opérés ;
- Organiser un appel à commentaires sur la base d'un prototype standard ;
- Traiter les retours de façon transparente, y compris après la publication du standard ;
- Penser l'évolution du standard en collaboration avec la communauté.

LA DIFFICULTÉ DE CHOISIR LE « BON » FORMAT

À la fin du processus de conception d'un schéma, une fois les champs et les valeurs possibles déterminés, des choix techniques s'imposent. Il est nécessaire de traduire ces arbitrages en langage informatique, notamment de choisir si les données seront publiées sous un format CSV ou JSON, qui sont deux formats parmi les plus courants actuellement. Chacune de ces options comporte des avantages et inconvénients.

	CSV	JSON
En bref	Il s'agit d'un format minimaliste pour échanger des données simples, qui a été développé dans les années 1970, dans un contexte de faible interopérabilité des données. Il est donc la réponse minimale à un vaste problème contextuel. Il est facilement compréhensible par les humains mais exprime mal la complexité (notamment les relations hiérarchiques) et est ambigu pour les machines, notamment car il ne permet pas le typage des champs (contient-il des nombres ou une chaîne de caractères ?).	Ce format, développé dans les années 2000, dans un contexte de plus grande interopérabilité des données, permet de représenter des structures plus complexes de données, compréhensibles sans ambiguïté par les machines. Il est cependant moins lisible pour les humains et nécessite plus de compétences techniques pour être manipulé.
Standard(s) techniques (utilisés par le concepteur de schéma)	Table schema	JSON schema

Représentation des données	 *Extrait du jeu de données des arbres urbains de l'eurométropole de Metz — CSV*	 *Extrait du jeu de données des DECP de la région Bretagne 2022 — JSON*
Avantages	Simplicité de publication des données par le producteur et de manipulation par le réutilisateur.	Expression de données plus complexes, d'arborescence dans les données. Expression de données plus complexes, d'arborescence dans les données.
Inconvénients	Difficulté d'exprimer des données complexes — il est certes possible de lier plusieurs fichiers en s'appuyant sur des données correspondantes (des données pivots), mais cela nécessite quelques compétences en informatique Ambigu pour les machines : il est notamment impossible de préciser le type du champ, d'indiquer s'il contient par exemple des nombres ou une chaîne de caractères	Difficulté de saisie des données manuellement en cas de « faible » niveau d'expertise (importance des logiciels métiers pour aider à l'export) Manipulation des jeux de données demandant plus de compétences techniques : maîtrise d'une syntaxe complexe, de librairies spécifiques…
Un exemple : représenter les données d'une famille	Au format CSV, chaque membre de la famille sera représenté dans une ligne. Pour chaque ligne, une série de caractéristiques sont renseignées dans le même ordre : nom, lieu de résidence, âge… Chaque personne est donc décrite de manière comparable à toutes les autres. Malgré les redondances d'informations qu'il contient, le fichier sera facilement compréhensible.	Au format JSON, les données de la famille seront représentées en branches. À la racine se trouveront les éléments communs, qui ne seront donc indiqués qu'une fois. Puis des branches représenteront chaque membre, avec ses propres informations. Ce format représente donc plus simplement des informations complexes.
Exemple de schéma utilisant le format de publication	Schéma des arbres urbains	Schéma des DECP

Dans le cadre du schéma des DECP, le choix du format n'a pas été évident. La tentation du CSV, de par sa facilité, s'est heurtée à la réalité des données : un marché public peut par exemple comporter plusieurs titulaires ou être sujet à modifications. Si un marché est conclu avec cinq titulaires, le CSV devrait donc comporter cinq lignes pour un seul marché, ce qui impliquerait des fichiers particulièrement longs. Un acteur des DECP explique l'intérêt d'un format JSON, permettant une représentation arborescente, compte tenu des données :

> *« On peut certes exprimer [une hiérarchisation d'informations] sous forme tabulaire, mais ça fait beaucoup de lignes pour un même marché. Donc ce n'est pas très pratique. L'idée a donc été que le format réglementaire soit XML et JSON. »*[43]

La possibilité de publier au format CSV était cependant plébiscitée par les producteurs de données. OpenDataFrance a répondu à ce besoin, en proposant une version légèrement simplifiée du schéma. Cette version ne correspondait cependant pas aux exigences légales imposées aux collectivités, comme le rappelle le même interviewé :

> *« Ce format CSV visait à répondre à un besoin. Et il a eu un certain succès, il a été plutôt bien adopté par les collectivités. […] Mais des collectivités pensaient qu'en publiant sous en format CSV ils répondaient aux exigences réglementaires, comme si elles avaient publié sous un format XML ou JSON, alors que non. »*

Une solution alternative pour assurer un respect des obligations réglementaires tout en évitant l'utilisation du format JSON serait la publication de plusieurs CSV, liés par des données pivots, c'est-à-dire des données communes à plusieurs bases, permettant d'identifier un même objet[44]. Cette alternative suppose cependant des compétences techniques.

> *« Une autre option aurait été un format regroupant plusieurs fichiers. Au lieu d'avoir un fichier contenant toutes les données on pourrait avoir deux fichiers. Par exemple, un avec la liste des marchés publics et un autre avec la liste des titulaires de marché, liés grâce à une clef commune. Mais cette solution implique également de la complexité. »*

43. Le format XML, développé à la fin des années 1990 et permettant de représenter un grand nombre de structures de données, a été très utilisé à sa création mais l'est aujourd'hui moins que JSON.
44. Un numéro SIRET permettra par exemple d'identifier une entreprise dans n'importe quelle base de données.

Aujourd'hui, la question du format idéal n'est pas totalement tranchée. Les arguments de la simplicité du CSV sont toujours opposés à ceux de la meilleure représentation des données en format JSON.

Fondé sur la norme JSON, le format GeoJSON permet par ailleurs la publication des données géographiques. Il permet de décrire des objets géométriques (points, segments, polygones…). C'est par exemple le format choisi pour les aménagements cyclables.

PENSER LA MISE EN ŒUVRE COMME CONDITION DE SUCCÈS D'UN STANDARD

PUBLIER un standard à la suite d'un processus de conception participatif favorise l'adoption par la communauté mais ne ne la garantit pas. Une phase d'accompagnement des producteurs de données est cruciale et nécessite des moyens financiers et humains conséquents. La sous-estimation des efforts à fournir post-publication du schéma risque d'entraver sa réutilisation et donc de mener à un nombre jugé décevant de jeux de données publiés. La production collaborative d'un schéma peut paraître exigeante, mais ce n'est que le début du chemin.

Un membre du comité de pilotage du schéma des registres d'entrées d'archives, qui s'adresse à tous les services d'archives (principalement publics) de France, regrette ainsi le manque de temps et de moyens alloués à la promotion du standard, qui deux ans après sa publication reste peu adopté :

> « *L'organisation d'ateliers est restée en* stand by *faute de temps. […] Pour encourager l'adoption du schéma on aurait aimé monter un workshop avec des gens qui ont leur logiciel métier à jour, qui ont leurs données, qui viennent avec ce qu'ils ont pour qu'on voit ensemble si c'est conforme au standard, ce qu'il faut retoucher… […] Mais pour le faire il nous faut du temps, malgré tout, et ça on n'arrive pas à le dégager.* »

Quelques pratiques permettent de faciliter la mise en œuvre d'un nouveau standard.
En septembre 2022, Datactivist a organisé des ateliers collaboratifs en partenariat avec Etalab[45]. Lors de ceux-là, les participants ont suggéré qu'une documentation à destination des producteurs, comprenant notamment des exemples de cas d'utilisation, était un facteur de succès important pour un schéma de données. Ils pointaient également l'importance de s'appuyer sur des promoteurs/ambassadeurs pouvant partager des retours d'expérience, ainsi que sur des éditeurs de logiciel impliqués dès la conception du standard.

45. Les organisations suivantes étaient représentées dans ces ateliers, nous les remercions chaleureusement pour les contributions : Etalab, transport.data.gouv.fr, OpenDataFrance, Futuro Cité, Nord Ouvert, Open Contracting Partnership, le CEREMA, data.inclusion, La Fabrique des Mobilités Québec, La MedNum, multi.coop, Colin Maudry.

PRINCIPAUX FACTEURS DE SUCCÈS D'UN STANDARD DE DONNÉES OUVERTES

CONTEXTUELS

- Moyens matériels et humains sur le long terme ;
- Soutien d'organisations faisant autorité dans le domaine du standard ;
- Encouragement institutionnel, voire obligations légales ;
- Instances de gouvernance du standard claires, au fonctionnement transparent.

COMMUNAUTAIRES

- Conception concertée dans des instances représentatives ;
- Participation d'acteurs métiers, techniques et d'éditeurs de progiciels (selon les secteurs) ;
- Interlocuteurs identifiés pour accompagner la production de données ;
- Acteurs relais/ambassadeurs du standard.

COMMUNICATIONNELS

- Communication le long du processus de conception ;
- Présentation publique du standard ;
- Documentation comprenant si possible des cas d'utilisation du standard et des cas de réutilisations de données standardisées ;
- Diffusion de retours d'expérience de réutilisateurs du standard ;
- Documentation du suivi de l'adoption.

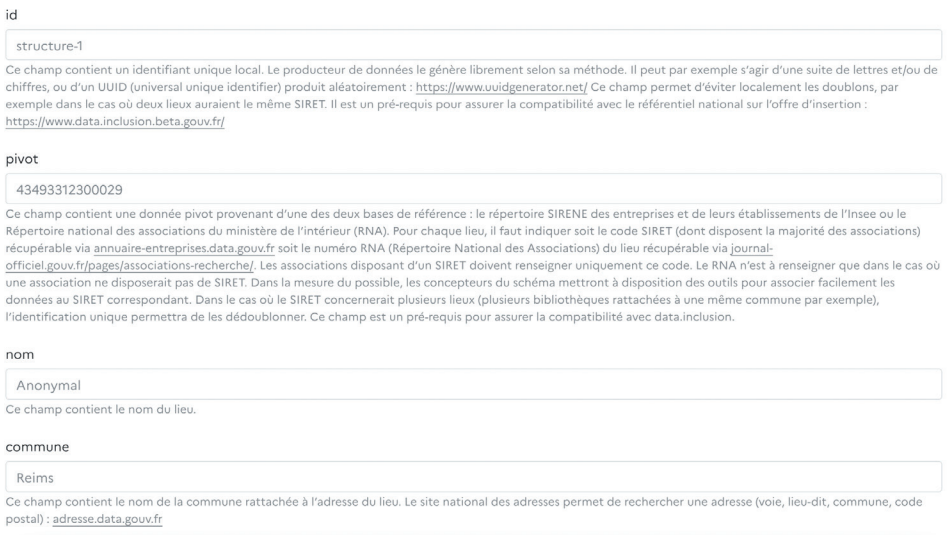
TECHNIQUES

- Utilisation de champs standards (respect de normes quand cela est possible, par exemple pour la description des horaires d'ouverture) ;
- Outillage pertinent (outils d'aide à la saisie, de validation…) ;
- Parcours de saisie des données pertinent au regard des habitudes des producteurs de données.

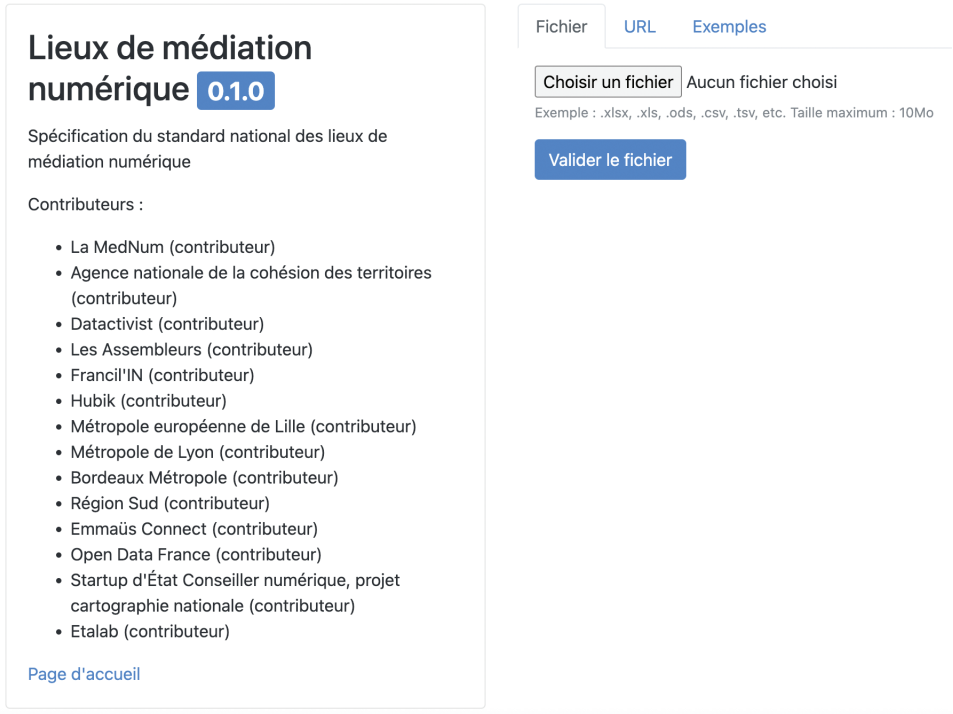
Nous recommandons ainsi de mettre à disposition en ligne une documentation, en plus de celle concernant les champs et valeurs du schéma, comprenant :
la méthodologie de conception
un guide d'aide à la production de données, sous forme de FAQ par exemple
Ce dernier peut renvoyer vers des outils, comme publier.etalab.studio ou Validata, qui facilitent la saisie et le contrôle des jeux de données.

deux outils d'aide à la saisie

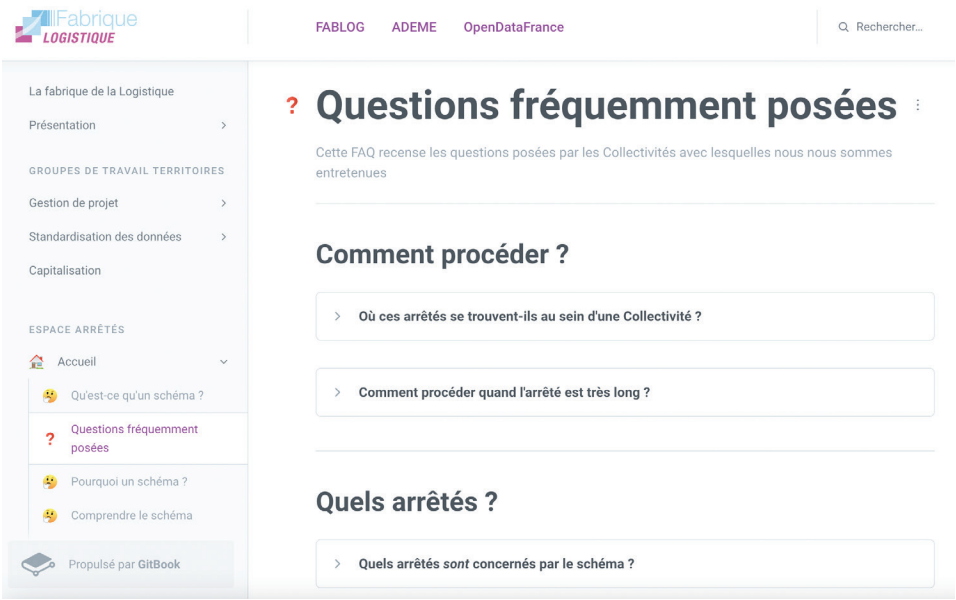
publier.etalab.studio : Cet outil développé par Etalab permet aux producteurs de données de saisir ou de charger leurs données en vue d'une publication sur la plateforme data.gouv.fr. Il propose un accompagnement sous trois formats : chargement des données déjà existantes, saisie via formulaires, saisie via tableurs. Ci-dessous, un exemple de formulaire généré à partir du schéma de données des lieux de médiation numérique, permettant au producteur de données de s'assurer de la conformité de son jeu de données.



Validata : Il s'agit d'un outil de validation développé par OpenDataFrance. Initialement créé pour les schémas du Socle Commun des Données Locales, il permet aujourd'hui de s'assurer qu'un jeu de données est conforme à n'importe quel schéma publié sur schema.data.gouv.fr ou publié en ligne au format JSON. Contrairement à publier.etalab.studio, l'outil ne propose cependant pas d'aide à la saisie de données en vue d'une publication directe sur data.gouv.fr.



La publication du schéma de données des arrêtés permanents de circulation en ville pour le transport de marchandises[46] a par exemple été accompagnée d'une documentation dense, expliquant entre autres l'intérêt d'un schéma de données, le mode de conception de celui-ci et donnant accès aux supports des réunions. Elle propose également une série d'exemples de jeux de données standardisés et une aide au remplissage de certains champs. Des data visualisations (cartographies) ont aussi été intégrées pour illustrer des réutilisations possibles des données standardisées.



46. Les arrêtés de circulation réglementent l'accès à des rues pour certains types de véhicules (poids lourds…) et certains usages (livraison, urgence…).
Ce schéma a été conçu par le CEREMA, la Région Sud, OpenDataFrance et Etalab. La documentation est accessible en ligne : https://opendatafrance.gitbook.io/fablog/espace-arretes/accueil

Captures d'écran du gitbook du schéma des arrêtés permanents de circulation — 10/08/2022.

Des temps d'échange avec les producteurs de données sous forme de webinaires/workshops en ligne sont également recommandés. En plus de présenter formellement le schéma, ils peuvent permettre de travailler concrètement au formatage des données — ne serait-ce qu'en comparant les jeux de données déjà publiés aux attendus du standard, afin de clarifier les modifications nécessaires.
Dans le cas où les données seraient produites par un grand nombre d'acteurs répartis sur tout le territoire, il est intéressant de s'appuyer sur des intermédiaires territoriaux pour aider à la diffusion du standard et éventuellement à la production de données. Dans le domaine de la médiation numérique, les collectivités sont ainsi moins productrices de données que relais essentiels de la promotion de la standardisation. Dans celui de la mobilité, les régions ont un rôle important d'animation de la politique open data et d'aide au partage d'expérience.
Au moment de la conception du standard il est par ailleurs important d'avoir des échanges avec les éditeurs de logiciels métiers, s'il en existe dans le domaine. L'intégration rapide du schéma à ces progiciels est le moyen le plus sûr de s'assurer de l'export des données au bon format. Le dialogue entre concepteur et éditeurs est généralement facilité s'il peut se reposer sur l'intermédiation des clients de ces entreprises, généralement les collectivités. Un acteur du schéma des DECP explique la difficulté de contraindre les

éditeurs de logiciel autrement que via leurs clients : « *Il était important que les prestataires adoptent ce format [de données]. Le problème c'est que la seule autorité vis-à-vis des éditeurs c'est leurs clients, qui sont généralement les communes. Même les régions n'entretiennent pas forcément un lien très équilibré avec eux et subissent souvent leurs prestataires.* »
Il importe de considérer cette phase de mise en œuvre du standard en amont du projet, de prévoir du temps et un budget. Il faut en somme envisager un projet de standardisation sur le temps long plutôt que s'arrêtant au moment de la publication d'une première version d'un standard.

recommandations pour la mise en œuvre d'un standard :

- Prévoir du temps et des moyens financiers post-publication du standard ;
- Présenter publiquement le standard ;
- Publier une documentation présentant la démarche de conception et des cas concrets d'utilisation ;
- Solliciter les participants à la concertation pour se faire relais du standard ;
- Proposer une aide à la saisie des données (sur demande et/ou lors d'ateliers) ;
- Documenter l'évolution de l'adoption du standard, communication sur les réutilisation de données standardisées.

PRODUIRE DES DONNÉES STANDARDISÉES

LES standards imposent un travail important aux producteurs de données publiques ouvertes[47]. Reprendre des données selon les spécifications d'un standard, ou produire de nouveaux de jeux de données demande en effet du travail, souvent peu visible. Dans cette dernière partie, nous vous proposons de faire un point sur ce qu'implique la standardisation des données côté producteur et sur les manières de faciliter ce processus.

47. Goëta S., Davies T., 2016, « The Daily Shaping of State Transparency: Standards, Machine-Readability and the Configuration of Open Government Data Policies », *Science, Technology and Innovation Studies*.

STANDARDISER DES DONNÉES DÉJÀ PUBLIÉES

Il arrive parfois qu'à la publication d'un nouveau standard une collectivité, ou un acteur ayant à disposition des données d'intérêt public, ait déjà entamé une démarche d'open data. Il peut tout de même se conformer à ce standard — voire le *doit* s'il s'agit d'un standard réglementaire, par exemple relatif à la mobilité collective.
Cela nécessite parfois un important travail de reprise des données et/ou de collecte de nouvelles données. Celui-ci peut être facilité par l'intégration rapide d'un schéma aux progiciels, si tant est qu'il en existe dans le domaine, et donc à la création de fonctions d'export de données formatées. Par ailleurs, le développement d'outils de diagnostic à destination du producteur de données pourrait permettre d'objectiver et de faciliter le travail nécessaire à la standardisation des données, en listant les informations manquantes mais nécessaires au remplissage des champs (en particulier des champs obligatoires) et en indiquant des sources possibles pour les collecter.
Si le manque de données peut être un problème pour le producteur, il existe le cas où le standard serait moins exigeant que le format qu'il utilise actuellement. Dans ce cas, la collectivité peut être tentée de ne pas s'y conformer, nous l'avons mentionné précédemment. Il est cependant parfois possible de l'adopter sans pour autant renoncer totalement à renseigner des données supplémentaires. S'agissant des lieux de médiation numérique, le standard conçu par La MedNum est par exemple moins précis dans la description des services proposés que le format utilisé par Bordeaux Métropole. Afin d'apparaître dans le guide métropolitain des lieux ressources pour l'inclusion numérique les lieux doivent actuellement indiquer leur niveau de maîtrise pour 22 services possibles, répartis en 4 catégories[48]. Le standard exige quant à lui à ce que les lieux listent les services proposés parmi 15 valeurs admises, sans préciser leur niveau d'expertise. Il serait possible pour la collectivité, qui tient à conserver un niveau de précision plus élevé, de prendre le schéma pour base tout en ajoutant des champs supplémentaires. Il est en effet possible de faire coexister des données standardisées et des données locales plus complètes. Cela nécessite un travail non négligeable mais permettrait

48. Les trois niveaux sont : expert : c'est notre cœur de métier (3 étoiles) ; maîtrise : sans être notre spécialité, nous maîtrisons ce service (2 étoiles) ; basique : nous répondons aux besoins les plus simples mais ne pouvons pas aller loin dans la maîtrise de ce service (1 étoile).

aux lieux d'être référencés au niveau national tout en étant plus finement documentés dans un guide et/ou une cartographie métropolitaine.

> *« [À Bordeaux métropole] on est rentré dans les détails. On a utilisé le référentiel APTIC, avec les 126, ou 136, typologies de services et on en a extrait 22 qui nous paraissaient les plus demandées, les plus utilisées. […] Quand j'ai soulevé la question d'avoir plus de services en COPIL [de conception du schéma des lieux de médiation numérique] pas mal de collègues ont trouvé que c'était très compliqué 22 items, que quelques-uns suffisaient. Donc on se retrouve un peu tous sur les usages de base, sur l'insertion professionnelle, les démarches en ligne, l'aspect créatif, mais nous ce qui nous intéresse vraiment, et on le perdra pas à Bordeaux, c'est le deuxième niveau. On s'adaptera au schéma le plus possible, mais on perdra pas ce niveau de détail. […] On restera compatible avec le format national, tout en l'enrichissant avec nos propres données. »*

CE LIEU PROPOSE UN ACCOMPAGNEMENT POUR :

Découvrir les usages de base du numérique

	Niveau d'expertise
Fonctionnement tablette ou smartphone	
Fonctionnement d'un ordinateur	★★☆
Internet : fonctionnement navigation, recherches web	★★☆
Mails : créer, envoyer, recevoir	★★☆

S'insérer professionnellement

Réaliser son CV	★★☆
Diffuser son CV en ligne	★★☆
Organiser sa recherche d'emploi	★★☆
Découverte et usage de l'emploi Store	★☆☆

Faire ses démarches en ligne

Faire ses déclarations pôle emploi	★★★
Déclarer ses revenus et découvrir des services proposés par les impôts	★★☆
Accéder à ses droits sociaux et les gérer (RSA, CAF)	★★★
Ouvrir et gérer son dossier de retraite	★★☆
Gérer ses droits d'assuré social en ligne/sur internet (ameli.fr)	★★★
Gérer son abonnement et ses factures d'électricité/gaz	★★☆
État civil : titre de séjour, carte identité, passeport	★★☆
Permis et carte grise	★★☆

Créer avec le numérique

Images : retoucher ses photos	☆☆☆
Découverte et utilisation imprimante 3D	☆☆☆
PAO : faire des présentations, des diaporamas	☆☆☆
Créer un site web de base	☆☆☆
Découvrir et expérimenter la programmation informatique (code)	☆☆☆
Découvrir et participer à des MOOCs (formation en ligne)	☆☆☆

*expert : c'est notre cœur de métier (3 étoiles), maîtrise : sans être notre spécialité, nous maîtrisons ce service (2 étoiles), basique : nous répondons aux besoins les plus simples mais ne pouvons pas aller loin dans la maîtrise de ce service (1 étoile).

Capture d'écran du Guide des lieux ressources numériques de Bordeaux Métropole — le 12/10/2022.

ACCOMPAGNER LA PRODUCTION DE NOUVELLES DONNÉES

DES entretiens avec des concepteurs de schémas et des réutilisateurs ont soulevé l'importance de l'accompagnement post-publication d'un standard. Lors des ateliers mentionnés précédemment, les participants ont également estimé qu'une documentation à destination des producteurs était un facteur de succès important pour un standard de données.

Face à une nouvelle obligation réglementaire et/ou conscients de l'intérêt de la publication de données standardisées, les producteurs ont parfois besoin d'aide, qui peut être apportée par une documentation écrite ou via la sollicitation directe d'« experts » du schéma. Responsable de la publication des données en open data de sa collectivité, une interviewée regrettait par exemple le peu d'aide à la production de données respectant le format réglementaire des DECP. Cela a dû être compensé par un effort de la part du producteur, notamment pour identifier les sources de données vers lesquelles se tourner : « *Ça manquait de documentation, la description des données à fournir n'était pas une évidence. On ne savait pas où se sourcer pour produire de la donnée, ça nous a demandé du travail.* »
Consciente de la difficulté que peut causer la production de données normalisées, l'équipe de transport.data.gouv.fr propose un accompagnement aux collectivités qui le souhaitent, en plus d'organiser des webinaires pour les schémas qu'elle estime complexe. Miryad Ali résume ainsi l'aide apportée par son équipe (qui produit par ailleurs une documentation conséquente : « *S'il y une certaine complexité dans le schéma on va faire un webinaire. Mais même pour les schémas faciles à prendre en main, par exemple pour les lieux de covoiturage, qui demandent juste un CSV, on envoie un mail en disant bien aux collectivités de revenir vers nous en cas de questions. Alors ça nous arrive de faire plusieurs appels avec les collectivités.* »

QUELLES ALTERNATIVES À LA STANDARDISATION ?

Si elle rend dans l'ensemble possible une meilleure interopérabilité des données, la standardisation n'est pas toujours la solution la plus indiquée pour favoriser la collecte et le partage de données. Cela est particulièrement vrai dans le cas de grandes bases de données constituées sur un mode collaboratif (grâce au *crowdsourcing*), c'est-à-dire sur la base de la bonne volonté d'individus n'ayant parfois que des informations limitées à disposition et des compétences techniques très inégales — des exigences strictes en termes de format de données peuvent donc être particulièrement dissuasives et entraver la collecte. Dans ce cas, les gestionnaires de la base de données peuvent estimer que des données hétérogènes, parfois incomplètes et collectées sur une temporalité longue valent mieux que pas de données du tout. C'est par exemple le cas de la base Open Food Facts, qui regroupe des données de produits alimentaires commercialisés dans environ 150 pays.

open food facts : toute contribution est bonne à prendre

Plus de 25 000 contributeurs ont permis de répertorier plus de 1,7 million de produits alimentaires dans la base de données internationale de l'association Open Food Facts. Ce répertoire de produits donne des informations sur leur composition, leurs qualités nutritionnelles (avec indication du nutri-score) ou encore leur niveau de transformation (en se basant sur l'indicateur expérimental Nova). En créant un compte sur le site, n'importe qui peut ajouter un produit en prenant une photo du code-barre, en renseignant directement celui-ci, ou en remplissant un formulaire avec des informations tirées de l'emballage et de l'étiquette du produit. Dans ce formulaire, aucun champ n'est obligatoire et la plupart des valeurs sont libres. Des recommandations de format sont parfois faites mais la conformité de la valeur renseignée à celles-ci n'est pas vérifiée. Des directives sont par exemple données pour le champ « liste des ingrédients » (même s'il est possible de ne pas les suivre), alors que l'origine du produit et/ou de ses ingrédients peut être renseignée totalement librement, selon les formules très variées rédigées sur les emballages. Quelques algorithmes contrôlent la qualité des données mais ils sont en pratique peu contraignants (ils permettent par exemple d'empêcher l'édition d'une fiche qui remplacerait la valeur nutritionnelle du produit par « 0 »).

Capture d'écran d'openfoodfacts.org — le 12/10/2022.

Le producteur de données est donc orienté mais reste libre de renseigner ce qu'il peut, avec l'assurance que la fiche produit sera publiée immédiatement, même si elle est en partie vide. Cela conduit à une base aux données très hétérogènes mais qui serait probablement beaucoup moins dense si elle était construite selon un standard exigeant plus d'efforts de la part des multiples producteurs bénévoles de données.

Le recours à des API (*applications programming interfaces* ou, interface de programmation d'application en français) peut également permettre au réutilisateur de données d'en collecter en provenance de différentes sources. En France, l'API Entreprise permet par exemple aux administrations d'accéder aux données et documents administratifs des entreprises et des associations issues de différentes sources, notamment l'INSEE, la direction générale des finances publiques et l'URSSAF. Elle permet donc à son utilisateur de réutiliser facilement des données pourtant non standardisées. Si le recours aux API est pertinent dans de nombreux cas, il n'est pas toujours une solution idéale, par exemple dans le cadre du contrôle citoyen de la vie politique. L'API Civic Information développée par Google permet ainsi d'agréger des données relatives aux élus états-uniens, mais elle est loin de collecter des informations exhaustives pour les presque 90 000 collectivités du pays, comme le rappelle James McKinney, fondateur de l'organisation sans but lucratif canadienne Nord Ouvert, dans un billet de blog défendant l'utilisation d'un standard de données :

> *« Le défi, pour le contrôle du corps législatif, est qu'il n'y a pas que des centaines d'administrations dans le monde ; il y en a des centaines de milliers. Aux États-Unis seulement, il y a 89 000 gouvernements locaux. Google a une API pour les élus, mais elle est loin de recenser tous les gouvernements locaux aux États-Unis. Y parvenir est un problème pour tout le monde, y compris Google. »*[49]

Il estime qu'une alternative possible à la collecte de données provenant de différentes bases de données, autre que le *crowdsourcing* et la standardisation (lesquelles reviennent à confier une charge de travail à un grand nombre de personnes ou au producteur de données), est le recours à l'intelligence artificielle. Dans le cadre de la collecte d'information sur les élus, une machine va ainsi rassembler les *« noms, les photos et les informations de contact des élus à partir de pages web »*[50], plutôt que de bases de données organisées. Le grand nombre d'informations disponibles sur internet implique cependant un manque de fiabilité de celles-ci. McKinney précise ainsi qu'il ne *« [connait] personnes qui utilise l'IA dans les cas nécessitant une fiabilité à 100 % »*[51].

---

49. McKinney, J., octobre 2018, « Monitoring Legislatures: The Long Game », *Medium*, https://medium.com/@jpmckinney/monitoring-legislatures-the-long-game-cefa4904ee81 (nous traduisons).
50. *Ibid*.
51. *Ibid*.

La production de données peut donc être confiée à une vaste communauté plutôt qu'à quelques acteurs. Quant à la collecte de données, en l'absence de standards elle peut s'effectuer en recourant à des API pour agréger différentes sources de ou à des l'IA pour collecter des informations sur internet. Si ces alternatives sont loin d'être toujours idéales et nécessitent un travail et parfois des compétences de la part des réutilisateurs, elles confirment que des données publiées, même non standardisées, sont toujours précieuses et socialement utiles.

CONCLUSION

LA standardisation des données est un processus essentiel pour accroître les impacts de l'open data. Le chemin vers l'homogénéisation de données largement dépendantes de leur contexte de production est cependant pavé de défis, allant de la conception de standard à leur adoption par les producteurs de données.

Les difficultés mentionnées dans ce cahier sont accrues par une relative méconnaissance du travail nécessaire à la standardisation. Celui-ci est souvent mal documenté, sous-estimé, et repose beaucoup sur la bonne volonté d'acteurs occupés à d'autres projets par ailleurs. La standardisation ne fait en effet pas figure d'exception au sein de la chaîne du travail des données et demeure une étape encore peu visible.

S'il est impossible de fournir une méthode assurant dans l'absolu le succès d'une démarche de standardisation, l'allocation de moyens financiers et humains et une gouvernance des standards ouverte, impliquant sur le temps long les producteurs et réutilisateurs de données, nous semblent être parmi les premiers facteurs de réussite. Il est peut-être même souhaitable de penser les standards de données comme des communs, qui pour répondre aux besoins de toutes les parties prenantes devraient d'abord rassembler et non s'imposer de fait, évoluer selon des modes de fonctionnement acceptés par des communautés ouvertes à tous[52].

52. Sur les principes des communs, voir par exemple le projet « géocommuns » porté par l'IGN. https://www.ign.fr/la-demarche-geocommuns. Pour aller plus loin, voir les travaux de recherche de Sébastien Shulz sur les communs numériques : Shulz S., 2021, *Transformer l'État par les communs numériques : Sociologie d'un mouvement réformateur entre droit, technologie et politique (1990-2020)*, thèse de doctorat en sociologie, Université Gustave Eiffel.

EN GUISE D'OUVERTURE : L'ENJEU DE LA STANDARDISATION DES DONNÉES OUVERTES AU PRISME DES TERRITOIRES INTELLIGENTS

LES territoires intelligents gagneraient, selon les acteurs interrogés dans le cadre d'une étude pour le compte de la direction générale des entreprises (DGE), à laquelle Data Publica a contribué, à s'appuyer sur des normes et standards dans chaque domaine d'action publique. Le cas de la standardisation des données ouvertes donne ainsi des pistes sur la façon dont pourraient être produites de nouvelles normes techniques, quel que soit leur secteur d'application. Les enjeux mentionnés au fil de ce cahier concernant la fabrication et la mise en œuvre de standards de données ouvertes soulignent l'importance de concevoir la standardisation et l'interopérabilité comme un objet nécessitant des instances de concertation ouvertes et des mécanismes de gouvernance clairs.
Si les normes au sens large pourraient aider au développement de territoires intelligents, celui-ci repose également sur une meilleure interopérabilité. Elle est en effet est perçue comme un prérequis par les acteurs de territoires intelligents[53]. La standardisation des données, en particulier ouvertes, est un outil de cette interopérabilité mais il en existe d'autres, parmi lesquels : l'interopérabilité des outils, l'homogénéisation de la sémantique et l'utilisation de protocoles de communication.

L'INTEROPÉRABILITÉ DES OUTILS

L'interopérabilité des données, ouvertes ou non, favorise le développement de territoires intelligents. Mais celle des outils est également essentielle. L'infrastructure technique d'un territoire intelligent est en effet généralement composée de trois couches :

- « inférieure » (outils de captation et de production des données : logiciels métiers, capteurs, IoT, applications diverses)
- « intermédiaire » (outils de stockage et d'échanges : plateforme de données, datalake, cloud…) ;

53. Data Publica, KPMG, 2021, « De la smart city à la réalité des territoires connectés. L'émergence d'un modèle français », Etude pour le compte de la DGE.

- « supérieure » (services et applications :
  outils de visualisation, applications de services numériques,
  outils d'analyse…).

L'un des principaux enjeux des territoires intelligents est de faire communiquer ces différentes couches, donc de rendre les outils et pas seulement les données interopérables.

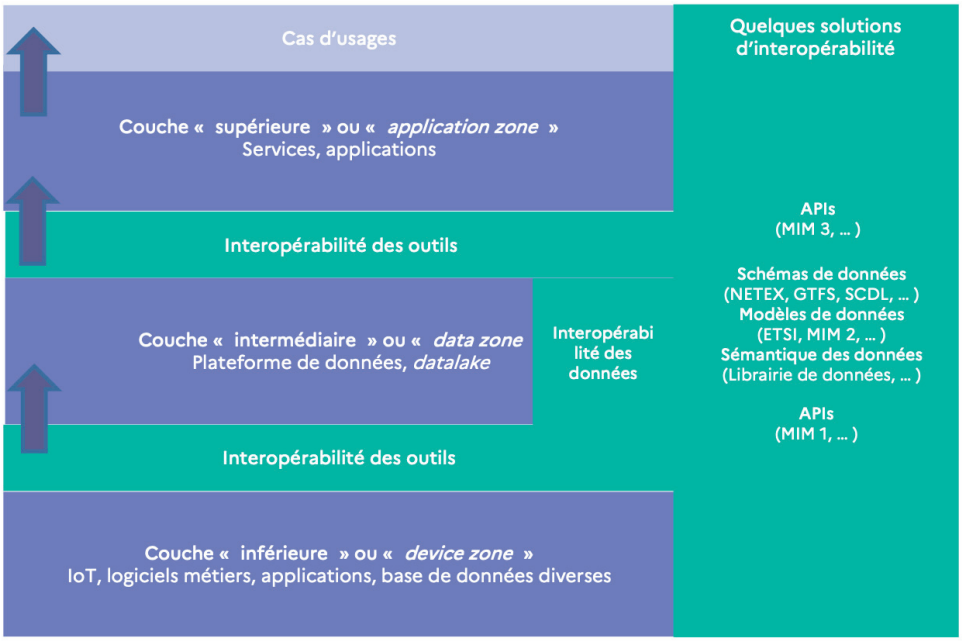
HOMOGÉNÉISER LA SÉMANTIQUE ET LES RÉFÉRENTIELS

L'interopérabilité des données et des outils est essentielle aux infrastructures techniques de chaque territoire. Mais un problème général se pose : à l'heure où la thématique de la ville intelligente s'est internationalisée, chaque territoire s'appuie sur des spécificités et une sémantique propre, ce qui entrave l'interopérabilité non pas au sein des systèmes et des infrastructures mais entre eux. Entre deux territoires, parfois même au sein d'un même espace national, un même objet (une route, un trottoir, un vélo…) peut être nommé de plusieurs façons. En l'absence d'un vocabulaire commun il est donc impossible pour les systèmes de communiquer.
Il apparaît donc nécessaire d'approfondir des réflexions sur l'ontologie des territoires intelligents au niveau français, européen, voire mondial.

LES PROTOCOLES DE COMMUNICATION, UN COMPLÉMENT À LA STANDARDISATION DES DONNÉES

La standardisation des données est une brique essentielle de l'interopérabilité des territoires : cela permet de les faire communiquer entre les différentes couches d'une infrastructure technique. Mais il existe d'autres outils que les standards pour échanger des données, notamment les protocoles de communication. Comme évoqué précédemment, les API (*Application Programming Interface*) permettent par exemple de faire communiquer deux logiciels sans avoir à développer de manière spécifique d'intégration entre les deux. Les auteurs de l'étude DGE estiment qu'elles sont désormais un incontournables des territoires intelligents :

« Elles permettent notamment d'interconnecter la couche "inférieure" et la couche "intermédiaire" ainsi que la couche "intermédiaire" et la couche "supérieure" d'une infrastructure technique de territoire intelligent » (p. 151)
Alors que certains territoires associent l'interopérabilité en priorité à la standardisation des données, d'autres travaillent avant tout au déploiement de protocoles de communication de plus en plus performants, notamment entre des capteurs de données (couche « inférieure ») et des plateformes de stockage (couche « intermédiaire »).



Représentation d'une infrastructure technique de territoire intelligent et solutions d'interopérabilité — Source : Data Publica-KPMG.

La priorité donnée aux standards ou aux protocoles de communication dépend bien sûr des objectifs, compétences, choix technologiques ou encore du réseau d'acteurs d'un territoire, dont dépendent l'infrastructure technique et les évolutions de celle-ci.

En bref, l'interopérabilité au sein et entre les territoires est un sujet complexe, qui repose sur des données standardisées mais pas uniquement. De plus amples recherches permettraient alors de mieux en cerner tous les enjeux et ainsi d'accompagner au mieux les évolutions des infrastructures des territoires intelligents.

REMERCIEMENTS

L'ÉQUIPE de l'Observatoire Data Publica tient à remercier celles et ceux qui ont participé aux ateliers en ligne des 8 et 15 septembre 2022, organisés par Datactivist avec l'appui d'Etalab. Ces temps d'échange réunissant acteurs français et nords-américains de la standardisation des données ouvertes ont nourri nos réflexions et enrichi notre propos.
Nous remercions également les personnes qui nous ont accordé de leur temps dans le cadre d'entretiens. À travers le partage de leurs expériences ils sont les premiers contributeurs de ce cahier.

L'OBSERVATOIRE DATA PUBLICA

CRÉÉ en janvier 2020, l'Observatoire Data Publica est une association loi 1901. Il s'appuie sur l'expertise de ses membres fondateurs, les cabinets de conseil CIVITEO, DATACTIVIST et INNOPUBLICA et le cabinet PARME Avocats. Pionniers de la gestion publique des données en France, ils ont souhaité mettre en commun et rendre disponibles des savoir-faire acquis auprès de collectivités et d'administrations publiques dans un cadre ouvert et non lucratif.

L'Observatoire Data Publica a été créé pour observer les pratiques nouvelles de gestion publique des données : émergence de « services publics locaux de la donnée », chartes éthiques, formes innovantes de gouvernance et de management de la donnée, prototypes de datascience et usages inédits d'algorithmes, recours à l'intelligence artificielle, etc.

L'Observatoire produit des récits et des analyses, il facilite les retours d'expérience et propose des enseignements sur ces nouveaux usages de la donnée. Ses fondateurs veulent faire de cette connaissance un bien commun.

Des études sont publiées dans *Les Cahiers de l'Observatoire* qui traitent de sujets souvent inédits ou proposent une grille de lecture nouvelle sur des problématiques connues. Les cahiers s'inscrivent dans la continuité des recherches les plus récentes au service d'une mise en œuvre opérationnelle et pragmatique.