GOUVERNEMENT

Liberté
Égalité
Fraternité

MISSION BOTHOREL

Pour une politique publique de la donnée

DÉCEMBRE 2020

Mission confiée par le Premier ministre

du 22 juin 2020 au 22 décembre 2020

Composition de la mission

Éric Bothorel, député des Côtes-d'Armor

Stéphanie Combes, directrice du Health Data Hub

Renaud Vedel, coordonnateur national pour l'IA

Avec l'appui de

Nicolas Amar, coordonnateur national adjoint pour l'IA

Lorien Benda, cheffe de projet au Health Data Hub

Delphine Chaumel, inspectrice des affaires sociales

Jean-Marie Chesneaux, inspecteur général de l'éducation, du sport et de la recherche

Olivier Dissard, chargé de mission données au commissariat général au développement durable

Maxime Donadille, collaborateur parlementaire

Florence Gomez, inspectrice des finances

Matthias de Jouvenel, administrateur civil au conseil général de l'économie

Robert Picard, ingénieur général des mines au conseil général de l'économie

Sophie Planté, inspectrice de l'administration

Avec la précieuse contribution de Jean-Baptiste Auger, Simon Chignard, Bastien Guerry, Mathilde Hoang, Perica Sucevic et Romain Tales.

Conditions de publication et de diffusion

Sommaire

Synthèse

Notre pays a besoin de plus d'ouverture – sous toutes ses formes : ouverture des données publiques (*open data*), mais aussi partage et accès sécurisé aux données sensibles. La France s'est placée à l'avant-garde européenne de la politique de la donnée et des codes sources depuis 2013, mais l'avance acquise est fragile. Cette politique est aujourd'hui bloquée dans un débat inapproprié, « pour ou contre l'ouverture », qui conduit à changer d'objectif alors qu'il faut changer de méthode. Cette inertie aboutit à une perte de chances pour notre société et notre économie, alors que le potentiel de connaissance et d'innovation est immense.

Ce rapport propose des réformes ambitieuses, notamment pour participer aux transformations en cours au niveau européen, mais aussi un grand nombre de mesures très raisonnables, qui n'appellent aucun « grand soir » et sont des actions réalisables sous un an, pouvant avoir des effets retentissants pour l'avenir. Si le gouvernement n'anime pas cette politique, la France manquera une occasion majeure de renforcer tout à la fois la confiance dans l'action publique, l'efficacité des politiques publiques, et la connaissance et l'innovation dans l'ensemble de l'économie, à un moment où la crise sanitaire et économique et le résultat du Grand Débat National en ont pourtant révélé le besoin impérieux.

Comme le dispositif des entrepreneurs d'intérêt général l'a montré, en permettant l'embauche de talents dans les administrations pour résoudre des problèmes concrets, un engagement même modeste peut produire des résultats considérables, quand il déploie la puissance de la donnée et des codes sources. Comme la loi pour une République numérique de 2016 l'a montré, la France peut aussi revoir plus fondamentalement les principes de son droit et maintenir son rang de patrie d'audace et d'innovation. Le présent rapport expose ce choix tel qu'il s'offre aujourd'hui à notre pays.

Il faut le redire : la politique de la donnée est utile à tous

L'intérêt de l'ouverture de la donnée et des codes sources a encore besoin d'être affirmé et démontré, même après la crise de la Covid19, où la preuve a été faite de l'importance de la donnée pour nos politiques publiques.

De nombreux acteurs, en particulier au sein de l'État, ne comprennent pas cette politique et ses objectifs : « on ne nous dit pas pourquoi il faut faire de l'*open data* ». Ils ne perçoivent pas non plus l'impact des réutilisations des données et des codes, qui n'est pas mesuré. Au mieux, l'ouverture est perçue comme une obligation ; au pire, les acteurs ne se sentent pas concernés : ainsi, une direction dit ne pas traiter de données « car [elle] n'est pas un service statistique ».

La donnée et les codes source ne sont pas juste un enjeu « tech », mais d'abord et surtout une question politique, démocratique, scientifique et économique. À cet égard, les prémices d'une politique de la donnée sont à rechercher dans la loi de 1978 posant les bases du droit d'accès aux documents administratifs, dont font partie les données et les codes source.

Scientifique, la donnée est un vecteur de connaissance, par le partage des données et des codes exploités par les chercheurs : la « science ouverte ». Si cette culture du partage entre équipes de recherche était mieux ancrée, la prise en charge et le traitement de la Covid19 auraient été certainement plus performants et plus réactifs pendant la crise, par exemple. Plus largement, dans l'ensemble des domaines de connaissance, la donnée est aussi la condition *sine qua non* des technologies d'intelligence artificielle, dont nous commençons tout juste à apprécier le potentiel.

Économique, la donnée est un levier d'innovation. L'ouverture de la base des valeurs foncières (DVF) en 2019 l'a prouvé en permettant la création de nombreux services et entreprises innovantes, notamment françaises, sur l'analyse des prix de l'immobilier. Une étude de la Commission européenne de 2019 chiffre la valeur de l'*open data* conduit à ce jour en France à 28 milliards d'euros. L'utilisation des logiciels libres est aussi un facteur de croissance, et devrait être le principe même d'une politique d'innovation interne de l'action publique. L'ANSSI est un des fleurons dans ce domaine et fait la démonstration que le partage du code n'est pas un facteur de vulnérabilité des systèmes d'information.

Démocratique, la donnée améliore le service public en interconnectant la puissance publique et l'usager. Les citoyens de Taïwan disposent d'une plateforme pour gérer leurs données partagées avec l'administration, décider de les ouvrir ou non à un service, et mettre à jour des justificatifs pour tous les guichets en un seul clic. Ce service ne pourrait-il pas être imaginé en France ?

Politique, la donnée est un moyen puissant de restaurer la confiance dans l'action publique. Le défi est immense : la consultation publique conduite par la mission a révélé une défiance majeure dans la donnée utilisée par l'État. À cet égard, la crise de la Covid19 a provoqué un éveil de la nation à la donnée. Les difficultés pour établir des statistiques en temps réel de l'épidémie, les conflits d'interprétation des courbes, la fragilité des hypothèses des modèles épidémiologiques, auront eu une vertu pédagogique : celle de révéler que toute donnée est une construction et qu'elle doit être expliquée. Ouvrir la donnée, c'est enrichir le débat public. Cette transparence est le meilleur remède à la défiance et au complotisme.

La crise a aussi montré que gouverner par la donnée nécessite de plus en plus souvent de recourir à des données détenues par des acteurs privés, sans lesquelles prendre le pouls du pays ne serait pas possible : les données de Google sur la fréquentation des lieux, celles d'Orange sur les déplacements à travers le territoire, ou celles du Crédit Mutuel sur l'utilisation des comptes bancaires.

Enfin, la donnée est un moyen d'évaluer correctement nos politiques publiques. D'une part, faire de la « data », ce n'est jamais que fiabiliser et permettre en temps réel le bon vieux contrôle de gestion ; et c'est aussi se donner enfin les moyens de suivre l'exécution des dépenses publiques. Le baromètre des résultats de l'action publique en est une traduction immédiate. D'autre part, le partage sécurisé des données des politiques publiques entre administrations et avec les chercheurs est la condition indispensable à une évaluation fine des politiques publiques. La statistique publique n'est plus un simple outil de comptabilité nationale, elle est un réseau de services statistiques qui ont toute la compétence pour évaluer rigoureusement les actions des ministères où ils sont implantés.

La donnée est très loin d'avoir produit tous ses effets. Il est faux de croire qu'un bilan décevant peut être tiré des retombées de l'ouverture des données initiée en 2016, et qu'il faudrait fermer le ban. Cette ouverture est largement incomplète, voire insatisfaisante à certains égards dans la manière dont elle a été mise en œuvre, et les acteurs publics ont aujourd'hui une faible connaissance des réutilisations permises. Encore une fois, c'est la méthode qu'il faut changer, pas l'objectif.

Ni fanatiques, ni réfractaires de l'ouverture

Il faut un équilibre entre ouverture et protection. Cet équilibre semble avoir été trouvé au niveau européen dans le règlement général de la protection des données (RGPD) pour ce qui concerne les données personnelles, mais cet équilibre n'est pas toujours atteint dans le droit et la pratique français, qui n'utilisent pas toutes les marges de manœuvre prévues par le règlement.

Le régime européen de protection des données personnelles doit être effectif en pratique, et son interprétation excessive ne devrait pas être systématisée.

La sécurité ne doit pas masquer une mauvaise foi. Certains acteurs publics prennent prétexte de dispositions de sécurité, qu'ils interprètent dans leur seul intérêt, pour ne pas ouvrir. La mission propose ainsi d'associer plus étroitement l'ANSSI à la politique d'ouverture, pour apporter l'expertise dans le domaine de la sécurité des systèmes d'information, et rappeler que l'ouverture du code est une meilleure protection que sa fermeture.

Certaines fermetures sont « politiques », par crainte d'une remise en cause ou d'un mauvais usage : ainsi les modèles de prévisions de l'Institut Pasteur sur la Covid19, ou certaines statistiques de la délinquance, que la mission propose d'ouvrir. La donnée ne porte pas un propos politique ; seule sa réutilisation le fait. Là où il y a peu d'ouverture, l'attention se cristallise sur le moindre chiffre disponible ; tandis que l'ouverture favorise un débat précis et moins houleux. Les administrations doivent être en mesure de faire entendre leur voix dans ce débat, en apportant leur expertise dans la manière d'appréhender la donnée, sans empêcher que d'autres acteurs puissent s'en saisir et contribuer, eux aussi, au débat.

L'ouverture doit être plus large. Le principe d'ouverture par défaut, édicté en 2016, donne à l'administration l'initiative de l'ouverture, et non plus au seul citoyen, comme prévu par la loi

de 1978. En pratique, beaucoup d'administrations ne jouent pas le jeu. Il est donc nécessaire de rendre ce droit plus effectif, notamment en renforçant les pouvoirs de la CADA. Dans 80 % des cas[1], l'administration ne répond même pas aux demandes, et le délai d'attente d'une réponse de la CADA atteignait en moyenne 176 jours en 2019.

Il faut enfin lever les barrières injustifiées à l'ouverture de données et de codes déjà financés par l'argent public, en examinant les redevances encore pratiquées par les administrations et les régimes de propriété intellectuelle des agents publics, dont la paye seule devrait rétribuer le travail. Dans le cas des données d'acteurs privés chargés d'une mission de service public, l'effectivité de la mise à disposition des données du service public doit être garantie, en anticipant l'impact économique de cette ouverture pour les producteurs (greffiers des tribunaux de commerce par exemple).

La qualité et l'accessibilité doivent être améliorées

L'ouverture doit gagner en maturité, sans renoncer aux principes fondamentaux, notamment l'ouverture par défaut et la libre réutilisation. Il n'y a pas lieu d'établir des critères de sélection des jeux de données à ouvrir en priorité, car il n'est pas possible de connaître par avance les réutilisations qui seront faites de données et de codes, et par conséquent de préjuger totalement de l'utilité d'une ouverture. L'*open data*, c'est aussi être à l'écoute des besoins de la société civile – sans attendre qu'ils se manifestent par le biais d'un contentieux. De la même façon, les licences limitant la réutilisation ne sont donc pas souhaitables.

Cependant, l'*open data* doit changer d'ère et viser une plus grande qualité et fiabilité de la donnée : par la documentation, souvent trop pauvre, par la définition de standards interopérables, par des métadonnées plus homogènes, entre autres. Le service public de la donnée doit ainsi être étendu à de nouveaux jeux de données de référence pour en élever la qualité et la disponibilité. L'enjeu de la qualité est crucial pour le développement de l'intelligence artificielle, qui peut aussi bien se nourrir de données publiques que de données sensibles, mais a dans tous les cas besoin d'une donnée abondante et de qualité.

La qualité adviendra par l'écoute des réutilisateurs. À date, la relation entre le producteur et le réutilisateur n'existe pas, le plus souvent. Ainsi, la direction productrice de la base la plus réutilisée de *data.gouv.fr* (DVF) ne participe pas régulièrement aux travaux de la communauté de réutilisateurs, qui pourraient pourtant lui apporter un indice majeur sur la qualité du service public à l'origine de la donnée. Une exception notable est la relation qu'ont souvent su instaurer les services statistiques ministériels, qui pratiquent ce « retour utilisateur » et alimentent le travail des chercheurs, mais ne sont pas eux-mêmes les services métiers.

Au-delà de la qualité intrinsèque de la donnée, la qualité de sa diffusion garantit que la donnée puisse circuler : par l'usage naturel des labels de qualité dans les services producteurs, son appropriation par tout autre utilisateur est facilitée. Les infrastructures doivent répondre à ce besoin, par un cadre sécurisé par l'interopérabilité et des services de diffusion adaptés. Leur gouvernance doit embarquer les réutilisateurs. Les investissements de l'État, à commencer par les actions financées par le plan de relance, doivent prendre en compte la circulation des données.

L'offre d'*open data* doit aussi gagner en accessibilité et en visibilité. La donnée doit être exposée au travers de catalogues visibles et fédérateurs pour être enfin identifiée au plus près de sa production. Le service de *data.gouv.fr* doit être repensé, pour améliorer l'exploration de l'offre et permettre une plus grande accessibilité des données. Le recours en pratique nécessaire à une API pour accéder à des bases, comme pour la base SIRENE gérée par l'INSEE, peut créer dans certains cas un frein pour les usages et doit donc être proportionné, même s'il permet de suivre finement et de mieux analyser les réutilisations et, dans une certaine mesure, de les encadrer.

[1] Échantillon de 98 demandes adressées par l'association L'Ouvre-Boîte réalisées entre 2017 et 2020.

S'agissant de l'ouverture des codes et de l'utilisation de logiciels libres, il faut structurer la communauté du secteur public et renforcer l'appui qui lui est apporté. Le logiciel libre n'est pas une idéologie déconnectée des besoins des administrations et ses enjeux ne se résument pas à la question de l'utilisation de LibreOffice. Il est au contraire le moyen de créer enfin du partage et de la mutualisation dans le secteur public, d'éviter que deux administrations s'épuisent sur un même problème sans le savoir et sans se parler, enfin de permettre à l'administration et à l'économie de s'enrichir mutuellement en développant ensemble des outils d'intérêt général. Il constitue aussi une réponse au manque d'attractivité de l'État pour les compétences numériques. La mission considère que la création d'un *Open Source Program Office* (OSPO), visible et pérenne, au sein de la DINUM, serait une première pierre pour relever ce défi.

Le partage entre acteurs publics doit être un impératif d'efficacité de l'action publique

Certaines données ne peuvent être ouvertes à tous. Là commence le domaine du partage et de l'accès, ces deux notions permettant de distinguer le cas où l'utilisateur possède une copie physique des données sur son serveur (partage) et le cas où il ne peut l'exploiter que par un accès au serveur du producteur des données, sans en garder une copie physique (accès).

Il est très regrettable que de nombreux acteurs réduisent la notion d'ouverture à celle d'*open data* et ne considèrent pas même l'éventualité de partager de manière limitée et sécurisée certaines de leurs données. La difficulté naît le plus souvent d'un manque de confiance à l'égard de la réutilisation, et trouvera donc sa réponse dans le portage politique et à haut niveau administratif de cette démarche.

Le partage de données entre administrations de l'État est scandaleusement faible, au point que certaines directions ressaisissent des données disponibles dans une direction du même ministère, ou que l'*open data* est parfois le seul moyen pour une administration de connaître l'existence puis d'accéder aux données d'une autre administration – ce qui plaide encore pour l'encouragement de cette ouverture intégrale, quand elle est possible.

Et lorsque le partage est acté, la procédure est parfois trop contraignante : c'est le cas lorsque des administrations souhaitent croiser deux bases de données et utiliser le numéro d'inscription au répertoire (NIR), procédure sécurisée assouplie par la loi pour une République numérique de 2016, mais qui n'est toujours pas opérationnelle et mise en œuvre quatre ans après. Il est ainsi aujourd'hui impossible de connaître la situation d'emploi des nombreux nouveaux bénéficiaires du RSA enregistrés cette année.

Le partage d'informations est également limité entre État et collectivités, en dépit de certaines coopérations sur des plateformes de données régionales par exemple. Mais les collectivités ne donnent généralement aucun accès aux données d'exécution des dispositifs nationaux dont elles ont la gestion, par exemple, y compris lorsqu'ils sont intégralement financés par l'État, comme dans le cas du développement économique. Par ailleurs, dans le cadre de la gestion de crise de la Covid19, l'enrichissement de l'information des collectivités territoriales sur l'évolution des données épidémiologiques concernant leur territoire constitue une demande forte, pour mieux comprendre par exemple l'application des mesures de confinement ou de couvre-feu locales.

L'accès sécurisé aux données sensibles doit renforcer notre indépendance en matière d'intelligence artificielle

L'accès sécurisé est une modalité qui permet d'analyser des données sans que ces dernières ne sortent du serveur propriétaire, soit une garantie maximale de protection des données sensibles et notamment personnelles. L'accès sécurisé permet d'exploiter tout le capital de la donnée et constitue un enjeu d'autonomie stratégique pour la recherche. Cet usage est en effet particulièrement utile pour les chercheurs, qui ont besoin d'exploiter des données qui ne soient pas anonymisées mais nominatives ou pseudonymisées (niveau moindre d'anonymisation, empêchant la ré-identification sans avoir recours à des informations supplémentaires).

Cette modalité se développe, grâce à l'émergence de plateformes et de gouvernances sectorielles ou intersectorielles, puissant vecteur de décloisonnement de la donnée, tel que le *Health data hub* dans la santé, le *Ag-data hub* en matière agricole, ou encore le centre d'accès sécurisé aux données (CASD), conçu d'abord comme un démembrement de l'INSEE à destination des chercheurs. Ce développement par secteur est souhaitable mais ne doit pas conduire à une situation de « silos » qui ne pourraient absolument plus être interconnectés à l'avenir.

Par ailleurs, les besoins en jeux d'apprentissage pour l'entraînement d'algorithmes d'intelligence artificielle ne sont pas satisfaits aujourd'hui en France, ce qui contraint les start-up françaises à aller chercher ailleurs les ressources pour développer des outils et des services que nous utilisons déjà dans notre quotidien. **Nous devons nous donner les moyens de garantir notre autonomie stratégique dans la technologie d'intelligence artificielle, au risque de voir bientôt nos vies dictées par des algorithmes formés à partir de données d'apprentissage qui ne reflètent pas nos valeurs et nos choix de société.**

Enfin, l'accès aux données est encore trop restreint pour les chercheurs. De nets progrès ont été réalisés, mais le système français est loin des standards internationaux. Il est nécessaire d'améliorer la prise en charge des demandes des chercheurs, associant les administrateurs des données et les services statistiques ministériels, car la recherche est un vecteur essentiel de l'évaluation de l'action publique. Une chercheuse, française, souhaitant évaluer le travail détaché en France, n'a ainsi à ce jour pas reçu les données de l'administration malgré une demande effectuée il y a plus de deux ans, et l'accord du comité du secret statistique. Elle n'a en revanche eu aucune difficulté à obtenir ces mêmes données de la part de la Belgique, du Luxembourg et du Portugal.

L'utilisation à grande échelle de données du secteur privé est devenue incontournable pour la puissance publique

Le concept de « donnée d'intérêt général » vise à définir le cas où la mise à disposition d'une donnée d'un acteur privé peut être justifiée par un « motif d'intérêt général ». En l'absence de définition de la notion, il convient de distinguer deux cas de figure, qui appellent des interventions différentes de la puissance publique : d'une part, l'utilisation par les administrations de données produites par le secteur privé (B2G), dont la crise a donné un exemple récent avec l'utilisation de données d'Orange et des opérateurs de carte bancaire pour suivre l'activité du pays pendant le confinement ; d'autre part, des initiatives de partage de données entre acteurs privés, par exemple au sein d'une filière (B2B). Les données d'opérateurs privés mises en *open data* parce qu'ils opèrent un service public ne sont pas rigoureusement parlant des données d'intérêt général mais une ouverture de données publiques. C'est le cas de la diffusion des données des services de transport public (horaires, prix, tarifs et itinéraires).

Cette réflexion sur les données accessibles à la puissance publique révèle une spécificité de l'histoire française, celle d'un État ayant garanti l'autonomie de sa prise de décision non seulement par des producteurs d'information nationaux, comme l'IGN, qui remonte aux besoins militaires du XVII^{ème} siècle, mais en leur donnant parfois un statut supérieur aux producteurs privés : l'INSEE a ainsi été créé en 1946 comme « monopole » de production de l'information économique, particularité française par rapport à l'Allemagne ou aux États-Unis, et reflet, à l'époque, de sa conception politique de la donnée, sans pluralisme[2].

Les acteurs privés transmettent déjà de manière obligatoire de nombreuses informations à la puissance publique, ou entre acteurs privés, dans le cadre de politiques de régulation notamment. La nouveauté dans la notion de donnée d'intérêt général est de considérer une pratique d'exploitation à grande échelle de jeux de données massives, et notamment de *Big Data* et de l'élargir à d'autres finalités que la production d'enquêtes statistiques.

[2] Pierre Rosanvallon, *L'État en France de 1789 à nos jours*, 1993.

Ce besoin pose moins une question de légitimité d'intervention de l'État, qu'un problème de sécurité juridique. La mise à disposition de ces données ne peut s'envisager que dans un cadre respectueux de la liberté d'entreprendre et du droit de propriété qui pourrait s'attacher aux données, et en garantissant la transparence de leur réutilisation par l'État. Une clarification des cadres juridiques du B2G et du B2B est nécessaire pour rassurer les acteurs privés sur cette démarche.

Il est désormais nécessaire de passer à l'échelle en matière de données d'intérêt général, notamment dans le cadre des travaux européens (*Data Governance Act, Data Act, Digital Services Act*), qui constituent une fenêtre d'opportunité à ne pas manquer. La mission considère que ce travail d'identification des secteurs où ce partage est pertinent, partiellement réalisé par une mission en 2016, doit devenir la mission d'un organe pérenne, et entrer dans un mode de gestion courant. Il est ainsi nécessaire de se doter d'un mécanisme transversal pouvant jouer le rôle de révélateur des besoins d'un plus grand partage de données privées. Si la CADA a permis de mettre l'administration à l'écoute des besoins de la société civile, il manque aujourd'hui le moyen de recueillir les besoins de données privées par les administrations.

Plutôt qu'un cadre juridique transversal et unique, qui n'est pas envisageable à court terme, faute de maturité suffisante, il convient donc de créer une gestion en « mode projet » de ces besoins, en associant les acteurs concernés pour définir les modalités et la gouvernance de ce partage, sous l'égide de la puissance publique. Les principes de transparence et de redevabilité doivent être placés au cœur de la démarche. La question de l'utilisation des données par la puissance publique ne peut pas uniquement se traiter entre la puissance publique d'un côté et le secteur privé de l'autre, *a fortiori* quand les données dont il est question sont à l'origine des données concernant les individus. Il faut veiller à intégrer la société civile et leurs représentants dans la démarche.

Dans la même perspective, les initiatives de portabilité citoyenne des données au service de l'intérêt général doivent gagner en ampleur, pour permettre un meilleur contrôle et mise à disposition choisie par les citoyens de leurs données personnelles, y compris au service de l'intérêt général.

Enfin, le développement d'infrastructures sécurisées de partage de données, comportant dès le stade de leur conception des outils de gestion des droits et de la réglementation, est une condition indispensable pour renforcer la confiance des acteurs privés dans ces nouveaux modes de collaboration et de création de valeur par la donnée.

Cette politique doit être incarnée et diffusée

Que se passera-t-il après ce rapport ? En l'état actuel des choses, la mission craint que ses recommandations ne soient pas portées et suivies. Elle formule donc plusieurs recommandations pour donner les moyens à cette politique d'être transformée en actes.

Premièrement, un portage politique et administratif est nécessaire : les questions soulevées dans ce rapport doivent l'être plus régulièrement, sans attendre une mission spécifique. Il manque une priorité gouvernementale, un administrateur général de la donnée, des administrateurs ministériels plus visibles et plus soutenus. Il ne s'agit pas d'une démarche incantatoire. La donnée doit être un objet politique et gouvernemental, et à ce titre, portée par le Premier ministre, dans le cadre d'un comité interministériel présidé par lui. Ce portage politique doit être aussi à haut niveau administratif : c'est le rôle de la DINUM, en tant que responsable de la mise en œuvre, et de la DITP, pour le suivi de l'exécution.

Deuxièmement, la donnée est une nouvelle mission des services publics et appellent, à ce titre, des moyens, humains et financiers : la CNIL et la CADA doivent pouvoir répondre à des sollicitations toujours plus nombreuses et complexes, la DINUM doit pouvoir apporter un appui pour répondre aux besoins des directions, les services statistiques doivent pouvoir être disponibles et réactifs. Cette recommandation n'est pas un gouffre budgétaire, dès lors qu'il est bien établi que la politique de la donnée est un vecteur puissant de productivité, quand on lui donne les moyens de se déployer.

Troisièmement, la politique de recrutement des talents du numérique doit être adaptée. La gestion des compétences et les outils RH pour attirer et maintenir les profils spécialisés progressent, mais ne sont pas encore suffisants pour garantir une filière technique de haut niveau dans le domaine de la donnée et du code.

S'agissant du code, qui sait que trois Français figurent parmi la liste très américaine des dix-huit *Debian leaders* mondiaux, rôle éminent dans le monde du logiciel libre ? La France doit accompagner ses talents, comme elle le fait pour des sportifs de haut niveau, et s'appuyer sur eux pour renforcer l'attractivité du secteur public grâce au logiciel libre.

Quatrièmement, la fonction publique a besoin d'une culture de la donnée et du code. Ce besoin fait partie de ceux qui ont été le plus mis en avant par la consultation publique. L'enjeu est d'agir sur le stock, et non seulement par les recrutements : trop de fonctionnaires, hauts placés et aux postes de commandement, éprouvent aujourd'hui de la peur à l'égard de l'ouverture des données, le plus souvent par ignorance. L'ouverture ne pourra progresser en attendant que des générations plus averties et sensibilisées aux enjeux parviennent jusqu'à ces postes. En général, les agents n'ont aujourd'hui que peu d'incitations à se former et à contribuer à la transformation de leur service par la donnée.

La mission a développé dans son rapport plusieurs « cas d'usage », sélectionnés pour leur caractère emblématique de certains des constats et des propositions formulés, et ayant pu faire l'objet d'un contradictoire avec les acteurs concernés suffisant à établir les faits.

La mission a également souhaité mettre en avant les contributions au rapport inspirées de la consultation publique conduite entre le 8 octobre et le 9 novembre 2020, dont la synthèse est présentée en annexe. Sont ainsi signalés les constats et recommandations ayant fait l'objet d'une audience toute particulière dans le cadre de cette consultation.

Recommandations

Recommandations transversales

Recommandation n° 1 : Initier un débat public sur les conditions de la confiance dans le numérique, permettant de définir les principes fondamentaux de sécurité et de transparence qui doivent s'imposer à la puissance publique

Recommandation n° 2 : Associer la société civile, par les consultations citoyennes et le Forum du Partenariat pour un gouvernement ouvert, à l'identification des jeux de données et des codes sources à ouvrir

Recommandation n° 3 : Conduire une évaluation de l'impact économique, social et scientifique de l'ouverture et du partage des données et des codes sources

Portage de la politique

Recommandation n° 4 : Assurer un portage politique au niveau du Premier ministre des enjeux de la donnée et des codes source. Inscrire à l'ordre du jour des comités interministériels présidés par le Premier ministre le suivi et la mise en œuvre de cette politique. Édicter une circulaire établissant les principes (gouvernance, missions et responsables dans les administrations, interopérabilité, qualité, guides juridiques)

Mise en œuvre de l'ouverture des données et des codes sources

Recommandation n° 5 : Nommer un administrateur général de la donnée, des algorithmes et des codes sources (AGDAC), missionné par le Premier ministre, auprès du DINUM, ayant pour mission à temps plein de piloter la stratégie nationale d'ouverture de la donnée et des codes sources, en s'appuyant sur les administrateurs ministériels des données, des algorithmes et des codes source (AMDAC)

Recommandation n° 6 : Structurer le pilotage et le suivi de la politique d'ouverture des données et des codes sources au niveau interministériel (indicateurs de performance, insertion dans les études d'impact des projets de loi)

Recommandation n° 7 : Engager la puissance publique sur la voie d'une participation plus active aux communs numériques

Recommandation n° 8 : Créer un « Open Source Program Office » (OSPO) ou une mission logiciels libres au sein de TECH.GOUV, chargée d'aider l'administration à ouvrir et à réutiliser les codes sources publics, d'identifier les enjeux de mutualisation et de créer des liens avec les communautés open source exsitantes et d'accompagner les talents français dans ce domaine

Recommandation n° 9 : Élargir et renforcer la fonction d'administrateur ministériel des données, des algorithmes et des codes sources (AMDAC) :

- en redéfinissant leurs missions dans une fiche de poste type
- en dotant les AMDAC d'une lettre de mission signée par les ministres concernés après consultation des directions générales et de la DINUM
- en s'assurant que l'AMDAC a des moyens d'intervention suffisants
- en systématisant des formations conjointes entre AMDAC et délégués à la protection des données

Recommandation n° 10 : Confier à l'ANCT une mission d'accompagnement des collectivités territoriales dans la publication des données et des codes sources *via* des programmes cofinancés entre État et régions

Recommandation n° 11 : Prendre davantage en compte les démarches d'open source et d'open data pour le rayonnement de la recherche française dans les évaluations et le financement des projets

Droit et régulation

Recommandation n° 12 : Faire évoluer le droit d'accès aux documents administratifs pour renforcer l'effectivité de la loi en confiant un pouvoir de sanction à la CADA en cas de non-respect des dispositions du CRPA relatives à la communication et à la publication des données et documents et pour alléger l'activité de la CADA sur les saisines simples, et pour fluidifier la gestion des dossiers récurrents devant la CADA

Recommandation n° 13 : Évaluer les besoins en ressources humaines de la CNIL pour renforcer son rôle de conseil et d'accompagnement et assortir l'augmentation des moyens correspondant d'un suivi au travers d'indicateurs de performance sur la satisfaction des usagers (dans le cadre du PLF)

Recommandation n° 14 : Prévoir dans les collèges de la CNIL et de la CADA deux personnalités qualifiées compétentes, l'une en matière de sécurité des systèmes d'information et l'autre sur les nouveaux usages de la donnée

Recommandation n° 15 : Associer l'ANSSI à la mise en œuvre de la politique d'ouverture des données et des codes sources afin d'assurer que cette politique n'entre pas en contradiction avec les impératifs de sécurité des systèmes d'information :

- prévoir que la CADA et la CNIL puissent saisir l'ANSSI pour avis quand il y a un doute sérieux en matière de sécurité des systèmes d'information ;
- prévoir la possibilité, pour l'AGDAC de solliciter l'ANSSI pour un audit de bibliothèques et de logiciels libres sensibles

Recommandation n° 16 : Vérifier que la loi garantit l'ouverture de toutes les données de services publics mis en œuvre par des acteurs privés (professions réglementées de la justice notamment)

Acculturation et politique RH

Recommandation n° 17 : Développer une politique de formation de la fonction publique plus ambitieuse sur les enjeux du numérique (obligation de formation des cadres dirigeants aux enjeux du numérique, séminaires de cadres dirigeants, offre de formation pour tous les niveaux hiérarchiques, plans de formation ministériels complémentaires à l'offre interministérielle, modules dans l'ensemble des cursus de formation de la fonction publique)

Recommandation n° 18 : Poursuivre les travaux relatifs à la gestion des emplois et des compétences du numérique et structurer dans la formation initiale une filière technique de la fonction publique pour les métiers experts du numérique, en créant des parcours pour les corps techniques et en pérennisant en CDI les agents contractuels apportant des compétences non disponibles dans les corps existants

Recommandation n° 19 : Diversifier les parcours des administrateurs et des attachés de l'INSEE dans l'ensemble des administrations, au-delà des services statistiques ministériels, et valoriser le travail et la carrière des agents choisissant ces parcours

Recommandation n° 20 : Accroître l'attractivité de l'État pour les métiers du numérique en tension (rendre le référentiel de rémunération obligatoire, développer la communication auprès des formations spécialisées)

Recommandation n° 21 : Passer à l'échelle et inscrire dans la durée le programme d'entrepreneurs d'intérêt général

Recommandation n° 22 : Proposer une offre de formation dédiée aux élus sur les enjeux de la donnée et des codes sources dans les politiques publiques

Qualité de la donnée

Recommandation n° 23 : Créer un label de service producteur de la donnée pour reconnaître les efforts investis dans la donnée, par exemple dans le cadre du service public de la donnée

Recommandation n° 24 : Définir et mettre en œuvre une politique interministérielle d'interopérabilité et de qualité de la donnée (démarches de standardisation, label FAIR, doctrine sur les métadonnées, catalogage)

Recommandation n° 25 : Encourager les écosystèmes à définir des principes de gouvernance de la qualité, en désignant un référent qualité et en créant des communautés de réutilisation avec participation active des producteurs de la donnée

Infrastructures, partage et accès sécurisé

Recommandation n° 26 : Orienter les investissements du plan de relance vers les infrastructures favorables à la circulation de la donnée (appels à projets de la DINUM et appels à projets sectoriels)

Recommandation n° 27 : Encourager la création de « hubs » sectoriels ou intersectoriels, selon des modalités adaptées à chaque secteur, et dans des conditions assurant leur interopérabilité

Recommandation n° 28 : Créer un dispositif de bac à sable expérimental permettant à la CNIL de déroger aux textes existants pour autoriser la réutilisation de données personnelles dans des jeux d'apprentissage d'algorithmes d'intelligence artificielle, et leur conservation pour une durée plus longue que celle autorisée lors de leur collecte initiale

Recommandation n° 29 : Mettre en œuvre les dispositifs techniques permettant d'utiliser la procédure d'appariement de fichiers sur la base du code statistique non signifiant à des fins de statistique publique et de recherche scientifique et historique

Recommandation n° 30 : Améliorer la prise en charge des demandes des chercheurs, en associant les AMDAC et les SSM (délai de réponse obligatoire, création d'un recours, recours à la consultation du comité du secret statistique à titre facultatif)

Données d'intérêt général

Recommandation n° 31 : Privilégier une approche incitative et concertée, le recours à d'éventuels dispositifs coercitifs devant être dûment justifié et faire l'objet d'une évaluation préalable

Recommandation n° 32 : Sécuriser le cadre juridique du partage volontaire de données d'intérêt général concernant l'utilisation des données à caractère personnel (par un guide de conformité de la CNIL) et l'application du droit d'accès et de réutilisation applicable aux données du secteur privé reçues par les administrations

Recommandation n° 33 : Encourager les initiatives de portabilité citoyenne des données au service de l'intérêt général, notamment par l'organisation de campagnes de mobilisation citoyenne

Utilisation par le secteur public de données issues du secteur privé (B2G)

Recommandation n° 34 : Clarifier le régime juridique de la réquisition pour permettre à la puissance publique d'accéder à des données du secteur privé en cas de motif impérieux d'intérêt général et d'urgence

Recommandation n° 35 : Confier au réseau de l'AGDAC et des AMDAC une mission de facilitation et de médiation de l'accès et de l'utilisation des données du secteur privé par le secteur public (B2G), en lien avec la direction générale des entreprises (DGE)

Recommandation n° 36 : Garantir l'effectivité des dispositions relatives aux données d'intérêt général de la loi pour une République numérique qui rencontrent des difficultés d'application :

- en matière de données détenues par les concessionnaires et délégataires du service public (clausiers types pour les acteurs publics)
- en matière d'utilisation des données privées à des fins statistiques, étudier l'opportunité d'élargir l'article 19 à certains services fondés sur les données

Partage de données entre acteurs privés (B2B)

Recommandation n° 37 : Développer le partage de données privées au service d'intérêts partagés (B2B) au sein des comités stratégiques de filières, dans les appels à projets publics (PIA), et en soutenant les initiatives associatives et privées

Introduction

Par une lettre du 22 juin 2020, le Premier ministre a confié à Éric Bothorel, député des Côtes-d'Armor, une mission relative à la politique publique de la donnée, associant Stéphanie Combes, directrice générale du Health Data Hub, et Renaud Vedel, coordonnateur national pour l'intelligence artificielle.

Entre juillet 2020 et décembre 2020, la mission a conduit plus de 200 auditions avec les acteurs concernés (cf. liste en annexe), recueilli par écrit les observations de plus de 50 administrations et opérateurs publics, sollicité l'expertise des services économiques de la direction générale du Trésor dans 9 pays, et celle des services juridiques de la direction interministérielle du numérique (DINUM) et de la direction des affaires juridiques des ministères économiques et financiers, entre autres (les contributions juridiques et l'enquête de comparaison internationale sont rassemblées dans un document annexé au rapport).

En outre, entre le 8 octobre et le 9 novembre 2020, la mission a conduit une consultation publique, sur un site Internet accessible à tous[3], pour recueillir, de la part de citoyens et d'organisations, tout à la fois des commentaires sur son rapport d'étape, publié le même jour que le début de la consultation, des propositions d'action pour répondre aux problèmes identifiés, ainsi que des réactions à plusieurs situations d'utilisation de données d'acteurs privés. Dans le cadre de cette consultation, 545 comptes utilisateurs ont été enregistrés, 108 contributions libres ont été formulées, 418 commentaires ont été publiés, 1 753 soutiens aux contributions ont été recueillies, et 5 954 visites ont été enregistrées au total sur le site.

Ainsi, à la demande du Premier ministre, le présent rapport :

- dresse un état des lieux des enjeux dans la politique publique de la donnée (partie 1) ;
- identifie les limites actuelles ainsi que les progrès accomplis et formule des recommandations en matière d'ouverture des données et de codes sources publics (partie 2) ;
- dégage les enjeux et les besoins en matière de qualité de la donnée, et de partage et d'accès sécurisé pour les besoins qui ne sont pas satisfaits par l'*open data* (partie 3) ;
- précise les moyens humains et techniques nécessaires à la mise en œuvre de la politique publique de la donnée (partie 4) ;
- propose un état des lieux et formule une doctrine en matière de données d'intérêt général, notamment de partage de données du secteur privé avec la puissance publique et de partage de données entre acteurs privés (partie 5).

Le rapport a été finalisé le 11 décembre 2020, avant la présentation du *Digital Services Act* par la Commission européenne, dont il n'a donc été tenu compte qu'en possession des éléments publics à cette date.

[3] Sur le site mission-open-data.fr

Partie 1
Une politique au service de toutes les autres

1. Les données et les codes sources : de quoi parle-t-on ?

1.1. Données, bases de données et métadonnées

On trouve une définition juridique de **la donnée** dans un arrêté du 22 décembre 1981, comme « *représentation d'une information sous une forme conventionnelle destinée à faciliter son traitement* »[4]. Une **base de données**, quant à elle, est définie par le droit européen comme un « *recueil d'œuvres, de données ou d'autres éléments indépendants, disposés de manière systématique ou méthodique et individuellement accessibles par des moyens électroniques ou d'une autre manière* »[5]. **La donnée, et *a fortiori* l'*open data*, ne se limite donc pas au champ des documents administratifs** défini par la loi CADA de 1978.

Les données sont habituellement traitées afin d'en extraire de l'information au sein d'une base de données ; on parle alors de données « structurées ». Toutefois, l'évolution des techniques d'analyse de données a permis de traiter un éventail plus large des données telles les données non structurées (par exemple les données générées par le passage sur une page internet).

À toute donnée comme à toute base de données peuvent être associées des **métadonnées**, c'est-à-dire des données qui peuvent indiquer la date, le lieu, la méthode de collecte ou toute autre information relative aux données elles-mêmes.

Plusieurs organisations productrices de données font la distinction entre **données « brutes » et données « travaillées »** (correspondant par exemple à des statistiques calculées ou estimées à partir des bases de données). Si d'un strict point de vue juridique il n'y a pas lieu d'opérer une telle distinction, elle peut fonder des stratégies d'ouverture différentes.

1.2. Les codes sources, à distinguer des algorithmes

Un code source peut être défini comme un ensemble d'instructions exécutables par un ordinateur. Un code source se différencie d'un algorithme, qui est défini par la CNIL comme « *la description d'une suite finie et non ambiguë d'étapes (ou d'instructions) permettant d'obtenir un résultat à partir d'éléments fournis en entrée* »[6], par le fait que l'algorithme n'est pas nécessairement informatisé. De manière simplifiée, l'algorithme est une recette de cuisine, et le code sa réalisation concrète : ouvrir l'algorithme est un premier acte de transparence, mais rien ne garantit que la recette a été réellement suivie. C'est en cela que l'ouverture du code est un gage de transparence supplémentaire, en permettant de voir comment le processus est réellement mis en œuvre. L'application ou le programme, enfin, est le produit de cette recette, prêt à être utilisé.

Un « algorithme public » est une procédure administrative dont tout ou partie est informatisée et qui intervient dans un processus de décision pour les citoyens. Un algorithme de répartition des places en crèches avec l'intervention humaine d'un comité d'attribution, combinée au tri automatisé d'une machine qui permet de décider qui en bénéficie, constitue un exemple.

[4] Vocabulaire des technologies de l'information et de la communication (TIC), Délégation générale à la langue française et aux langues de France, 2017.

[5] Directive 96/9/CE du Parlement européen et du Conseil du 11 mars 1996 concernant la protection juridique des bases de données.

[6] Rapport de la CNIL de décembre 2017, *Comment permettre à l'homme de garder la main ?*

Certains codes sources participent à la mise en œuvre « d'algorithmes publics », mais cela ne concerne pas la majorité des cas. À titre d'exemple, le code de la plateforme *code.etalab.gouv.fr* est publié sans que cela ne fonde un processus de décision administrative. La loi pour une République numérique a par ailleurs introduit des obligations en matière de transparence et d'explicabilité des décisions administratives individuelles prises sur le fondement d'un traitement algorithmique, c'est-à-dire d'un algorithme traduit par un langage de programmation informatique.

Par ailleurs, ouvrir le code source de la part automatisée d'un algorithme public ne suffit pas à fournir une explication aux citoyens concernant la mise en œuvre des algorithmes. La Bonne Boîte, algorithme développé par Pôle Emploi pour aider les demandeurs d'emploi dans leurs démarches, constitue une illustration. En effet, en plus de publier son code source, celle-ci a publié une vidéo explicitant l'algorithme utilisé[7].

1.3. Données et codes sources publics, documents administratifs et information publique

D'après un groupe de travail de l'Observatoire juridique des technologies de l'information[8], une donnée est dite « publique » si elle est créée sur un fonds public. Cette définition matérielle peut s'étendre pour qualifier les codes sources publics : sont publics les codes sources créés sur un fonds public.

Les codes sources et données publics sont considérés comme des documents administratifs et des informations publiques. En effet, l'article L.300-2 du code des relations entre le public et l'administration (CRPA) précise la définition des documents administratifs : « *Sont considérés comme documents administratifs, au sens des titres Ier, III et IV du présent livre, quels que soient leur date, leur lieu de conservation, leur forme et leur support, les documents produits ou reçus, dans le cadre de leur mission de service public, par l'État, les collectivités territoriales ainsi que par les autres personnes de droit public ou les personnes de droit privé chargées d'une telle mission. Constituent de tels documents notamment les dossiers, rapports, études, comptes rendus, procès-verbaux, statistiques, instructions, circulaires, notes et réponses ministérielles, correspondances, avis, prévisions, codes sources et décisions* ». Une donnée publique peut donc être produite par une entité privée, le critère étant la mission de service public[9]. Le droit français considère ainsi, comme le permet le droit européen[10], qu'un code source est un document administratif.

Le terme d'information publique est défini de manière indirecte dans l'article L.321-1 du CRPA comme *« les informations publiques figurant dans des documents communiqués ou publiés par les administrations* ». Ainsi, une information correspond au contenu du document administratif.

Les données et codes sources publics sont considérés à la fois comme un document administratif et comme une information publique. L'avis n° 20144578[11] de la Commission d'accès aux documents administratifs (CADA) affirme ainsi, s'agissant en l'espèce du code source du logiciel simulant le calcul de l'impôt sur les revenus des personnes physiques, qu'un code source est un document administratif, et doit donc être ouvert suivant des modalités décrites ci-dessous, et qu'en tant qu'information publique, il doit aussi être librement réutilisable.

---

[7] Disponible sur la plateforme Dailymotion, « Expliquer l'algorithme de La Bonne Boîte en 1'30 ».

[8] *Commercialisation des données publiques*, P.Gaudrat, Observatoire juridique des technologies de l'information, 1992.

[9] En étaient alors exclues « les données produites dans le cadre d'une mission de service public à caractère industriel et commercial. Cela concerne non seulement les établissements publics à caractère industriel et commercial mais pour la part de leur activité effectuée selon les règles du commerce. »

[10] Directive (UE) 2019/1024 du Parlement européen et du Conseil du 20 juin 2019 concernant les données ouvertes et la réutilisation des informations du secteur public, considérant (30) : « *la définition du terme « document » n'est pas destinée à couvrir les programmes informatiques. Les États membres peuvent étendre l'application de la présente directive aux programmes informatiques* ».

[11] Avis 20144578 du 09/01/2015 de la CADA. Communication du code source du logiciel simulant le calcul de l'impôt sur les revenus des personnes physiques.

1.4. Les données d'intérêt général, une notion floue qui recouvre des situations très différentes

Parmi les données publiques, certaines peuvent être produites par des organismes de droit privé, lorsqu'ils sont en charge d'un service public par délégation ou lorsque l'État est actionnaire. Le caractère privé du producteur ne fait donc pas obstacle, dans certains cas, à l'ouverture de ses données. En revanche, lorsque la puissance publique souhaite accéder, dans un but d'intérêt général, à des données produites ou collectées par la sphère privée indépendamment d'une mission de service public, la question est plus difficile à instruire, du fait de l'absence de critère consensuel et stabilisé pour encadrer cette pratique. La partie 5 du présent rapport traite spécifiquement des données d'intérêt général et propose des repères pour clarifier la notion.

2. L'ouverture des données et des codes sources : pour quoi faire ?

La politique d'ouverture des données et des codes sources répond à de multiples enjeux à la fois pour les citoyens, les entreprises, les acteurs publics et les chercheurs.

La crise de la Covid19 a mis en évidence l'importance de la politique publique de la donnée, à la fois dans un souci de transparence vis-vis des citoyens, pour le pilotage des politiques publiques, mais aussi pour mieux informer les entreprises des dispositifs leur étant destinés ou encore pour contribuer à l'amélioration des connaissances au travers de la recherche. Au plus fort de la crise sanitaire, la fabrique numérique des ministères sociaux a par exemple mis en place une plateforme destinée aux professionnels de santé et aux acteurs de la gestion de la crise pour accéder à une vision globale des services et ressources disponibles pour combattre l'épidémie. Ces développements, réalisés avec la direction de la recherche, des études, de l'évaluation et des statistiques (DREES), l'agence nationale de la sécurité des systèmes d'information (ANSSI) et le ministère des armées, ont permis par exemple de mettre à disposition une plateforme donnant l'accès aux tests de la Covid19 disponibles pour les personnes autorisées. Elle comprend également un outil développé par le ministère des armées recensant la recherche scientifique relative à la Covid19, ainsi qu'un module sur les projets innovants.

La crise a aussi démontré que l'urgence politique peut accélérer la collaboration entre les services de l'État en confirmant la nécessité de casser les silos et en partageant les informations entre les services. Les travaux de ciblage des entreprises réalisés dans le cadre du contrôle de l'activité partielle a ainsi été permis via le croisement des données de la délégation générale à l'emploi et à la formation professionnelle (DGEFP), de la direction générale du travail (DGT) mais aussi de la déclaration sociale nominative (DSN), de l'institut national de la statistique et des études économiques (INSEE) et de l'agence centrale des organismes de sécurité sociale (ACOSS) par exemple.

2.1. Pour les citoyens, un enjeu de transparence de la vie publique et de simplification des démarches

La politique d'ouverture des données et des codes sources publics vise en premier lieu à accroître la transparence de la vie publique vis-à-vis du citoyen. Cet objectif, qui vient renforcer la vie démocratique du pays, s'inscrit directement dans l'application de l'article 15 de la Déclaration des droits de l'Homme et du Citoyen : *« La société a le droit de demander compte à tout agent public de son administration »*. La transparence de la vie publique, qui passe notamment par l'ouverture des données publiques, figure également dans les principes de la déclaration du Partenariat pour un gouvernement ouvert, de septembre 2011, auquel la France adhère (cf. partie 2, titre 2). Du reste, ce principe ressort comme la motivation principale dans l'ensemble des pays interrogés dans l'enquête menée par la mission[12].

[12] La mission a sollicité l'expertise des services économiques de la direction générale du Trésor pour neuf pays : le Canada, la Corée du Sud, les États Unis, l'Estonie, l'Irlande, Israël, le Japon, le Royaume-Uni et Taïwan. Leurs réponses complètes figurent dans les annexes au rapport.

Les données sont le vecteur d'une information publique de meilleure qualité

La donnée favorise la transparence de l'action publique, d'abord au niveau de la politique nationale conduite par le gouvernement. Le plus récent et le plus éloquent est certainement le cas des données sur l'épidémie de la Covid19. L'ouverture de ces données répond à cet objectif de transparence sur les informations fondant la prise de décisions par le gouvernement ; il s'agit d'un élément essentiel pour susciter la confiance des citoyens et l'adhésion aux décisions politiques. Ces données, d'abord rassemblées au travers d'une initiative citoyenne début mars 2020, font dorénavant l'objet d'un tableau de bord de suivi de l'épidémie publié quotidiennement sur le site du gouvernement[13] (cf. cas d'usage sur les données et modèles épidémiologiques).

La donnée produit déjà des effets importants en matière environnementale. Les informations relatives aux ventes de produits phytosanitaires sont rendues disponibles à travers un outil de visualisation : fruit d'une collaboration entre l'Office français de la biodiversité (OFB) et le service de la donnée et des études statistiques (SDES) du ministère de la Transition écologique et solidaire, il est possible de suivre la tendance d'évolution des ventes et de dresser un bilan sur les données les plus récentes. Les données sont issues de la banque nationale des ventes de produits phytopharmaceutiques par les distributeurs (BNV-d), qui comporte les quantités vendues et déclarées par les distributeurs de produits phytosanitaires aux fins du calcul de la redevance pour pollution diffuse qui leur est appliquée par leur agence de l'eau.

Les données sont aussi un vecteur de démocratie locale et de transparence sur l'action des collectivités. L'ouverture des données des finances publiques locales répond à l'enjeu de contrôle de l'action de l'administration par les citoyens. L'Observatoire des Finances et de la Gestion publique Locales (OFGL) a ainsi lancé, le 4 février 2020, le portail de données financières et de gestion du secteur public local[14] : celui-ci *« vise à mettre à disposition, dans une plateforme en ligne, unique et ouverte, des données relatives aux finances et à la gestion des collectivités locales françaises, d'en permettre l'accès, la compréhension et l'analyse, d'en faciliter les réutilisations »*. Les données utilisées dans le portail sont toutes issues de jeux de données mis à disposition de manière ouverte par différents producteurs, en particulier la direction générale des finances publiques (DGFiP), la direction générale des collectivités locales (DGCL) et l'INSEE.

Les résultats des actions conduites par les services déconcentrés sont ensuite un exemple de transparence « de proximité », montrant l'intérêt de la donnée y compris dans la vie quotidienne. L'ouverture des résultats de contrôles sanitaires, disponibles sur internet depuis 2017[15], participe à la transparence des informations vis-à-vis du public. Ces résultats, issus des contrôles effectués tout au long de la chaîne alimentaire, sont mis à jour quotidiennement et restent visibles pendant un an. Ils permettent de connaître le niveau d'hygiène des établissements de production, de transformation et de distribution des produits alimentaires.

Au-delà de l'objectif de transparence de la vie publique, **l'ouverture des données participe aussi à la simplification des démarches administratives des citoyens** dans la mesure où elle facilite l'échange de données entre administrations. La démarche « dites-le nous une fois » illustre l'intérêt de ces échanges de données (cf. *infra*).

[13] https://www.dashboard.covid19.data.gouv.fr

[14] https://data.ofgl.fr

[15] https://www.alim-confiance.gouv.fr/

Comment ça se passe à l'étranger ?

En Estonie, une plateforme permet de suivre toutes les recettes et dépenses liées au gouvernement pendant une période donnée. Ce projet appelé Riigiraha a été lancé en 2014 et donne une vision au niveau national et local (base du contribuable, les résultats opérationnels, les activités d'investissement, les salaires mensuels des fonctionnaires élus et nommés, les dépenses en programmes sociaux et d'autres statistiques financières).

Au Royaume-Uni, le National Health Services (NHS) Digital a rendu accessible au public les données relatives à la santé et à l'aide sociale sous l'Open Government Licence sur son site web et sur data.gov.uk. Ces données, anonymisées et regroupées conformément aux normes nationales, comprennent des publications statistiques, des données produites en réponse aux demandes de renseignement découlant du Freedom of Information Act de 2000 et des données structurelles et de dépenses.

L'ouverture des codes sources est un gage de transparence de la décision publique

L'ouverture des codes sources constitue également un enjeu essentiel de la transparence de la vie publique par la mise à disposition des citoyens de la traduction informatique des algorithmes de décision. Cela permet ainsi d'expliciter la méthode de prise de décision. Quelques exemples permettent d'illustrer l'intérêt de cette démarche et l'attention grandissante des citoyens.

L'exemple le plus notoire est celui de l'ouverture du code source de Parcoursup. L'outil de gestion des parcours des étudiants, Parcoursup, s'inscrit dans la loi n°2018-166 du 8 mars 2018 relative à l'orientation et à la réussite des étudiants (ORE) qui définit les modalités d'admission dans l'enseignement supérieur. Il a alimenté de nombreux débats concernant la transparence de la procédure d'affectation des futurs étudiants et a même été source de contentieux[16].

En 2017, le code source de l'implémentation des algorithmes mis en œuvre dans la plateforme Parcoursup ont été publiés[17] et ils font depuis l'objet d'une mise à jour annuelle. Cette publication répond en partie à la demande de transparence des étudiants et de leur famille, mais elle est doublement incomplète[18] : d'une part, il s'agit là d'une petite partie du code source de la plateforme, et d'autre part, cette partie de l'implémentation ne correspond qu'à une partie des algorithmes qui déterminent l'orientation des étudiants.

[16] En juin 2018, l'Union nationale des étudiants de France (UNEF) a demandé à l'université des Antilles de lui communiquer les documents informatiques qu'elle utilisait pour l'examen des candidatures présentées via la plateforme Parcoursup. L'université ayant refusé, l'UNEF a contesté ce refus devant le tribunal administratif de la Guadeloupe. Par un jugement du 4 février 2019, le tribunal administratif a jugé que l'université devait délivrer à l'UNEF les documents demandés. L'université s'est pourvue en cassation contre ce jugement devant le Conseil d'État. Par décision du 12 juin 2019, celui-ci a estimé que, lorsqu'un établissement reçoit des demandes supérieures à ses capacités d'accueil et met en place une sélection des candidatures, il est seulement tenu d'informer les candidats qui en font la demande des critères et modalités d'examen de leur candidature ainsi que des motifs pédagogiques qui justifient la décision prise à leur égard, conformément à ce que la loi du 8 mars 2018 a prévu.

[17] https://framagit.org/parcoursup/algorithmes-de-parcoursup

[18] Nous insistons sur l'importance de maintenir la distinction entre les codes sources et les algorithmes, ces derniers pouvant faire l'objet d'une publication dédiée (hors implémentation) et nécessitant des explications propres, lorsqu'ils interviennent dans une décision administrative individuelle.

En effet, les « algorithmes locaux », désignant les critères qui servent à pré-ordonner les dossiers des candidats avant que les commissions d'examen des vœux ne les examinent, ne sont pas publiés. Cette pratique est justifiée par l'article L. 612-3 du code de l'éducation selon lequel, si les candidats peuvent obtenir la communication des critères, modalités d'examen et motifs pédagogiques justifiant la décision prise à leur égard, la communication des algorithmes utilisés est réservée aux seuls candidats qui en font la demande, une fois la décision prise, et pour la seule décision les concernant. Dans son rapport de février 2020, la Cour des comptes[19] préconisait cependant d'abroger cette disposition législative, *« et à défaut et à tout le moins, […] obtenir des universités, des écoles et des lycées qu'ils recourent eux-mêmes à cette publication » en recommandant finalement de « rendre publics les "algorithmes locaux" utilisés par les commissions d'examen des vœux pour l'ensemble des formations proposées ».* Cette recommandation fait écho aux demandes de nombreux acteurs de l'enseignement supérieur réclamant une plus grande transparence dans le processus d'affectation des étudiants.

Le Conseil constitutionnel, saisi le 16 janvier 2020 par le Conseil d'État d'une question prioritaire de constitutionnalité (QPC), a toutefois reconnu, dans sa décision du 3 avril 2020[20], la conformité de cette disposition législative à la Constitution, tout en soulignant que cela ne dispense pas *« chaque établissement de publier, à l'issue de la procédure nationale de préinscription et dans le respect de la vie privée des candidats, le cas échéant sous la forme d'un rapport, les critères en fonction desquels les candidatures ont été examinées et précisant, le cas échéant, dans quelle mesure des traitements algorithmiques ont été utilisés pour procéder à cet examen ».*

De manière plus ancienne, en 2016, l'ouverture du code source utilisé pour le calcul de l'impôt sur le revenu a permis de renforcer la transparence de l'action de l'État, notamment par le biais de travaux de recherche réalisés sur ce sujet. Le code source est publié dans ses versions allant des revenus de 2010 à 2019. Ainsi, des chercheurs de l'INRIA[21] se sont emparés des données et du code source publiés par la DGFiP pour réaliser une étude sur l'implémentation informatique des dispositions légales du code des impôts. Si une convention avec l'administration fiscale a été nécessaire pour avoir accès à l'ensemble du code source (qui n'est pas disponible intégralement sur internet pour des raisons de sécurité, selon l'argument avancé par la DGFiP), cette étude propose une sémantique formelle du langage informatique utilisé par la DGFiP permettant d'inférer des méta-propriétés sur le calcul de l'impôt. Celles-ci peuvent ensuite compléter et affiner les analyses économiques existantes sur les effets redistributifs de l'impôt sur le revenu, mais aussi de diverses allocations. L'étude invite ainsi à *« une formalisation systématique des portions algorithmiques de la loi [qui] permettrait d'augmenter le niveau d'assurance sur la cohérence du système socio-fiscal français »*.

La fermeture des codes sources peut entretenir une méfiance à l'égard de l'État

A contrario, certains codes sources gagneraient à être publiés, en tout ou partie, afin d'offrir une transparence sur la conservation des données des citoyens.

C'est par exemple le cas du code source de FranceConnect, qui est un dispositif permettant de garantir l'identité d'un utilisateur en s'appuyant sur des comptes existants pour lesquels son identité a déjà été vérifiée. Il permet ainsi de faciliter les démarches en ligne des usagers sur un certain nombre de sites (impots.gouv.fr, ameli.fr, l'Identité Numérique La Poste, MobileConnect et moi, msa.fr et Alicem). L'ouverture du code source serait un gage supplémentaire, sur le fait qu'il ne conserve aucune trace de connexion et sur la confiance que les citoyens peuvent lui accorder.

[19] Un premier bilan de l'accès à l'enseignement supérieur dans le cadre de la loi orientation et réussite des étudiants, Cour des Comptes, février 2020.

[20] Décision du Conseil constitutionnel n° 2020-834 du 3 avril 2020.

[21] Étude formelle de l'implémentation du code des impôts, D. Mérigoux, R. Monat, C. Gaie, décembre 2019.

De la même façon, le code source de l'identité numérique régalienne (AliceM notamment), s'il est ouvert, pourra renforcer la confiance dans le dispositif. En lien avec la mise en œuvre du règlement (UE) n° 2019/1157 du Parlement et du Conseil du 20 juin 2019, relatif au renforcement de la sécurité des cartes d'identité des citoyens de l'Union et des documents de séjour délivrés aux citoyens de l'Union et aux membres de leur famille exerçant leur droit à la libre circulation, il vise à harmoniser et à renforcer la sécurité de la carte nationale d'identité. Dans ce cadre, il est essentiel qu'au moins certaines briques du code source soient ouvertes, afin d'appuyer la confiance des citoyens en explicitant la manière dont sont utilisées les données.

2.2. Pour les entreprises, susciter l'innovation et moderniser l'économie

La politique d'ouverture des données et des codes sources publics vise également à offrir un terrain fertile à l'innovation, et plus globalement à moderniser l'économie. La loi pour une République numérique avait ainsi pour ambition de *« libérer l'innovation en faisant circuler les informations et les savoirs, pour armer la France face aux enjeux globaux de l'économie de la donnée »*. Le service public de la donnée s'inscrit dans cet objectif au travers de la mise à disposition des jeux de données de référence, qui sont les données qui « *constituent une référence commune pour nommer ou identifier des produits, des services, des territoires ou des personnes* », qui sont « *réutilisées fréquemment par des personnes publiques ou privées autres que l'administration qui les détient* » et dont la réutilisation nécessite donc « *qu'elles soient mises à disposition avec un niveau élevé de qualité* »[22]. Le CRPA fait également mention de l'intérêt « *économique, social, sanitaire ou environnemental* » des documents administratifs[23].

Par ailleurs, la libération de jeux de données publiques constitue un enjeu essentiel du développement de l'intelligence artificielle. L'étude menée début 2020 par Cap Gemini, « *State of AI* », met en évidence l'importance pour les entreprises de bénéficier d'un accès décloisonné à un large patrimoine de données pour développer les cas d'usage (cf. partie 3).

Conduite pour le Portail européen des données, initiative de la Commission européenne, l'étude *« The Economic Impact of Open data, Opportunities for value creation in Europe »* de Cap Gemini évalue la valeur économique des produits, services et contenus enrichis ou rendus possibles par l'*open data* à 184 Mds€ à l'échelle de l'Union Européenne en 2019 et 334 Mds€ en 2025. À l'échelle de la France, cela représenterait **une valeur économique de 28 Mds€ pour 2019**, soit 1.19% du PIB[24].

La même étude estime que la croissance annuelle du PIB pourrait être de l'ordre de 4,3 % dans le scénario le plus conservateur à 15,7 % dans un scénario où le partage des données se développe, le patrimoine de donnée se déconcentre et où de multiples usages innovants émergent. Bien entendu, ces chiffres sont à prendre avec beaucoup de précaution compte tenu de la crise de la Covid19 et de ses impacts sur l'économie rendant délicat l'exercice de projection.

Cette étude est une synthèse des différentes estimations faites au niveau macroéconomique. La mission estime nécessaire de conduire une évaluation plus fine de l'impact économique, social et scientifique de l'ouverture et du partage des données et des codes sources, qui pourrait par exemple être confiée à France Stratégie (cf. recommandation n°3). Ce type d'étude pourrait s'appuyer sur des indicateurs permettant de mesurer les réutilisations des jeux de données et de codes sources.

Les retombées économiques de l'ouverture des données

Les producteurs publics de données peinent souvent à identifier les retombées économiques de l'ouverture de leurs jeux de données et de leurs codes sources. Plusieurs exemples illustrent toutefois l'impact de la publication des codes sources sur le développement de l'activité économique, comme la base SIRENE et la base des demandes de valeurs foncières (DVF).

[22] Article L. 321-4 du CRPA.

[23] Article L. 312-1-1 du CRPA.

[24] L'étude combine une démarche macro-économique (« top down ») de synthèse des études existantes et une démarche micro-économique (« bottom up ») d'évaluation des bénéfices de cas d'usages em-blématiques pour chiffrer la valeur économique des produits, services et contenus enrichis ou rendus possibles par l'Open Data.

La base SIRENE est une des bases les plus exploitées par le monde économique, et considérée comme une base pivot dans le cadre de la stratégie pour un État-plateforme. Après son ouverture en 2017, le nombre de réutilisateurs mensuels réguliers est passé de 500 en moyenne avant 2017 à 4 400 (cf. cas d'usage sur la base SIRENE). La base SIRENE représentait environ 10 M€ de recettes publiques lorsqu'elle n'était pas ouverte, pour 500 réutilisateurs, ce qui conduit à estimer, pour les 4 400 réutilisateurs actuels, la valeur libérée au bénéfice de l'économie à 88 M€.

L'ouverture de la base des demandes de valeurs foncières (DVF) en 2019 est un autre exemple d'ouverture à fort impact, dont la mise à disposition a d'ailleurs été d'abord demandée par les usagers. En 2011, la base (dénommée Patrim) a été ouverte aux collectivités et à certains acteurs professionnels sous l'impulsion du législateur, souhaitant renforcer la transparence des marchés immobiliers, et d'une association d'usagers, devenue le groupe national « demande de valeurs foncières » (GNDVF), qui fédère la communauté de réutilisateurs et mutualise leurs connaissances.

Outre son succès auprès du grand public, la base crée de l'activité économique en stimulant l'innovation et le développement de nouveaux services. Elle est aujourd'hui pleinement exploitée par les producteurs de données de prix et de loyers (SeLoger, Meilleurs Agents, PriceHubble, Yanport) ou encore par les acteurs accompagnant les collectivités territoriales (publics comme le CEREMA, privés comme Spallian). Elle permet également le développement de la PropTech et des start-up cherchant à renouveler l'approche de la valeur des logements, comme Homedata, Homiwoo, Liberkeys, KelQuartier, Bien'Ici et CityScan[25].

Pourtant, le potentiel de réutilisation de la base DVF est loin d'être épuisé. D'une part, l'ouverture de la base DVF a représenté une diminution d'informations par rapport à celles déjà accessibles aux utilisateurs depuis 2011, certains champs[26] ayant été fermés par la DGFiP par rapport à la base Patrim, alors qu'ils permettent de connaître la nature de la transaction immobilière et de renforcer la fiabilité de l'exploitation de la base. D'autre part, et surtout, plusieurs acteurs plaident en faveur d'un enrichissement de la base DVF par d'autres données, comme celles contenues, y compris en stock, dans la base comportant la description des biens (Majic)[27], les plans locaux d'urbanisme (PLU) ou la base des permis de construire (Sitadel).

L'ouverture des codes sources, un vecteur majeur de mutualisations

Par ailleurs, l'ouverture des codes sources publics permet la création de valeur économique. S'il n'existe pas d'étude exhaustive sur cet impact, ni de méthodologie définitivement arrêtée de l'évaluation des effets de l'ouverture d'un code public, quelques exemples emblématiques permettent d'illustrer l'intérêt de cette démarche pour les entreprises.

Un premier exemple marquant est fourni par le succès de Scikit-learn, bibliothèque libre Python destinée à l'apprentissage automatique. À l'origine initiée par un chercheur de l'INRIA, celle-ci a été mise à disposition du public avec l'objectif d'appuyer son développement par de nombreux contributeurs. Aujourd'hui cette librairie fait partie des librairies les plus utilisées dans le monde. Financé à l'origine par l'INRIA, ce projet bénéficie de dons institutionnels et privés (parmi lesquels Microsoft, AXA, BNP Paribas CARDIF, Fujitsu, Intel etc.) témoignant de l'intérêt des entreprises pour ce type de librairie.

Un autre exemple est le code node-csv-string. Cette bibliothèque de code est celle que la plateforme *code.etalab.gouv.fr* présente comme étant la plus réutilisée (844 réutilisations d'après GitHub). Il s'agit d'un code source développé sous licence libre par un laboratoire de recherche du CNRS, l'Institut de l'information scientifique et technique (INIST), servant à analyser des fichiers csv à l'intérieur de programmes informatiques.

[25] Claire Juillard, Sciences Po, « Produire des données de prix et de loyers à l'heure de la PropTech : quel rôle pour l'État ? », janvier 2020.

[26] Trois champs ont été fermés : l'article applicable du code général des impôts (CGI), le code SAGES du service de la publicité foncière compétent, et la référence du document de publication.

[27] Le Cerema met à disposition ces informations issues de Majic dans la base enrichie « DV3F » mais elle n'est pas en *open data.*

L'accès limité ou payant aux données publiques est un frein à l'innovation

A contrario, des entrepreneurs interrogés par la mission regrettent de ne pas avoir accès à certains jeux de données, pourtant susceptibles de générer de l'innovation et, par là même, de la création de valeur.

C'est par exemple le cas de l'accès aux données des contrôles techniques des véhicules (base gérée par l'UTAC[28] sous tutelle du ministère chargé des transports) et aux données du système d'immatriculation des véhicules (SIV) relevant du ministère de l'Intérieur. Depuis 2016, Certificare**,** une start-up française, développe un outil permettant de sécuriser les transactions de véhicules d'occasion sous la forme d'un rapport apportant des garanties quant à l'absence de fraude au kilométrage, dont les acteurs du secteur et la Fédération Internationale de l'Automobile (FIA) s'accordent pour dire qu'elle concerne 10 à 15 % des véhicules d'occasion en France aujourd'hui. Le véhicule n'est alors plus en phase avec son programme de maintenance, d'où des répercussions directes en matière de sécurité routière et d'environnement.

Cette entreprise a développé des partenariats avec des acteurs privés détenant de l'information sur le kilométrage des véhicules (assureurs, réseaux de réparateurs, constructeurs, notamment) permettant de commercialiser cette offre auprès des professionnels du commerce automobile.

Toutefois, cette start-up voit son développement grandement freiné et sa pérennité remise en cause en raison de plusieurs difficultés :

- **en premier lieu, l'accès aux données concernant le kilométrage des véhicules relevés lors des contrôles techniques n'est pas prévu pour les propriétaires de véhicules**, en raison d'une disposition réglementaire du code de la sécurité routière qui prévoit que ces informations sont réservées à l'Organisme Technique Central (OTC) . Si des travaux sont en cours pour modifier l'article R. 323-20 du code de la route et autoriser les propriétaires de véhicules à y accéder, par exemple dans le cadre d'Histovec[29], cela ne lève pas pour autant les obstacles rencontrés par la start-up puisque l'administration doit accepter d'ouvrir ces données (par exemple par API) afin d'industrialiser le processus, conformément à la conformité RGPD développée et maîtrisée par Certificare (consentement, mandat et intérêt légitime) ;
- **en second lieu, le système d'immatriculation des véhicules (SIV) nécessite l'achat d'une licence annuelle auprès du ministère de l'intérieur** afin de pouvoir utiliser les données : ainsi, un arrêté du 11 avril 2011 fixe le montant de la redevance due en contrepartie de la mise à disposition des informations issues du système d'immatriculation des véhicules, avec trois niveaux de tarification variant en fonction du nombre de lignes et de champs de données livrés et en fonction des modalités de réutilisation (usage interne, vente de prestations à des tiers, et rediffusion de données à d'autres titulaires d'une licence à finalité commerciale). **Le coût pour la start-up pourrait atteindre 500 000 € par an.**

[28] Société de droit privé désigné par le décret 91-1021 du 4 octobre 1991 comme en tant qu'Organisme Technique Central (OTC) du contrôle technique des véhicules.

[29] Système d'information développé par le ministère de l'intérieur qui permet aux particuliers d'obtenir l'historique administratif du véhicule mais ne traite pas de l'ensemble des informations utiles pour informer l'acheteur quant à l'état du véhicule (historique d'entretien, véhicule hors d'usage, kilométrage, entre autres).

Par ailleurs, les données dont disposent les départements en matière sociale pourraient guider davantage l'activité privée. Par exemple, la publication des données sur les aides et services à destination des personnes âgées par les départements peut permettre aux acteurs privés de construire des offres de services adaptées dans les territoires concernés. La mise en place d'un portail national d'information pour les personnes âgées et leurs proches[30] par la caisse nationale pour la solidarité et l'autonomie (CNSA) peut répondre à ce besoin et participer à la visibilité de ces données pour l'ensemble des acteurs. A cet égard, la publication des jeux de données sur le prix hébergement et des tarifs dépendance des EHPAD constitue une première avancée en matière de transparence, d'abord pour les personnes âgées et leurs proches, mais aussi pour les acteurs privés. Cette information n'est cependant pas complète : s'agissant par exemple des EHPAD, ceux-ci ne remontent pas tous leurs tarifs et ces tarifs ne sont pas forcément complets (le prix de l'hébergement peut varier en fonction de nombreux critères, comme les loyers d'un logement). Il est toutefois à noter que le portail est complémentaire des sites internet des départements et autres points d'information locaux : l'annuaire des départements comporte ainsi une fiche détaillée de chaque département avec les coordonnées, les informations et les services proposés par chaque département sur son site internet.

De même, la mise à disposition des données agrégées relatives aux profils et au nombres de personnes en situation de handicap sur un territoire donné, issues des systèmes d'information des maisons départementales des personnes handicapées (MDPH), pourrait être de nature à faciliter la mise en accessibilité des infrastructures et services aussi bien publics que privés.

Plus globalement, l'ouverture des données de la CNSA doit faciliter le partage de données entre acteurs du secteur médico-social ainsi, plus globalement, qu'entre acteurs du champ de l'autonomie. Elle pourrait permettre également aux acteurs les exploitant d'innover et à la CNSA de profiter de ces innovations pour améliorer sa connaissance des offres et des besoins.

Comment ça se passe à l'étranger ?

Au Royaume-Uni, la Commission Géospatiale a été créée en 2017 afin de maximiser la valeur économique des données de localisation et l'ouverture gratuite des données de l'*Ordnance Survey Master Map* en 2018 (carte digitale unique inventoriant l'ensemble des données topographiques du Royaume-Uni, initialement payante), avec l'objectif d'augmenter les opportunités de développement de technologies basées sur les données de localisation, comme les voitures connectées

En Israël, le ministère des transports et l'ICT Authority ont organisé un hackathon sur les données et codes sources publics : le vainqueur de la compétition a développé une application de gestion en temps réel des horaires de passage des bus qui a donné naissance à la société MOOVIT, société qui a depuis été rachetée par Intel pour 900 millions de US$. Dans le cadre du programme national en faveur de la mobilité intelligente, les codes sources du ministère des Transports, des municipalités et des entreprises ont été mis à la disposition du public.

[30] www.pour-les-personnes-agees.gouv.fr

2.3. Pour les acteurs publics, éclairer la décision et améliorer la qualité des services publics

L'*open data* est encore parfois le meilleur moyen pour des administrations de partager des informations, d'abord parce que le partage nécessite généralement des procédures longues et complexes (signature de conventions), ensuite parce que le service producteur de la donnée n'est généralement pas à même de connaître les besoins des autres acteurs publics s'agissant de sa donnée. À titre d'exemple, la DGFiP ne peut pas soupçonner l'intégralité des usages qui peuvent être faits de ses données et ne peut donc pas les partager avec l'ensemble des réutilisateurs potentiels par le biais de conventions de gré à gré. Ce frein est majeur, car le partage de données entre acteurs publics – qu'il s'agisse des données produites par l'État, les opérateurs ou les collectivités territoriales – est un des principaux leviers pour améliorer l'efficacité de l'action publique et la conduite des politiques publiques, en permettant leur analyse systémique (tous acteurs et tous secteurs confondus).

L'ouverture des données impose aussi aux acteurs publics de replacer la donnée au cœur de leurs actions en les invitant à en rendre compte de manière plus précise et la plus étayée possible. En ce sens, cette démarche d'*open data* renforce l'exigence de pilotage de l'action publique par la donnée, et n'est qu'une version plus fiable et plus réactive du contrôle de gestion et de l'évaluation des politiques publiques (prise en compte des projections et des modèles permettant de comprendre et de mesurer l'impact d'une politique, mieux anticiper et gérer les crises en exploitant « les signaux faibles » et la remontée des alertes, etc.). Elle constitue également un facteur important de l'amélioration des services rendus au public. L'*open data* permet d'exposer la donnée ou le code à des acteurs externes, qui peuvent contribuer à les enrichir et donc à améliorer *in fine* l'action publique.

L'ouverture des données produites par les acteurs privés présentant un intérêt général aide également les acteurs publics à mieux comprendre les comportements et les besoins des différents acteurs dans le cadre de l'élaboration et du pilotage des politiques publiques (cf. partie 5).

Une amélioration du pilotage des politiques publiques

La donnée est d'abord un moyen de renforcer la fiabilité et la réactivité du contrôle de gestion et de l'évaluation des politiques publiques. Elle est le vecteur principal de la démarche de performance, la donnée clé étant l'indicateur d'une politique publique, au sens de la loi organique relative aux lois de finances (LOLF), qui a fondé cette démarche de performance des dépenses publiques. Taux d'exécution, taux de réalisation, taux de satisfaction sont autant de données qui servent directement les décideurs publics dans leurs stratégies et leurs orientations.

L'ouverture des données épidémiologiques dans le cadre de l'épidémie de Covid19 a mis en évidence l'importance de la collecte de données dans le pilotage de la crise. Par ailleurs, la recherche de transparence de l'action publique et la pédagogie nécessaires pour susciter l'adhésion aux décisions prises ont joué un rôle dans la place centrale réservée aux données (cf. cas d'usage sur les données et les modèles épidémiologiques).

Le baromètre de l'action publique mis en place par la ministre de la transformation et de la fonction publiques en 2020 répond à cet objectif d'amélioration du pilotage des politiques publiques, au niveau national et au niveau territorial, en constituant un tableau de bord de l'ensemble des politiques publiques et de leurs résultats. Il permet de redonner davantage de sens à l'action publique en l'engageant dans une démarche de performance, et de garantir la transparence sur le service rendu aux usagers. La ministre de la transformation et de la fonction publiques affirme ainsi que ce baromètre permet de refléter aussi bien les bons que les mauvais résultats ; il s'agit donc bien d'un outil de pilotage et d'un vecteur d'efficacité de l'action publique[31].

[31] « *Moi, ce que j'aimerais, c'est qu'on soit transparent avec les Français. Ça ne sert à rien de faire le tour de France de ce qui marche, ni faire le tour de France de ce qui ne marche pas. Ce qu'il faut c'est qu'on puisse dire : voilà les calendriers qu'on se fixe, voilà les jalons qu'on se pose, voilà les garanties qu'on vous donne. Le baromètre il est intéressant, il vous dit à la fois quand il fait beau, et quand il ne fait pas beau, quand il pleut ou quand il y a du soleil.* »
(Amélie de Montchalin, entretien à France Inter le 20 juillet 2020)

Une simplification des démarches et une amélioration de la qualité du service public

Au niveau national, la donnée est vecteur de simplification des démarches et de partage d'informations entre acteurs publics, avec la plateforme *demarches-simplifiees.fr* développée par la DINUM en 2018 afin d'offrir aux administrations un service clé en main pour numériser des démarches administratives et s'affranchir des formulaires papiers. L'ouverture du code source a ainsi permis à l'ADULLACT, association promouvant le logiciel libre au sein des collectivités, de déployer à son tour ce type d'outils auprès des collectivités territoriales.

Les opérateurs s'engagent aussi dans cette démarche. Dans le cas de Pôle Emploi, le partage des données permet de personnaliser et de simplifier le parcours des utilisateurs. Par exemple, E-dem CPAM de Vendée est un service de déclaration en ligne qui facilite le recours aux droits des bénéficiaires de leur territoire. En mettant à disposition le bouton « se connecter avec Pôle emploi » à cette branche de la CPAM et en remontant les expériences déclarées par les employeurs, Pôle Emploi permet à la CPAM de mieux accompagner les bénéficiaires des aides pour faire valoir leurs droits.

À l'échelon local, la collecte de données urbaines, publiques ou privées, est un élément essentiel pour le développement de la ville intelligente (« *smart city* ») et l'amélioration des services rendus aux usagers. De nombreux cas d'usage de la donnée sont ainsi développés par certaines municipalités pour le contrôle de stationnement (pour mieux déployer les équipes de la police municipale), sur le suivi des îlots de chaleur *via* des capteurs météorologiques permettant de guider les choix urbanistiques, sur le gaspillage alimentaire dans les cantines scolaires, sur le ramassage des déchets, etc.

Un autre exemple est le développement d'OpenCimetiere, logiciel libre de gestion des concessions de cimetières, développé par la commune d'Arles depuis 2003. Il permet la gestion de la place du défunt dans les concessions, la gestion des autorisations, la gestion du terme de la concession, la gestion des concessions libres, la gestion des opérations funéraires, l'archivage systématique de l'ensemble des données pour constituer une mémoire commune, etc. La publication de ce code source[32] permet ainsi de partager un outil avec l'ensemble des mairies.

L'ouverture des données favorise le partage entre acteurs publics

L'échange direct de données entre acteurs publics, *via* des interfaces de partage entre systèmes d'informations ministériels, est une composante de la politique d'ouverture des données publiques : elle vise notamment à simplifier des démarches pour les citoyens. Le programme *« Dites-le-nous-une-fois »* répond à cet objectif. Le fait d'avoir une vision large des demandeurs dans leurs relations avec les différents opérateurs et administrations doit conduire progressivement à une approche globale des services.

Pôle Emploi est un des opérateurs publics les plus engagés dans cette démarche de partage de la donnée avec ses partenaires. Afin d'améliorer la qualité de service auprès de ses usagers, Pôle Emploi met en place des échanges :

- avec la CNAV à la fois pour inciter le demandeur d'emploi à préparer son dossier de retraite le moment venu et éviter les doubles indemnisations ;
- avec des partenaires de Pôle emploi visent à mieux suivre les demandeurs d'emploi orientés vers des prestataires, organismes de formation, les Missions locales ;
- avec les caisses d'allocations familiales et les conseils départementaux afin de mieux repérer les allocataires du RSA ;
- avec l'OFII, les MDPH, les Cap Emploi ou les Missions locales pour mieux adapter l'offre de service, la fréquence des contacts ou sécuriser la poursuite d'un parcours d'insertion ou d'accompagnement.

[32] http://www.openmairie.org/catalogue/opencimetiere

A contrario, la mission a pu constater que de nombreuses directions de l'État, parfois même certains services au sein d'une même direction, ne partageaient pas la donnée, entraînant une perte d'efficacité significative pour la réalisation de leurs missions, comme dans le cas d'agents de la direction générale de la concurrence, de la consommation et de la répression des fraudes (DGCCRF) contraints de ressaisir des données de la direction générale des finances publiques (DGFiP).

Comment ça se passe à l'étranger ?

À Taïwan, « MyData Platform », lancée en juin 2020, permet aux citoyens de télécharger leurs données personnelles fournies aux services publics (carte d'identité, revenus, déclaration de propriété immobilière, immatriculation du véhicule, etc.), de se présenter à un guichet de service public sans pièces justificatives, et de donner accès aux justificatifs présents sur son espace personnel mais aussi d'autoriser le partage des documents téléchargés avec 40 services publics utilisant la plateforme, pour mettre à jour les informations en temps réel.

Un renforcement de la sécurité des systèmes d'information publics

Contrairement à une idée encore très répandue, l'ouverture des codes sources est un facteur de fiabilisation et de sécurisation des systèmes d'information, dès lors qu'elle permet de confronter le code à des retours utilisateurs. De manière complémentaire, le développement de logiciels libres permet d'enrichir le service. L'exemple le plus structurant est celui du **projet OpenCTI (Open Cyber Threat Intelligence),** un outil développé par l'ANSSI en partenariat avec l'équipe européenne de réponse aux urgences informatiques (CERT-EU) permettant de structurer, traiter et partager les connaissances en matière d'analyse de la cybermenace. En mettant à disposition les codes sources de l'application OpenCTI, l'ANSSI et le CERT-EU souhaitent faire évoluer l'outil en faisant contribuer les acteurs de la cybermenace au plus près de leurs besoins opérationnels ; cela vise à assurer un développement pérenne et toujours plus adapté aux besoins de la communauté.

Comment ça se passe à l'étranger ?

En Irlande, le All-Island Research Observatory (AIRO) fournit des ressources cartographiques permettant de comprendre la dynamique des régions et des comtés. Cet outil de cartographie est désormais régulièrement utilisé par les autorités nationales ou locales pour élaborer des Plans de développement des comtés ou des Plans économiques et communautaires locaux à travers l'Irlande.

En Israël, le site « hidavroot » (logiciel ouvert) a facilité le dialogue entre le comité Trajtenberg, nommé par le premier ministre en 2011 à la suite de manifestations de masse contre l'augmentation du coût de la vie, et les citoyens. Le code a ensuite été diffusé au sein de la communauté des logiciels libres afin de continuer son développement avec le logiciel Ideal Management.

Cet exemple amène à s'interroger sur la reproductibilité de ce type de service en France. Le manque d'appropriation de l'application StopCovid par les citoyens, avant un intérêt plus grand pour TousAntiCovid, a mis en évidence la réticence de la population française à partager des données personnelles avec la puissance publique. Dans ce contexte, il apparaît nécessaire de créer une confiance minimale dans les services de l'État en associant davantage la société civile pour établir les prérequis indispensables permettant d'engager ce type de démarche (par exemple, publication des codes sources utilisés pour rendre transparente les modalités d'utilisation des données personnelles des usagers, renforcement de la sécurité informatique, avis sans réserve de la CNIL, etc.).

2.4. Pour la recherche, fertiliser l'activité et favoriser les collaborations

Les exemples de réutilisation des données et des codes sources sont innombrables et les quelques illustrations ci-dessus soulignent l'intérêt de la recherche fondée sur la réutilisation de données (cf. cas d'usage sur les données et modèles épidémiologiques) ou de codes sources (cf. exemple de réutilisation du code source de calcul de l'impôt sur le revenu).

Au-delà de la mise à disposition des données et des codes sources publics, la science ouverte constitue également un enjeu important pour le partage des connaissances (cf. encadré). Le comité pour la science ouverte, créé en 2018, précise ainsi que *« La science ouverte est la diffusion sans entrave des publications et des données de la recherche. Elle s'appuie sur l'opportunité que représente la mutation numérique pour développer l'accès ouvert aux publications et -autant que possible- aux données de la recherche. »*

Une étude publiée dans *Plus One* en 2020 a étudié les impacts sur le nombre de citations pour un article de recherche lorsque des métadonnées et données y étaient associées. L'étude se fonde sur 531 000 articles dans le domaine de la biosanté et montre que lorsqu'un article est associé à des données en annexe, il est cité 25 % de fois en plus par rapport à un article sans données associées. Ainsi, avoir des données associées est un facteur de publicité et de crédibilité accrue des résultats. Si la mission n'a pas connaissance d'une étude similaire sur le partage des codes sources, on peut cependant supposer que cette démarche peut avoir le même type d'effet.

Par ailleurs, différents projets de logiciels libres incubés dans la sphère publique participent de cette dynamique et visent à faciliter l'accès à l'analyse de données pour un ensemble d'acteurs, aussi bien publics que privés. L'exemple de **Scikit-learn** mentionné supra témoigne de l'intérêt de ce type de démarche.

Les axes du comité de la science ouverte

Le Plan national pour la science ouverte a été annoncé par la ministre de l'enseignement supérieur et de la recherche le 4 juillet 2018.

Le premier axe est de généraliser l'accès ouvert aux publications, (i) en rendant obligatoire la publication en accès ouvert des articles et livres issus de recherches financées par appel d'offres sur fonds publics, (ii) en créant un fonds pour la science ouverte, et (iii) en soutenant l'archive ouverte nationale HAL et en simplifiant le dépôt par les chercheurs qui publient en accès ouvert sur d'autres plateformes dans le monde.

Le deuxième axe est de structurer et d'ouvrir les données de la recherche, (i) en rendant obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics ; (ii) en créant la fonction d'administrateur des données et le réseau associé au sein des établissements ; (iii) en créant les conditions et promouvant l'adoption d'une politique de données ouvertes associées aux articles publiés par les chercheurs.

Le troisième axe est de s'inscrire dans une dynamique durable, européenne et internationale. L'enjeu est de transformer les pratiques scientifiques pour qu'elles contribuent à la structuration du paysage international de la science ouverte par la diffusion des meilleurs usages et des meilleures pratiques. Le but est aussi de généraliser les pratiques quotidiennes de la science ouverte, dans le domaine des publications, des données, de la propriété intellectuelle et de l'évaluation par les pairs, et de contribuer à un écosystème à la fois résilient, régulé et transparent, œuvrant dans le sens des intérêts de la communauté scientifique. Le plan prévoit ainsi de (i) développer les compétences en matière de science ouverte notamment au sein des écoles doctorales ; (ii) d'engager les opérateurs de la recherche à se doter d'une politique de science ouverte ; (iii) de contribuer activement à la structuration européenne au sein du *European Open Science Cloud* et par la participation à GO FAIR.

Source : Site internet du MESRI

3. L'ouverture des données et des codes sources : quels risques ?

Si les apports de l'ouverture des données et des codes sources sont nombreux, il convient de ne pas négliger les risques. Protéger la vie privée des individus, ne pas porter atteinte à la sécurité des systèmes d'information, par exemple, justifient de ne pas procéder à la publication des données et des codes sources : une voie intermédiaire entre l'ouverture au public et la fermeture complète peut toutefois être imaginée pour faciliter l'accès sécurisé ou l'échange de ces données ou codes sources (cf. partie 3). D'autres risques avancés par les acteurs rencontrés renvoient davantage à une question d'acculturation et d'appréciation des risques par les différentes parties prenantes, et nécessitent de conduire une réflexion associant la société civile.

Une sensibilité particulière sur la protection des données personnelles

Le risque d'identification des données personnelles est le frein à l'ouverture le plus souvent mis en avant par les différentes personnes auditionnées par la mission. Ainsi, la CNIL souligne à juste titre que *« la protection des données personnelles n'est pas un sujet annexe à l'ouverture mais une condition essentielle du partage pour assurer la confiance »*. Parmi les données personnelles, certaines revêtent une sensibilité particulière ; il s'agit plus particulièrement des données listées à l'article 9 du RGPD, à savoir les données révélant *« l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique »*. À cet égard, la réglementation nationale spécifique et plus restrictive que le droit européen concernant les traitements des données de santé (cf. partie 2) révèle la sensibilité particulière des Français sur ce sujet.

La crainte de divulguer les données personnelles conduit du reste de nombreux acteurs à s'autocensurer dans la mise en œuvre de la politique d'ouverture des données. La CNIL souligne ainsi que le RGPD n'est pas un règlement qui interdit le traitement de données, mais qui les encadre en édictant des principes de travail et non des principes d'abstention. Afin de faciliter l'appropriation du RGPD par l'ensemble des acteurs, le site internet de la CNIL dispose d'une page web dédiée pour les start-up sur ce sujet.

Dans le cas des données personnelles, la publication en ligne est possible sous réserve notamment de pouvoir anonymiser le document administratif : l'anonymisation est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et ce de manière irréversible. Dans ce cas, la législation relative à la protection des données ne s'applique plus, car la diffusion ou la réutilisation des données anonymisées n'a pas d'impact sur la vie privée des personnes concernées. Sur son site, la CNIL donne un certain nombre d'éléments sur l'anonymisation des données, qui ne doit pas être confondue avec le principe de pseudonymisation (cf. encadré).

Anonymisation et pseudonymisation

L'anonymisation ne doit pas être confondue avec la pseudonymisation. La pseudonymisation est un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans information supplémentaire.

En pratique, la pseudonymisation consiste à remplacer les données directement identifiantes (nom, prénom, etc.) d'un jeu de données par des données indirectement identifiantes (alias, numéro séquentiel, etc.).

La pseudonymisation permet ainsi de traiter les données d'individus sans pouvoir identifier ceux-ci de façon directe. En pratique, il est toutefois bien souvent possible de retrouver l'identité de ceux-ci grâce à des données tierces : les données concernées conservent donc un caractère personnel. L'opération de pseudonymisation est parfois réversible, contrairement à l'anonymisation.

La pseudonymisation constitue une des mesures recommandées par le RGPD pour limiter les risques liés au traitement de données personnelles.

Source : CNIL.

Selon les critères définis par les autorités de protection des données européennes un jeu de données est véritablement anonyme dès lors qu'il n'est pas possible d'isoler un individu dans le jeu de données (principe d'individualisation), qu'il n'est pas possible de relier entre eux des ensembles de données distincts concernant un même individu (principe de corrélation) et qu'il n'est pas possible possible de déduire, de façon quasi certaine, de nouvelles informations sur un individu (principe d'inférence). À défaut de remplir parfaitement ces trois critères, le responsable de traitement doit démontrer, via une évaluation approfondie, que le risque de ré-identification avec des moyens raisonnables est nul. Les techniques d'anonymisation et de ré-identification étant amenées à évoluer régulièrement, la CNIL rappelle la nécessité, pour tout responsable de traitement concerné, d'effectuer une veille régulière pour préserver, dans le temps, le caractère anonyme des données produites.

Des risques liés à la sécurité informatique

De nombreux acteurs publics mettent en avant l'argument de la sécurité informatique pour ne pas publier les données. Cet argument peut s'entendre, mais en partie seulement, comme le souligne l'ANSSI. Quand des données sont rendues publiques, l'agence recommande en effet d'avoir systématiquement une analyse de risque pour vérifier que cette démarche ne conduit pas à faciliter des réutilisations malveillantes. Par exemple, elle cite une pratique ancienne consistant à publier sur internet, sur tous les sites des collectivités, les versions utilisées par les serveurs, les outils de publication. Cela revient à pointer du doigt les failles de sécurité. Or, il n'y a pas d'intérêt particulier à publier ce genre de données. Si les attaquants disposent d'autres moyens pour obtenir ces informations, leur publication leur facilite la tâche.

En réalité, les acteurs faisant valoir la sécurité des systèmes d'information semblent méconnaître la possibilité de renforcer leur résilience offerte par la démarche d'ouverture des codes sources, et que le directeur général de l'ANSSI a confirmé de manière expresse à la mission. Selon les termes Guillaume Poupard, en effet, « *l'identification des failles est plutôt utile sur un mode défensif* », et la réutilisation d'un outil accroît ses chances d'être sécurisé (y compris pour des solutions propriétaires, donc). Le directeur général souligne que c'est le phénomène de « *faux sentiment de sécurité* » qui est le plus néfaste pour les acteurs qui n'ouvrent pas leur code source, de « *sécurité par l'obscurité* », qui se privent en réalité d'une connaissance de leurs propres vulnérabilités. Dans cette perspective, l'ANSSI propose depuis quelques années un outillage générique : la méthode « Ebios Risk Management ». Elle se distingue par l'idée que cette analyse de risque doit renoncer à l'exhaustivité et surtout adopter une méthode tournée vers les métiers, c'est-à-dire qu'elle est conçue non pas pour les experts de la cybersécurité mais pour les gens qui ont des responsabilités métiers et qui sont à même d'exprimer eux-mêmes quels sont les risques, quelle est la gradation dans les risques, ce qui est acceptable ou l'est moins. L'ANSSI a labellisé plusieurs sociétés qui proposent de l'outillage logiciel pour guider les utilisateurs et leur permettre de se passer au maximum d'experts de l'analyse de risque.

Par ailleurs, plusieurs acteurs, en particulier les directeurs des systèmes d'information des institutions auditionnées, ont mis en avant le risque d'exposition de failles éventuelles au travers des codes sources, ouvrant la voie à une exploitation malveillante. Selon eux, certains codes sources, tels que les algorithmes de chiffrement, devraient être exclus de la démarche d'ouverture. Néanmoins, l'ANSSI confirme que ce risque lié à l'ouverture des codes sources est minime, et à relativiser du fait de l'expérience : les attaquants s'appuient très peu sur les codes sources. Du reste, l'ANSSI, elle-même a adopté un choix stratégique résolu en faveur du logiciel libre et de l'ouverture des codes sources.

La crainte d'une interprétation erronée ou à mauvais escient

Nombreuses sont les administrations à avoir souligné le risque d'interprétation des données brutes, nécessitant un traitement statistique pour être correctement comprises et réutilisées. C'est particulièrement le cas des services statistiques ministériels, qui insistent sur leur travail de construction et de qualification des données et sur l'importance de faire de la « pédagogie ». La DARES souligne ainsi que les statistiques concernant l'activité partielle pendant la crise de la Covid19 publiées sur *data.gouv.fr* correspondent aux données liées aux dépôts de demandes préalables d'activité partielle fournies par la direction métier (à savoir la DGEFP), et non aux demandes d'indemnisation effectives. Pour autant, la DARES est en mesure de contribuer et de documenter ces données sur *data.gouv.fr* et accompagner leur publication, pour jouer ce rôle de pédagogie. En effet, il est difficile de justifier la fermeture de ces données au motif qu'elles sont insuffisamment précises, car les données jugées « fiables » par la DARES (le nombre d'indemnisations effectives) ne sont disponibles que dans un délai d'un an (les entreprises disposant d'un délai d'un an pour réaliser cette demande d'indemnisation).

Sans remettre en cause le rôle essentiel de la statistique publique dans la qualification et le traitement de la donnée, cet exemple révèle combien l'émergence d'une demande de données publiées en temps réel est telle que les contraintes de la statistique publique ne permettent pas toujours d'y répondre. En effet, le retraitement statistique des données implique un délai de traitement peu compatible avec l'objectif de réalisation d'un tableau de bord en temps réel. Il montre la nécessité d'accompagner la publication des données d'un effort de pédagogie, de la part des acteurs les plus à même d'éclairer l'utilisation de la donnée, notamment sur des plateformes collaboratives comme *data.gouv.fr*. Ces acteurs devraient s'emparer de la démarche d'ouverture plutôt que de l'invalider *a priori*. Cet effort est celui qui permettrait de prévenir les interprétations fallacieuses par les différents utilisateurs (journalistes, citoyens, etc.).

Certaines administrations vont jusqu'à évoquer le risque d'un usage néfaste des données mises à disposition du public. Par exemple, le ministère de l'éducation nationale souligne le risque d'utilisation des données de résultats scolaires à des fins de classements des établissements sur le seul critère de réussite au baccalauréat. Le ministère considère que cela ne donne aucune information pertinente, et est pénalisant, voire stigmatisant, pour des lycées qui font des efforts et ont des élèves en grande difficulté scolaire à l'entrée : *« un lycée avec un taux de réussite élevé peut simplement avoir sélectionné au fur et à mesure de leur scolarité les meilleurs élèves et fait partir les autres ou même avoir accueillis des élèves qui avaient déjà un très bon niveau scolaire ».* C'est pour cette raison que les indicateurs de valeur ajoutée (IVAL) ont été élaborés. Du fait de la multiplicité des indicateurs, ils ne peuvent donner lieu à des classements bruts mais soit à des familles d'établissements soit à des classements en fonction de préférences personnelles. **La mission note toutefois que la presse reconstitue chaque année des classements sur les établissements scolaires, phénomène qui préexiste largement à l'*open data*, et ce, malgré l'absence de publication des données brutes par le ministère de l'éducation nationale.**

D'autres acteurs ont évoqué un risque de réputation en mettant au jour l'imperfection des données et des codes sources publics, et redoutent ainsi une certaine défiance de la part de la société. C'est le cas des directions métiers qui disposent souvent de données brutes, non travaillées. La crise de la Covid19 a révélé les difficultés pour établir des statistiques en temps réel de l'épidémie, les conflits d'interprétation des courbes, la fragilité des hypothèses des modèles épidémiologiques. Malgré les critiques, la publication des données a aussi contribué à acculturer les citoyens aux problématiques des données et des modélisations en révélant que toute donnée est une construction et qu'elle doit être expliquée.

Les administrations sont légitimes à souhaiter que la publication de leurs données et codes sources s'accompagnent d'une communication fondée sur leur expertise, par exemple au travers d'analyses sans empêcher que d'autres acteurs puissent se saisir de ces matériaux pour réaliser des études. En effet, la transparence est nécessaire à la confiance ; il apparaît préférable d'associer et d'acculturer les citoyens aux problématiques de la donnée et des codes sources. C'est pourquoi la mission recommande qu'un débat sur les enjeux du numérique soit organisé. Il doit permettre d'établir les prérequis indispensables pour l'ouverture et le partage des données et des codes sources.

L'ouverture de données remet en cause des situations de rente économique

L'ouverture des données obéit à un principe de réutilisation libre et gratuite, et justifie l'interdiction de pratiquer des redevances depuis 2016 (cf. partie 2). Néanmoins, plusieurs données sont encore aujourd'hui mises à disposition à titre payant, soit que ces exceptions soient prévues par la loi ou que les opérateurs ne se conforment pas totalement avec les dispositions, soit que ces données sont produites par des acteurs privés dans le cadre d'un service public, sans que ce service public ne s'étende aujourd'hui explicitement à la mise à disposition de ces données.

L'ouverture des données, en remettant en cause des rentes économiques, conduit à une perte de recettes, mais cette perte ne fait pas obstacle à l'ouverture : la loi du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public, transposant la directive PSI2, a ainsi précisément limité les redevances aux cas où les administrations « *sont tenues de couvrir par des recettes propres une part substantielle des coûts liés à l'accomplissement de leur mission de service public* »[33].

[33] Article L. 324-1 du CRPA.

À la fin de l'année 2020, les trois opérateurs concernés tirent encore plus d'un million d'euros de recettes de ces redevances :

- si les redevances de l'IGN sont aujourd'hui en passe d'être définitivement supprimées[34], elles ont représenté une perte de recettes non négligeable sur la période récente : passées de 10,2 M€ en 2012 à 3,2M € en 2019 (notamment par la réduction des redevances de la base de données des adresses ou des scans), elles représentent encore aujourd'hui 1,1 M€ de recettes ;
- Météo-France perçoit encore des recettes estimées à 2 M€, après avoir déjà vu ses recettes « grand public » passer de 17,8 M€ en 2012 à 9 M€ en 2017, en raison de la baisse de la vente de données brutes. Comme le remarque la Cour des comptes[35], l'enjeu de cette diminution est moins importante que pour l'IGN. L'enjeu majeur perçu par Météo-France aujourd'hui est plutôt la prise en charge du coût de la mise à disposition des données[36] ;
- le service hydrographique et océanographique de la Marine (SHOM) tire 11 % de ses recettes de la vente de données, pour un montant d'environ 7 M€ par an[37]. Ces redevances sont perçues dans le cadre d'un mécanisme licence spécifique, par lequel le SHOM met à disposition certaines données sous licence CC-BY-SA 4.0 afin d'imposer aux réutilisateurs la condition de rediffusion à l'identique. Pour les utilisateurs qui ne souhaitent pas respecter cette condition, une autre licence, payante, leur est proposée[38]. Le SHOM souhaite ainsi maintenir un paiement sur les données pour lesquelles il possède des droits de propriété intellectuelle (cartes marines par exemple)[39].

Dans ces cas, le modèle économique fondé sur la donnée est constitutif d'une rente de situation, comme cela a été identifié dès 2013[40]. Cependant, ces acteurs ne sont pas parvenus à prendre en compte cette évolution dans leur modèle. Généralement, il semble difficile d'envisager le développement de nouvelles ressources propres à partir de ces données. Dans son référé du 11 décembre 2018, la Cour des Comptes explique cette réduction des ventes de données au secteur privé par plusieurs facteurs : les nouvelles concurrences des plateformes du numérique, l'apparition de données gratuites ou libre sur le marché, la demande en données de qualité à laquelle les opérateurs ont des difficultés à répondre, et la fin du cycle d'acquisition des données pour les grandes entreprises.

Un exemple de données aux retombées économiques importantes comme la base DVF l'illustre bien aussi : la donnée brute ne pouvant être mise à disposition à titre payant, la seule marge restant au producteur pour développer une recette à partir de la donnée serait celle de la production d'un service lié à cette donnée ; or, les acteurs publics producteurs de la donnée ne semblent pas aujourd'hui en mesure d'investir et de fournir ce service, dès lors qu'ils n'ont parfois pas même suffisamment de moyens pour remplir leurs obligations de mise à disposition gratuite.

[34] Le conseil d'administration de l'IGN du 4 décembre 2020 a acté l'évolution de la tarification et des licences, prévoyant que les données éditées par l'IGN, sans droit de tiers, disponibles en téléchargement ou accessibles par flux, le seront en licence ouverte, selon les termes de la licence ouverte Etalab 2.0 ou ses versions ultérieures.

[35] Référé de la Cour des Comptes, 11 décembre 2018, « La valorisation des données de l'IGN, de Météo-France et du Cerema ».

[36] Selon Météo-France, ces coûts représentent aujourd'hui 100 000 € par an environ pour l'hébergement, la mise à disposition et la diffusion par le portail de données publiques. Une augmentation éventuelle n'a pas été chiffrée.

[37] Informations communiquées à la mission par le SHOM.

[38] Cette pratique s'inspire d'une pratique déjà ancienne chez certains éditeurs de logiciels libres, qui publient le code source sous une licence libre et sous une licence plus restrictive pour laquelle certains rétuilisateurs sont prêt à payer.

[39] Contrat d'objectifs et de performance 2021-2024.

[40] Rapport au Premier ministre de M. A. Trojette.

Le risque d'une « captation de la valeur » de la donnée

Plusieurs acteurs auditionnés par la mission ont mentionné la crainte d'une réutilisation lucrative des données et codes sources publiés gratuitement, et d'une captation de la valeur par des acteurs privés. La perception de ce risque est généralement associée à une appréhension vis-à-vis de grandes plateformes numériques, surtout étrangères, et à leur capacité à valoriser la donnée, par rapport à des acteurs nationaux, dont la légitimité est beaucoup plus rarement questionnée.

Cette crainte d'une privatisation de la valeur ajoutée est d'abord évoquée pour la publication des codes sources produits par les agents publics en *open source*. Comme évoqué dans le paragraphe 2.1, l'ouverture de ces codes favorise la mutualisation des ressources entre les acteurs publics ou privés, comme le prouve le succès de la bibliothèque Scikit-learn. Néanmoins, ce type de démarche peut nécessiter une maintenance ainsi qu'une grande réactivité lorsque des problèmes de sécurité sont identifiés, notamment en cas de déploiement. Aussi, pour les cas de codes les plus structurants (une minorité seulement des codes sources ouverts aujourd'hui), ce type de projet doit s'appuyer sur une communauté capable de remplir ce rôle : le dynamisme et la force d'une communauté se mesurent à travers les contributions régulières de ses membres. Les exemples de QGIS, Géotrek et Prodige illustrent bien cette problématique (cf. cas d'usage sur trois logiciels libres en matière d'information géographique).

Ensuite, plusieurs acteurs auditionnés, issus aussi bien de la sphère publique que privée, soulignent le risque de captation des réutilisations des données et des codes sources par les plateformes du numérique, qui disposent de moyens sans commune mesure, en comparaison d'acteurs européens de taille plus modestes pour développer des projets innovants. En effet, si la donnée n'est pas un bien « rival » au sens économique (son utilisation par un acteur ne réduit pas l'utilisation potentielle par un nouvel acteur), il existe manifestement des rentes d'exploitation de la donnée, les acteurs occupant une position importante sur un marché ayant plus d'aisance à exploiter de nouvelles données et à proposer des services.

Face à cette crainte, il ne semble pas pertinent et réaliste de rechercher un accès restreint ou limité aux données, qui ne serait par exemple qu'au bénéfice d'acteurs nationaux ou européens. Une telle « priorité » n'est de fait pas nécessaire pour permettre la réutilisation des données par des acteurs nationaux et de taille plus modeste, comme cela a été confirmé à la mission par plusieurs acteurs. Ensuite, cette restriction serait inefficace pour espérer éviter une « captation » de la valeur par les plateformes du numérique, qui disposent déjà d'un accès à des données personnelles en volume significatif, et dont le modèle économique ne peut réellement être interrogé que par des outils de régulation adéquats.

L'option d'une intervention en aval, qui consisterait à contrôler la réutilisation des données faite par ces grandes plateformes, dans le but de prévenir des usages illégaux, est plus réaliste, mais ne peut pas impliquer de sélectionner les réutilisations de données en fonction de leur finalité. Donner la faculté à un producteur de contrôler *a priori* la réutilisation de sa donnée est par essence contraire à la démarche d'*open data*, qui ne préjuge pas des usages et permet ainsi de libérer complètement le potentiel social et économique de la donnée. En revanche, le recours à des API pour la mise à disposition des données permet de connaître les réutilisations qui en sont faites et de détecter plus rapidement les éventuels usages illégaux qui pourraient en être faits. Si l'usage est illégal, le contrôle *a posteriori* permettra de garantir le respect du droit. C'est aussi le sens de l'approche retenue par la Commission européenne dans le cadre du *Digital Services Act*, en cours de rédaction à date de rédaction du présent rapport, mais qui doit permettre une meilleure régulation de l'utilisation faite par les plateformes du numérique des données personnelles.

La fermeture de nos données et de nos codes crée pourtant une inégalité pour les acteurs nationaux et une perte de compétitivité pour notre pays

Pourtant, la fermeture des données et des codes publics crée de fait une vraie inégalité pour les acteurs nationaux et une perte de compétitivité pour notre pays : l'accès à certaines données publiques étrangères, par exemple en matière de santé, est plus aisé que l'accès aux données françaises pour les entreprises ou chercheurs français. À titre d'exemple, les chercheurs de l'Inserm privilégient parfois l'utilisation des données de pharmacovigilance de la FDA, faute de pouvoir disposer facilement des données de l'ANSM (cf. encadré).

La législation française ne permet pas l'*open data* de la pharmacovigilance

L'Agence nationale de la sécurité du médicament (ANSM) a souhaité publier la base de données de pharmacovigilance, après avoir pris soin d'appliquer un processus d'anonymisation (y compris pour ce qui concerne la maille géographique retenue pour la présentation des résultats). **Cette publication permettrait de renforcer la transparence et la confiance des patients et des professionnels de santé.**

En France, cette publication est aujourd'hui impossible, la loi pour une République numérique de 2016 ne prévoyant la publication de documents comportant des données personnelles qu'au respect d'une des trois conditions suivantes : i) l'existence de dispositions législatives expresses, ii) l'accord des personnes concernées ou bien iii) la mise en œuvre d'un traitement permettant de rendre impossible l'identification de ces personnes (application d'un processus anonymisation). Dans le cas de la pharmacovigilance, l'anonymisation opérée par l'ANSM permet notamment une réidentification par le patient lui-même (risque qualifié de très probable). Sauf à anonymiser au point de ne permettre aucune réidentification (mais avec le risque de perdre l'intérêt de la donnée), seule une modification de la législation permettrait aujourd'hui à l'agence de mettre à disposition ses bases de données en matière de vigilance, ou bien le fait de prévoir le recueil du consentement des patients (mais grevant d'autant l'intérêt de la transparence si un grand nombre de données sont retirées de ce fait).

Cette démarche existe pourtant déjà au niveau européen, où des données individuelles anonymisées de pharmacovigilance sont publiées par l'agence européenne du médicament (EMA) dans sa base EUDRA Vigilance[41]. Les données que l'ANSM prévoyait de publier sont de même nature.

Cet intérêt scientifique est majeur aujourd'hui, et ne peut être comblé qu'en ayant recours à des données de pays étrangers. L'Inserm a indiqué à la mission que « *beaucoup d'unités ont cherché à utiliser les données de l'ANSM (médicament), pour réaliser des études épidémiologiques sur les effets secondaires des médicaments. Celles-ci n'étant pas couvertes par le référentiel SNDS, les chercheurs se sont tournés vers la base de données américaine de la FDA qui est de moindre qualité mais qui est directement accessible.* »

Afin de pallier cette difficulté, la CNIL recommande une évolution du cadre juridique afin de donner un caractère contraignant au principe de transparence des données de pharmacovigilance. L'ANSM soutient cette évolution du droit, considérant en effet que la publication des données de pharmacovigilance présente un risque faible et proportionné par rapport l'intérêt général de la publication de ces données, pour permettre l'analyse des données et l'information des patients sur les risques liés à l'utilisation des médicaments.

De la même façon, une société financée par Bpifrance rencontre aujourd'hui des difficultés pour accéder à des données des établissements de santé, au point de devoir concéder des pertes de marché en France, au bénéfice de sociétés chinoises dont les données utilisées sont de moindre qualité, mais qui y ont accès dans leur pays (cf. encadré).

[41] http://www.adrreports.eu/fr/

Dans la santé, les obstacles à la recherche favorisent le recours à des services chinois

La start-up TheraPanacea engage depuis 2017 des travaux de recherche pour améliorer le traitement des cancers par radiothérapie, grâce à l'intelligence artificielle, en permettant de fiabiliser le protocole de traitement, par exemple le contour des organes et des tissus à épargner, et en divisant par vingt le temps de préparation d'une radiothérapie. La start-up collabore avec trois institutions publiques, deux centres de lutte contre le cancer et un réseau d'hôpitaux publics.

Les recherches de la start-up sont aujourd'hui ralenties par des obstacles administratifs qui limitent l'accès aux données, à la fois par le manque d'harmonisation des règles d'accès aux données entre acteurs de la santé et par la longueur des procédures de la CNIL. Alors que le projet a été validé par Bpifrance en juillet 2020, la start-up n'a pas pu aboutir à un accord général avec ses partenaires sur le partage de données, et doit engager en décembre 2020 des négociations individuelles avec chaque établissement partenaire du projet, et n'aura ainsi probablement pas accès aux données avant l'été 2021, dans la meilleure hypothèse. Il n'existe en effet aujourd'hui pas d'harmonisation des règles de propriété intellectuelle qui prévalent dans chaque établissement, et la start-up doit ainsi négocier et gérer des règles de propriété intellectuelle très hétérogènes d'un établissement à l'autre, avant de pouvoir traiter les données par le biais du *Health Data Hub*.

Outre la perte de chances pour l'amélioration du traitement des patients, cette situation conduit à favoriser la concurrence de sociétés chinoises, qui accèdent aux marchés des hôpitaux français, grâce aux données, pourtant de moindre qualité, qu'elles exploitent en Chine. La société Infervision propose par exemple un produit de diagnostic de la covid19, dont la qualité scientifique est contestée par plusieurs acteurs, mais qui a été financée au niveau européen ; tandis que l'accès aux données de haute qualité en France (10 000 scanners de plus de 20 institutions) qui permettraient de développer ces produits, en garantissant notre autonomie stratégique, n'a pas été possible pour la start-up avant la fin de l'année 2020.

Dans la recherche, la crainte d'un risque réputationnel et de « mal faire »

Aujourd'hui, la culture de la science ouverte dans l'enseignement supérieur et la recherche n'atteint pas un niveau suffisant. Il subsiste de nombreux freins. **L'enjeu le plus important est sans doute la reconnaissance de la démarche de science ouverte.** Dans de nombreuses disciplines, la production de code n'est pas valorisée, au sens où elle n'apporte rien aux chercheurs pour leur carrière. Les chercheurs ont trop souvent le sentiment de perdre du temps s'il faut travailler sur leurs codes uniquement pour le rendre accessibles aux autres. Les mécanismes de reconnaissance et de valorisation passent par l'acceptation, dans chaque discipline, de l'importance des codes sources comme c'est déjà le cas en informatiques ou en mathématiques appliquées.

Un autre frein est la peur de « mal faire » et, parfois, le manque de compétence. Écrire un code pour ses propres besoins de recherche est sans comparaison avec le fait d'écrire un code pour, potentiellement, les besoins des autres chercheurs. Il y a la crainte d'être jugé sur la qualité de son code et, ainsi de s'exposer. Cela pose la question de la formation des chercheurs. Une autre approche est possible. Les chercheurs qui souhaitent développer des codes de qualité, pour en faire profiter la communauté par la suite, devraient pouvoir s'appuyer sur des départements spécifiques dont la mission est de venir en soutien à ce type de démarche. C'est déjà le cas dans de grandes universités à l'instar des « bureaux européens » pour le montage de projet européen. Faute de quoi, les chercheurs ont le plus souvent recours à des post-doctorants dont, par définition, la présence est limitée dans le temps et qui ne seront plus disponibles pour assurer la pérennité du code quand cela est nécessaire. Cette approche pose la question du recrutement pérenne d'ingénieurs de recherche avec l'ouverture de postes dédiés à la montée en qualité des codes sources issus des laboratoires.

Par ailleurs, dans le cas de recherche partenariale, par exemple pour les thèses CIFRE, il est fréquent que l'industriel exige la confidentialité des codes sources (voire des données) et il est parfois difficile aux petits établissements ou laboratoires d'opposer une démarche de science ouverte, compte tenu des enjeux de financement.

En outre, dans le domaine de la recherche, certains savoirs font l'objet d'une protection spécifique, dans le cadre du dispositif de la protection du potentiel scientifique et technique de la nation (PPST) piloté par le Secrétariat général de la défense et de la sécurité nationale (SGDSN). Ce dispositif a pour but de protéger les savoirs et savoir-faire stratégiques et les technologies sensibles, pour prémunir les services de recherche contre des tentatives de captation d'informations (espionnage industriel et scientifique). La PPST a vocation à s'appliquer aux savoirs identifiés et reconnus comme stratégiques, et n'est donc pas incompatible avec une démarche de science ouverte, au contraire : elle incite les services de recherche à s'interroger, dès le départ et sans attendre une sollicitation externe, sur le caractère stratégique ou non pour l'État du savoir et, par voie de conséquence, à ne pas protéger sans justification des savoirs qui, eux, ne le sont pas, et peuvent être partagés.

Enfin, le risque réputationnel pour les chercheurs et la difficulté d'accéder aux données et aux codes sources a été concrètement mesuré par la mission, lorsque plusieurs chercheurs interrogés ont fait part de leur souhait d'anonymiser leur situation dans ce rapport, afin d'éviter une mise en cause de leur travail et de leurs demandes d'accès aux données.

Finalement, la mission formule deux recommandations transversales pour que la société civile puisse se saisir des enjeux de la donnée et des codes sources, dans une démarche de transparence et de responsabilité de l'action publique dans ce domaine. La consultation publique conduite par la mission, dont la synthèse est présentée en annexe, a révélé un besoin important de restaurer la confiance dans l'action numérique de la puissance publique, auquel il est nécessaire de répondre.

Recommandation : Initier un débat public sur les conditions de la confiance dans le numérique, permettant de définir les principes fondamentaux de sécurité et de transparence qui doivent s'imposer à la puissance publique

Recommandation : Conduire une évaluation de l'impact économique, social et scientifique de l'ouverture et du partage des données et des codes sources

CAS D'USAGE – Données et modèles épidémiologiques dans le cadre de la gestion de crise de la Covid19

Dans le cadre de la crise sanitaire liée à la Covid19, la publication des données et des modèles épidémiologiques constitue un enjeu majeur en période de gestion de crise :

- d'une part, **pour éclairer la décision publique** : par exemple, les mesures de confinement, le recours aux équipements de protection et l'application des gestes barrière , les adaptations locales nécessaires (par exemple, fermeture des bars et restaurants dans certaines métropoles), le pilotage de la politique de dépistage ;
- d'autre part, **pour alimenter le débat public, dans un souci de transparence vis-à-vis des citoyens**, à un moment où il est nécessaire de gagner leur confiance et de susciter l'adhésion aux décisions prises. La publication des indicateurs permet également de gagner un temps précieux en évitant aux institutions de répondre individuellement à des demandes identiques.

Les chercheurs ont aussi un rôle à jouer pour **améliorer la connaissance des impacts de la Covid19** à la fois sur la santé des citoyens, mais aussi sur les inégalités économiques et sociales : l'accès aux données à une maille plus fine que l'information disponible publiquement est parfois nécessaire.

Des initiatives citoyennes accélératrices : OpenCOVID et CovidTracker

Début mars 2020, une initiative citoyenne OpenCOVID collecte et agrège les données communiquées par les ministères, les ARS, les préfectures, etc, témoignant ainsi de la pression des citoyens pour obtenir des informations précises sur l'épidémie. Lorsque l'initiative openCOVID a débuté, il était uniquement possible de suivre la propagation du virus à partir des bilans hebdomadaires de Santé publique France et des allocutions officielles mais aucun fichier consolidé n'était publié par l'État. L'ensemble des travaux d'OpenCOVID peuvent être consultés sur la page Github de l'organisation. Les échanges entre les membres de l'initiative se déroulaient sur l'outil Slack.

L'initiative OpenCOVID a constitué un levier essentiel pour l'ouverture des données. Si des échanges étaient déjà en cours pour obtenir l'ouverture des données, la découverte de *veille-coronavirus.fr* au sein de l'administration a accéléré les discussions et a rendu indispensable la communication des chiffres officiels exhaustifs sur la situation sanitaire. Le développement d'un outil de visualisation a également été plébiscité pour faciliter l'accès et la compréhension des données. Depuis le 26 mars, le tableau de bord de suivi de l'épidémie (veille-coronavirus.fr) est disponible sur gouvernement.fr[42] et est actualisé quotidiennement. En parallèle, Santé publique France met également à disposition du public 79 indicateurs liés à la Covid19 au travers de son portail Géodes (géo-données en santé publique) qui peuvent être visualisés sous forme de cartes, tableurs, figures. Santé publique France et Etalab publient depuis fin mars les indicateurs sur *data.gouv.fr* dans un souci de transparence.

[42] https://dashboard.covid19.data.gouv.fr/vue-d-ensemble?location=FRA

L'intégration du tableau de bord sur *gouvernement.fr* a été largement facilitée par les liens étroits entre les membres de OpenCOVID et Etalab. Les liens historiques entre Etalab et la société civile ont été un élément clé pour accélérer l'ouverture des données et fournir des outils de visualisation.

La visualisation des données est également proposée par des acteurs privés comme CovidTracker, projet initié par un *data scientist* indépendant, Guillaume Rozier, dont le tableau de bord, alimenté notamment par les données en *open data* de Santé publique France, rencontre un grand succès en raison de son accessibilité et de son ergonomie. Cette initiative participe également à l'émulation entre acteurs publics et acteurs privés : ainsi, l'ergonomie d'outils tels que l'application TousAntiCovid a pu être améliorée notamment grâce à ce type d'échanges.

L'acculturation des citoyens aux problématiques de collecte des données

La qualité des données collectées et publiées a soulevé de nombreuses interrogations de la part des citoyens. Néanmoins, la crise a permis de révéler l'intérêt des citoyens porté à ce type de données, et a permis également une acculturation à cette problématique.

De nombreux utilisateurs ont souligné des incohérences entre l'agrégation des données granulaires publiées en *open data* et les indicateurs nationaux communiqués notamment sur le tableau de bord. Par exemple, la somme des patients hospitalisés dans chaque département est inférieure au nombre total de patients hospitalisés, écart qui s'explique par le fait que le nombre total de patients hospitalisés inclut les patients hospitalisés à l'étranger et les patients qui ne peuvent pas renseigner d'adresse de domiciliation.

Une prise de conscience sur la nécessité d'investir dans des infrastructures pour la remontée des données

Par ailleurs, la crise sanitaire a accéléré les investissements nécessaires dans les infrastructures pour permettre la mise à disposition du public des données épidémiologiques sur la Covid19. La publication des données a aussi mis en évidence les difficultés liées à leur collecte, ces données étant produites par de multiples acteurs et s'appuyant sur des systèmes d'informations hétérogènes et difficilement interopérables. En particulier, certains laboratoires (données SI-DEP) ou hôpitaux (données SI-VIC) peuvent communiquer des données avec du retard, ce qui conduit à fausser le nombre de nouveaux cas confirmés ou de nouveaux patients dans les chiffres quotidiens.

Santé publique France a été un acteur majeur dans la collecte des données de santé pour mieux comprendre l'épidémie au travers de deux bases de données SI-VIC, SI-DEP. Ces deux bases sont notamment à l'origine des bilans quotidiens et hebdomadaires sur les nouveaux cas, décès et taux d'occupation des lits à l'hôpital :

- le **système d'information pour le suivi des victimes d'attentats et de situations sanitaires exceptionnelles créé en 2016 suite aux attentats de Paris (SI-VIC**) est un outil initialement conçu pour répondre à un besoin de suivi des victimes d'attentats ou de situations sanitaires exceptionnelles de moyenne envergure (entre 5 et 8 000 dossiers patients saisis par évènement) : il a dû être adapté pour identifier les patients atteints par la Covid19 et recenser davantage de patients dans la perspective d'un déploiement dorénavant effectif dans la quasi-totalité des établissements hospitaliers de France ;
- le **système d'information national de suivi du dépistage (SI-DEP)** fournit les données épidémiologiques, permet de diffuser des conseils aux personnes atteintes de la Covid19 et sert de déclencheur aux enquêtes sanitaires réalisées par l'ARS et l'Assurance maladie. Ce SI a été développé dans un délai inférieur à un mois alors même que le ministère ne disposait pas d'une connaissance fine des milliers de laboratoires en France et qu'un projet de SI par Santé Publique France s'était heurté à de multiples obstacles depuis huit ans. Or, 90 % des laboratoires étaient connectés à SI-DEP au moment du déconfinement, illustrant la réussite de ce projet mené sous l'égide de la délégation ministérielle du numérique en santé (DNS) qui a réuni l'ensemble des parties prenantes et s'est appuyée sur la maîtrise d'œuvre de l'AP-HP. Cette réussite tient à l'exploitation de tous les leviers : implication du niveau politique, avec un fort engagement du secrétaire d'État au numérique aux côtés du ministre des Solidarités et de la Santé ; prise en charge financière des évolutions de logiciels rendues

nécessaires ; réglementation conditionnant le remboursement du test au laboratoire à la bonne saisie dans SIDEP.

Quant aux établissements et services médico-sociaux, aucun outil n'existait véritablement, même au niveau régional, pour suivre les données épidémiologiques : cette situation a conduit les ARS à mener des enquêtes *ad hoc*, par exemple pour recueillir les données auprès des EHPAD. Les données étaient saisies par les acteurs de terrain, disposant de peu de temps pour le faire, conduisant à des incohérences et nécessitant des relances téléphoniques. Face à l'impossibilité d'exploiter l'outil de signalement des cas, Santé publique France a construit un système d'information Voozanoo opérationnel à la fin mars pour les établissements sociaux et médico-sociaux, en premier lieu les EHPAD. Sa mise en place a permis de limiter les remontées concurrentes et parallèles demandées par les différentes autorités compétentes (ARS, préfet, conseil départemental, en particulier), même si celles-ci n'ont pas cessé dans un certain nombre de territoires. De nombreuses difficultés d'utilisation ont été cependant relevées par les utilisateurs et doivent encore faire l'objet d'améliorations. Ces données ne sont pas publiées en *open data*, seul le nombre de cas confirmés et le nombre de décès national sont communiqués deux fois par semaine.

Par ailleurs, l'enrichissement de l'information des collectivités territoriales sur l'évolution des données épidémiologiques concernant leur territoire constitue une demande forte afin d'adapter la réponse locale à la crise sanitaire. À titre d'exemple, disposer de données épidémiologiques à une maille géographique plus fine que la commune, ou de données ventilées selon les caractéristiques sociodémographiques de la population permettrait d'apporter une réponse plus ciblée en matière de prévention, d'implantation des barnums, etc.

Afin de répondre aux demandes de disposer d'indicateurs à une maille infradépartementale, Santé publique France a commencé à mettre des données à disposition sur Géodes, au niveau des communes, concernant par exemple le taux d'incidence.

Plus globalement, cela pose la question de l'anonymisation des données et du risque de réidentification : sans aller jusqu'à la mise à disposition du public de ces indicateurs, la question du partage des données entre Santé publique France et les collectivités territoriales pourrait être envisagée dans un souci d'efficience de l'action publique. Ce partage de données pourrait se faire dans le cadre du *Health Data Hub*.

Les jeux de données épidémiologiques

Les bases de données liées au suivi épidémiologique de la Covid19 ne font pas l'objet d'une publication en *open data*, étant donné qu'il s'agit de données à caractère personnel. Santé publique France a rendu public des indicateurs et statistiques notamment sur le portail Géodes.

Néanmoins, les données complètes et personnelles contenues dans les bases ne peuvent être accessibles que par le biais du *Health Data Hub* ou auprès des producteurs de données, sous certaines conditions. À la suite d'un arrêté du 21 avril 2020, le *Health Data Hub* a été autorisé à recevoir des données pseudonymisées (bases OSCOUR, SI-DEP, SI-VIC, SNDS Fast Track entre autres) en lien avec l'épidémie de la Covid19. La mise en production accélérée du HDH dans le contexte de la crise sanitaire répond à deux objectifs : d'une part, l'impérieuse nécessité d'assurer la centralisation des données de santé, afin d'en faciliter l'accès aux équipes de recherche sur des sujets identifiés comme étant prioritaires ; d'autre part, la construction, en partenariat avec les acteurs à la source de la collecte des données, des grands jeux de données dont il est apparu que certains font défaut dans la lutte contre l'épidémie de la Covid19.

Néanmoins, la mission a pu recueillir un témoignage d'une chercheuse en sciences sociales, révélateur des difficultés rencontrées dans la réutilisation des données de santé, y compris lorsqu'il s'agit de participer à la construction de connaissances dans le cadre de la Covid19. Celle-ci a souhaité avoir accès aux données du SI-VIC et SI-DEP, avec l'objectif de décrire sur une zone géographique la plus fine possible, le profil social des personnes infectées par la maladie.

En matière de santé, le projet de recherche doit poursuivre une finalité d'intérêt public et doit être validé par le CESREES (Comité Éthique et Scientifique pour les Recherches, les Études et les Évaluations dans le domaine de la Santé) et autorisé par la CNIL (sauf cas particuliers tels que les organismes disposant d'un accès permanent). Le projet d'étude doit être soumis au *Health Data Hub* qui joue le rôle de guichet unique. Le dossier est ensuite évalué par le CESREES qui dispose d'un mois, une fois renouvelable pour se prononcer sur la complétude du dossier. Après validation par le CESREES, le dossier est transmis à la CNIL qui dispose de deux mois renouvelables pour rendre sa décision. Cette procédure doit être effectuée par toute personne souhaitant accéder à des données de santé personnelles, que le projet utilise ou non la plateforme du *Health Data Hub*.

Constituer les dossiers CESREES et CNIL peut prendre plusieurs mois pour des chercheurs. En effet les délégués à la protection des données (DPD) des instituts de recherche sont en trop faible nombre par rapport aux besoins, et les chefs de services informatiques sont souvent mobilisés sur de nombreux sujets autres que l'accès aux données. La demande d'accès n'est pas allégée pour les chercheurs ayant déjà eu une autorisation sur une base de données et demandant à nouveau l'accès à la même base.

Ainsi, si la chercheuse affirme avoir été soutenue positivement par le HDH, elle n'a toujours pas accès aux données SI-VIC alors que la première demande a été faite en avril. En septembre, le dossier était en cours de constitution auprès de la CNIL. Ainsi, elle regrette le caractère « sclérosant » de la procédure : bien que faisant partie d'un organisme de recherche, la procédure demande aux chercheurs de s'engager sur des processus sur lesquels ils n'ont que peu de maîtrise (par exemple, les pare-feu de l'organisme de recherche au sein duquel ils travaillent) et nécessitant de recueillir des signatures auprès du délégué à la protection des données et du directeur des systèmes d'information, sans que cela ne puisse être généralisé pour l'ensemble des demandes. Ce travail de constitution de dossier consomme un temps précieux sur l'activité de recherche.

Au-delà de la situation de cette chercheuse, il est à noter que depuis l'ouverture de la plateforme du *Health Data Hub*, en date de novembre 2020, le *Health Data Hub* a reçu 52 demandes relatives à la Covid19, 19 projets ont été autorisés par le CESREES et la CNIL. Parmi ces 19 projets, deux projets vont utiliser la plateforme du HDH pour mener leurs travaux de recherche. Le délai d'instruction de la CNIL est en moyenne de 20 jours ouvrés pour les projets n'utilisant pas la plateforme du HDH (avec des réponses allant de 1 jour ouvré à 88 jours ouvrés, délai médian de 9 jours ouvrés). Pour les projets utilisant la plateforme du HDH, les délais d'autorisation de la CNIL sont de 46 jours ouvrés (45 et 47 jours ouvrés respectivement). La DREES et l'INSEE, en tant qu'organismes ayant un accès permanent, utilisent la plateforme depuis avril et novembre 2020 respectivement.

Les modèles épidémiologiques

De nombreux modèles épidémiologiques ont été mobilisés lors de la première vague de l'épidémie : les modèles de l'AP-HP, de l'Institut Pasteur, de l'Inserm, de l'EHESP, des acteurs privés, entre autres. Certains d'entre eux, tels que ceux permettant d'établir des projections de l'évolution de l'épidémie, réalisés par l'Institut Pasteur, ont été utilisés par le Conseil scientifique pour éclairer les décisions du gouvernement. La base de données SI-VIC a servi entre autres à alimenter les modélisations réalisées par l'Institut Pasteur pour anticiper l'évolution des situations dans chaque région au regard de l'épidémie.

Lors de la première vague de la crise, le partage de ces modèles entre chercheurs et institutions n'a pas fait l'objet d'une coordination et d'une stratégie suffisamment claires. En effet, plusieurs instituts ont proposé des modèles, sans les rendre accessibles à d'autres instances, ni *a fortiori* au grand public. Deux raisons principales ont manifestement présidé à ce manque d'ouverture, d'après les informations recueillies par la mission.

En premier lieu, le partage des modèles entre chercheurs a pu être freiné par les obstacles culturels identifiés dans le domaine de la recherche plus généralement : d'une part, du fait des faibles incitations individuelles pour les chercheurs à partager leurs modèles, étant donné qu'il n'est pas obligatoire de publier des données et des codes ayant servi à un article de recherche ; d'autre part, du fait du risque réputationnel parfois perçu par certains chercheurs qui hésitent à publier une version non aboutie de leurs travaux, notamment de leurs codes.

Par ailleurs, pour ce qui est de la France, aucun institut n'a pris la responsabilité d'ouvrir ses modèles comme ont pu le faire Etalab ou SpF s'agissant des données. À l'étranger, l'Imperial College a publié son modèle CovidSim le 22 avril 2020 sur Github, *via* une collaboration avec Microsoft.

Si le débat a eu lieu au niveau du conseil scientifique concernant l'opportunité d'ouvrir complètement les modèles épidémiologiques développés par l'Institut Pasteur utilisés pour établir les projections d'évolution de l'épidémie, cette publication a finalement été écartée, en raison du risque de mauvaise interprétation des résultats des modèles. Pour autant, les membres du conseil scientifique ont communiqué sur la construction du modèle et sur ses hypothèses, dont certaines sont, comme pour tout modèle épidémiologique, fragiles. La publication des résultats des modèles, dont la dernière est intervenue le 25 septembre 2020, mentionne ainsi en avertissement : « *ces scénarios sont faits sur la base de données incomplètes et d'hypothèses incertaines. La prorogation du virus SARS-CoV-2 reste difficile à anticiper, et la dynamique de l'épidémie peut changer très rapidement.* »[43]

S'il n'y a donc pas eu de publication des modèles de l'Institut Pasteur, le partage des modèles a eu lieu entre différents acteurs impliqués dans la gestion de la crise, en particulier avec les ARS, ce qui a permis de nourrir des échanges au sein d'un nombre très limité d'acteurs. Néanmoins, elle est intervenue assez tardivement durant la crise, amenant certaines ARS et l'AP-HP à développer leurs propres modèles.

La mission regrette que ces modèles ne soient pas intégralement publiés, dans une démarche de transparence (la structure du modèle étant déjà connue) et d'amélioration de la connaissance scientifique. La confiance dans l'action publique sera davantage confortée par une démarche de transparence et de pédagogie, expliquant les limites de la fiabilité de tout modèle épidémiologique, que par le maintien d'une opacité qui ne protège en réalité aucun secret.

Du reste, il est étonnant de constater que, malgré l'ampleur mondiale de cette pandémie, il n'y ait pas eu davantage de partage des modélisations au niveau international, même si des tentatives ont été initiées au niveau européen.

Il faut publier les modèles de l'Institut Pasteur dans une approche pédagogique et collaborative

En premier lieu, cette publication devrait s'accompagner d'une communication forte expliquant les biais et hypothèses du modèle, afin d'éviter une mauvaise compréhension du public. Rappelons, à cet égard, que la mise à disposition du public des données épidémiologiques de la Covid19 a participé à l'acculturation des citoyens, notamment sur les imperfections des données.

Par ailleurs, la publication de ces modèles gagnerait à intervenir dans le même temps que la publication d'autres modèles épidémiologiques, de sorte à ne pas cristalliser les débats sur une seule modélisation et à participer à sa robustesse au travers des débats qu'il nourrit.

[43] Modélisation Mathématique des Maladies Infectieuses, Institut Pasteur (A. Andronico, C. Tran Kiem, J. Paireau, P. Bosetti, S. Cauchemez), Santé Publique France, « Evolution possible du nombre de patients COVID-19 dans les services hospitaliers en France métropolitaine », 25 septembre 2020.

CAS D'USAGE – Les statistiques de la délinquance

Le ministère de l'Intérieur a réalisé un effort de transparence ces dernières années et commence à structurer son service statistique ministériel, mais tous les services ne comprennent pas encore le besoin de nourrir le débat public avec des données indépendantes et plus complètes sur la délinquance à travers le territoire.

Des statistiques structurellement incomplètes et difficiles à interpréter

Les statistiques de la délinquance sont structurellement imparfaites, car elles reflètent d'abord la propension des victimes à déposer plainte. Elles doivent donc être complétées par des enquêtes de victimation. En France, il a fallu attendre 1995 pour que l'INSEE, à la suite d'une orientation européenne, introduise un module fixe sur la victimation et le sentiment d'insécurité au sein de l'enquête permanente sur les conditions de vie des ménages (EPCVM) organisée chaque année. Puis à partir de 2007, l'INSEE réalise annuellement une enquête de victimation à part entière : l'enquête Cadre de vie et sécurité (CVS), en partenariat étroit avec l'ONDRP et le service statistique ministériel de la sécurité intérieure (SSMSI), depuis la création de ce dernier fin 2014. L'enquête est menée en face-à-face par des enquêteurs de l'INSEE auprès d'environ 23 000 ménages ordinaires résidant en France métropolitaine[44].

Les statistiques de la délinquance mises à disposition aujourd'hui sont incomplètes, le décompte des faits dits « constatés » par la police et la gendarmerie et transmis à la justice (appelé « état 4001 ») ne comprenant pas les infractions de circulation routière, et excluant les contraventions. Cela entraîne une différence à la fois avec les enquêtes de victimation, qui prennent en compte toutes les atteintes, quel que soit leur niveau de gravité, et avec les statistiques judiciaires, qui comptent les contraventions de cinquième classe avec les crimes et délits[45]. Le SSMSI a néanmoins commencé depuis 2019 à inclure les contraventions dans le contour de certains indicateurs qu'il produit : destructions et dégradations, discriminations (racisme, anti-LGBT, sexisme, notamment outrages sexistes, entre autres).

Ces indicateurs sont difficiles à interpréter, d'abord du fait de l'absence d'homogénéité en nombre entre le fait, l'auteur du fait et la victime[46]. Ensuite, l'évolution des statistiques dans le temps dépend beaucoup des conventions et des pratiques des plaignants et des forces de sécurité. C'est par exemple historiquement le cas pour les violences conjugales (comme la hausse du nombre de violences conjugales enregistrées depuis le Grenelle sur cette problématique) et les violences sexuelles (par exemple la hausse du nombre de violences sexuelles enregistrées après le mouvement « *me too* ») ou pour les affaires impliquant des mineurs[47]. Les variations des indicateurs ne reflètent donc pas nécessairement une évolution du phénomène que l'on cherche à observer, mais un changement de comportement des acteurs qui interviennent dans sa construction (du côté des

[44] A. Estival, O. Filatriau, « La mesure statistique de la délinquance », AJ Pénal, Dalloz, avril 2019.

[45] B. Aubusson, N. Lalam, R. Padieu, P. Zamora, « Les statistiques de la délinquance », France, portrait social, édition 2002-2003, INSEE, p. 141-158.

[46] Un seul fait peut avoir un ou plusieurs auteur (ayant agi en association ou en bande), peut ne faire aucune victime ou au contraire plusieurs ; et plusieurs faits peuvent être rassemblés dans une seule affaire, dès que le dossier est pris en charge par la justice (comme rappelé dans B. Aubusson, N. Lalam, R. Padieu, P. Zamora).

[47] Ibid.

plaignants, ou du côté des forces de l'ordre qui enregistrent les plaintes). Enfin, l'enquête CVS permet d'observer les évolutions de la délinquance sans l'influence des comportements de dépôt de plainte.

Un service statistique ministériel créé en 2014, notamment pour améliorer la qualité

Le service statistique ministériel (SSMSI) du ministère de l'Intérieur est le plus récent à avoir été créé, en 2014, notamment après qu'un groupe de travail interne au ministère a recommandé de « rompre avec une présentation des statistiques reposant sur des indicateurs trop globaux, trop imprécis et trop hétérogènes (chiffre unique de la délinquance, taux d'élucidation global) ». Le constat avait également été formulé par l'Observatoire national de la délinquance et des réponses pénales (ONDRP) et par une mission d'information parlementaire de 2013 relative à la mesure statistique des délinquances et de leurs conséquences[48].

Depuis 2014, le SSM a réalisé des travaux d'amélioration de la qualité des statistiques de la délinquance, comme la correction d'anomalies ou la création d'indicateurs permettant de comparer les variations mensuelles en corrigeant la saisonnalité des phénomènes[49]. Ce travail a permis, selon les interlocuteurs de la mission, de « diminuer l'hystérisation » du débat sur les statistiques de la délinquance.

Par ailleurs, la modernisation des systèmes d'information de la police nationale et de la gendarmerie nationale a aussi permis l'amélioration des statistiques, qui centralisent depuis 2016 toutes les infractions enregistrées par les forces de sécurité (crimes et délits mais aussi contraventions de 4e et 5e classes), enregistrées par les forces de sécurité, selon la nature d'infraction et avec plus de détails que dans l'état 4001 (lieu des faits, lien entre la victime et l'auteur, mode opératoire, entre autres). Cependant, les informations centralisées sur les contraventions enregistrées par la gendarmerie sont très limitées : les informations sur les victimes et les auteurs n'en font par exemple pas partie.

Un service statistique au rôle centralisateur encore trop peu reconnu

La montée en charge du SSM est progressive mais encore insuffisante. La fiabilisation des statistiques mensuelles de la délinquance, notamment par la mise en cohérence des pratiques d'enregistrement de la gendarmerie et de la police, représente une mission conséquente, qui obère les capacités du SSM pour centraliser davantage de données.

S'agissant de la mise à disposition de données auprès de publics externes, le SSM a d'abord fait le choix de concentrer ses efforts sur les chercheurs, mais cette fonction n'est pas encore suffisamment centralisée. Si le SSM vise à permettre aux chercheurs d'y accéder avec toutes les garanties de sécurité nécessaires, via le CASD, cette mise à disposition n'est pas exclusive, et les directions métiers continuent d'établir des partenariats en dehors de ce cadre avec des chercheurs, le plus souvent pour des données qui ne sont pas partagées avec le SSM. La centralisation du rôle du SSM dans l'utilisation des données à des fins de recherche semble nécessaire, d'autant plus que le SSM récupère en 2021 les missions de l'ONDRP.

Le SSM souffre plus généralement d'un défaut de positionnement, ne bénéficiant pas d'un soutien identique de la part de ses deux directions de rattachement (DGPN et DGGN), l'une d'entre elles ne l'associant pas aux travaux sur lesquels le SSM a compétence et contestant le périmètre de cette compétence par une lecture restrictive de son décret constitutif, ce qui nuit à la mise en œuvre de ses missions. À titre d'exemple, les données de la plateforme des violences sexuelles ne sont pas mises à disposition du service statistique car la direction considère que la protection des données

[48] A. Estival, O. Filatriau, « La mesure statistique de la délinquance », AJ Pénal, Dalloz, avril 2019.

[49] Par exemple dans le cas des cambriolages, qui connaissent des pics saisonniers, notamment en décembre, ce qui perturbe l'interprétation des variations mensuelles.

« sur la vie sexuelle » prévue par l'article 7 bis de la loi de 1951 sur la statistique publique empêche cette communication.

Au-delà de l'ouverture des données, le SSM a un rôle de partage et d'accès aux données pour les administrations partenaires du ministère de l'Intérieur (comme le ministère de la transition écologique dans la prévention des risques, des catastrophes naturelles, et la gestion de crise par exemple), qui ne peut être freiné pour des motifs de rivalité entre services. Le périmètre du SSM doit donc être officiellement élargi à l'ensemble des statistiques des principaux métiers du ministère, au premier titre desquels la sécurité intérieure (soit sécurité publique, enquêtes judiciaires, enquêtes administratives et renseignement), mais aussi la sécurité civile. Ses moyens, en regard, doivent être mis à niveau, y compris par des transferts d'effectifs depuis les directions où s'exercent déjà ces compétences à ce jour.

Le positionnement du SSM comme centre névralgique de la production de données du ministère est donc prioritaire, en lien avec l'administrateur ministériel de la donnée, responsable de sa valorisation. La prise en charge et la refonte à venir de l'enquête CVS, à la suite de l'INSEE, conforteront le SSM dans ce rôle, mais représente aussi un besoin de financement que le ministère doit anticiper, pour ne pas perdre la qualité d'information dont il dispose aujourd'hui. Mais les autorités politiques du ministère doivent rappeler que l'objectif d'une statistique de qualité et qui pare les erreurs d'interprétation passe par l'engagement de l'ensemble des directions et des opérateurs dans la démarche, comme l'Agence nationale de traitement automatisé des infractions (ANTAI) par exemple.

La mission recommande ainsi de renforcer le positionnement et les moyens du service statistique ministériel pour en faire un centre de production de la donnée, y compris des opérateurs, et de mise à disposition des chercheurs, et promouvoir la valorisation en lien avec l'administrateur ministériel de la donnée.

Une opacité persistante, que des enjeux de sécurité ne justifient pas toujours

En dépit des efforts réalisés, les données de la délinquance demeurent très peu accessibles, avec trois types de difficulté principales :

- un défaut d'ouverture, quand aucune information n'est donnée : c'est le cas des sanctions contre les agents de la gendarmerie nationale, par exemple (celles de la police sont disponibles) ;
- un « semblant » d'ouverture, quand des statistiques sont communiquées, par exemple dans le cadre de conférences de presse, mais ne sont pas mises à disposition dans un format exploitable (formats représentant des tableurs) : c'est le cas des verbalisations pour non-respect du confinement, par exemple, ou des résultats mensuels des services communiqués par le ministre de l'intérieur depuis le mois d'octobre 2020[50] ;
- une granularité ou une fraîcheur limitée, quand les statistiques sont publiées à un rythme annuel, alors que la donnée mensuelle est disponible, ou à un échelon national, alors que l'échelon départemental est disponible ; ce degré de granularité peut cependant représenter un coût de fiabilisation non négligeable.

Une cartographie partielle des données a été réalisée par le site *lepanieralasade.fr*, qui n'est pas neutre par ses prises de position sur l'action des forces de sécurité, mais dont la présentation reproduite ici est conforme aux constats de la mission (cf. illustration).

Les services de sécurité ont exprimé à la mission plusieurs craintes, par exemple que la publication de statistiques à des échelons locaux plus fins ne conduisent à une « compétition » de la délinquance à travers le territoire, à l'image de ce qui a été parfois constaté pour les voitures brûlées lors du Nouvel An. La mission estime que communiquer de manière à demi-transparente sur un

[50] Certaines de ses données ont toutefois été publiées par le SSMSI dans une publication de juillet 2020 (*Insterstats Analyse* n°28).

nombre très limité de données concentre l'attention publique sur des chiffres qui sont mal interprétés, là où une ouverture plus large élèverait le niveau du débat public et permettrait de mieux identifier les difficultés à travers le territoire. Actuellement, le manque d'informations peut conduire à mettre en cause la réalité statistique de certains phénomènes (par exemple l'absence de données disponibles sur les violences contre les agents et sur leur évolution dans le temps).

Les services font également valoir que le grand public n'est pas capable d'interpréter correctement les données mises à disposition. En réalité, ce problème se pose déjà, parce que la donnée est incomplète et insuffisamment documentée : comme l'a souligné une des directions rencontrées, l'Institut Montaigne a par exemple interprété à tort les effectifs présents par département et affirmé qu'en Seine-Saint-Denis il y aurait « deux fois moins de policiers par habitant » que dans l'Indre[51]. En réalité, Le Blanc, commune de l'Indre, héberge le Commandement du soutien opérationnel de la gendarmerie nationale, assurant une mission de logistique et de gestion administrative pour l'ensemble du territoire, ce qui biaise totalement les effectifs présents géographiquement dans ce département. L'erreur est née précisément du fait que les données publiées étaient incomplètes et insuffisamment documentées.

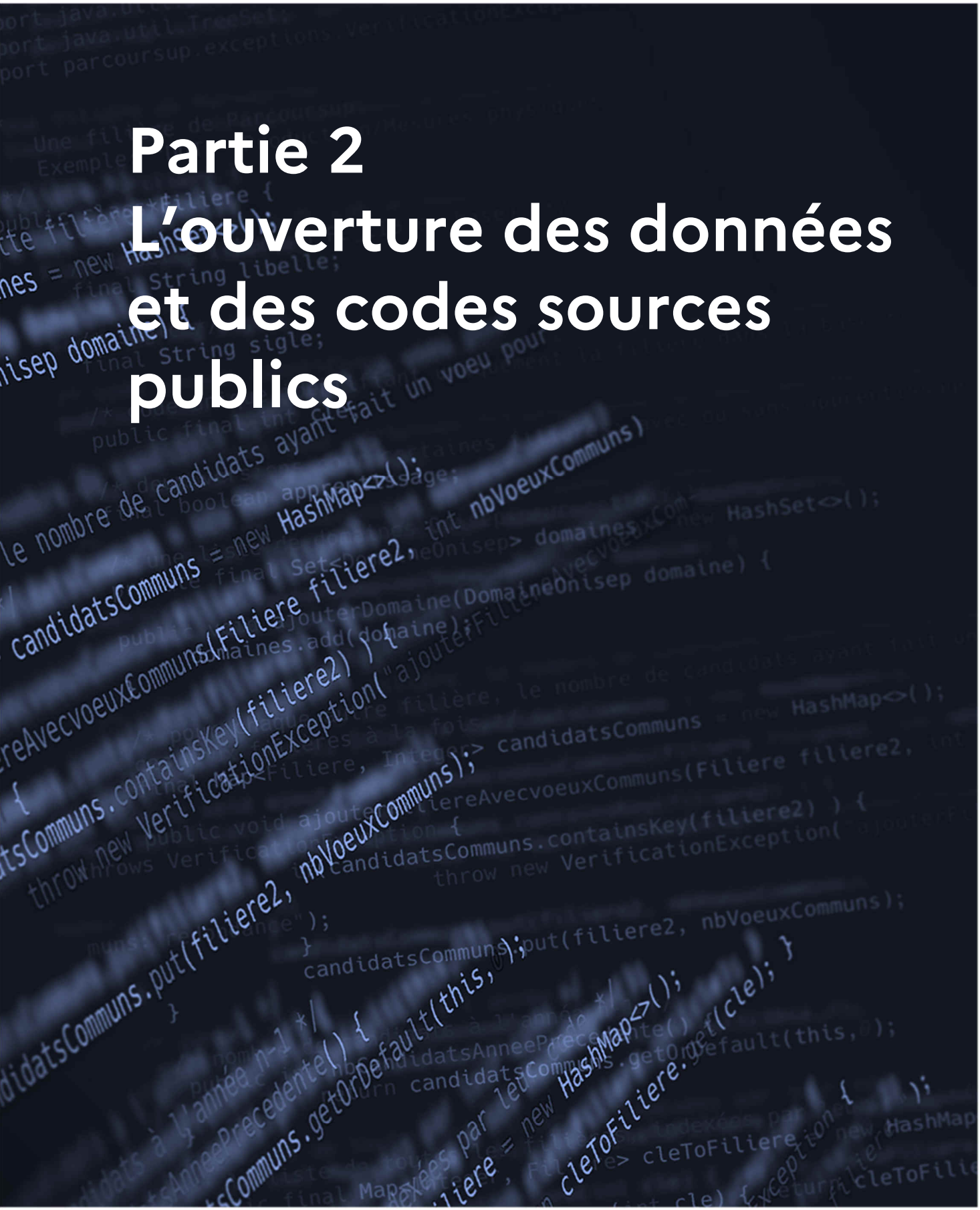
Enfin, les services mettent en avant le risque encouru par les agents des forces de sécurité intérieure. La mission tient à rappeler que l'ensemble de ses recommandations s'inscrivent dans le cadre législatif actuel, c'est-à-dire du respect des droits attachés aux données personnelles, et qu'il n'est pas envisageable de publier des données qui mettraient en danger des individus. En outre, le respect strict des règles de secret statistique par le SSMSI garantit qu'il n'est pas possible de connaître une personne dans les résultats qu'il diffuse, ce qui suppose que les cases des tableaux diffusés contiennent un effectif suffisant (qu'il s'agisse de mis en cause, de victimes ou d'agents des forces de sécurité intérieure).

La mission recommande ainsi d'ouvrir de nouvelles bases de données relatives à la délinquance, et d'en augmenter la fréquence et la granularité de publication.

[51] Institut Montaigne, Hakim El Karoui, *Les quartiers pauvres ont un avenir*, octobre 2020, p. 153.

Cartographie des données de la police nationale et de la gendarmerie nationale

Source : *lepanierasalade.fr/index*



Partie 2
L'ouverture des données et des codes sources publics

1. Un cadre juridique à l'avant-garde européenne, qui demeure cependant complexe

1.1. Le cadre général de l'ouverture des données et des codes sources publics

Le cadre juridique de l'ouverture des données et des codes sources s'est d'abord développé autour de la notion d'accès aux documents administratifs consacrée par la loi de 1978, dite « loi CADA »[52]. L'ensemble des textes législatifs et réglementaires français encadrant l'ouverture des données et des codes sources est codifié dans le code des relations entre le public et l'administration[53]. La loi CADA, qui était à l'origine une proposition de loi, a été introduite avec l'objectif de faciliter l'accès des citoyens aux documents administratifs qui les concernent, mais a aussi contribué à ce que l'administration rende compte de son action et à réduire le champ du secret administratif. Plus récemment, les lois Valter de 2015 et pour une République numérique en 2016 (dite « loi Lemaire ») ont encore renforcé la portée et l'effectivité de ce droit d'accès, et ont consacré un principe d'ouverture par défaut des données et des codes sources.

Ce cadre juridique est encore en évolution au niveau européen, avec notamment la directive du 20 juin 2019[54] qui n'est pas encore transposée en droit interne mais doit l'être avant le 17 juillet 2021. Cette directive, créée en 2003 et appelée « directive PSI » (pour *Public Sector Information*), avait déjà été modifiée en 2013, et devient désormais la directive « concernant les données ouvertes et la réutilisation des informations du secteur public », **évolution sémantique qui reflète l'importance prise par la donnée dans l'accès à l'information publique.**

Une ouverture dont l'initiative a été confiée à l'administration et non plus au seul citoyen en 2016

Le droit à la communication et à la publication des documents administratifs distingue trois cas de figure, pour lesquels l'obligation d'ouverture de données et de codes sources s'applique différemment : i) le partage de documents administratifs au sein de l'administration ; ii) le droit à la communication des documents administratifs des personnes privées ; iii) l'obligation de publication par défaut des données et des codes sources publics.

S'agissant du partage de documents administratifs au sein de l'administration, le principe est que les administrations communiquent entre elles les données et les codes sources qui possèdent un intérêt pour d'autres administrations, dans le respect des secrets légaux. Ainsi, la circulation des documents administratifs au sein de l'administration[55] est définie par l'article 1er de la loi pour une République numérique, qui n'est pas codifié :

[52] Loi n°78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal

[53] Livre III du Code des relations entre le public et l'administration : l'accès aux documents administratifs et la réutilisation des informations publiques.

[54] Directive (UE) 2019/1024 du Parlement européen et du Conseil du 20 juin 2019 concernant les données ouvertes et la réutilisation des informations du secteur public.

[55] Par administration on entend les administrations au sens de l'article L.300-2 du Code des relations entre le public et l'administration : « l'État, les collectivités territoriales ainsi que […] les autres personnes de droit public ou les personnes de droit privé chargées d'une […] mission » de service public.

- les administrations sont tenues de communiquer aux autres administrations qui en font la demande les documents administratifs concernés qu'elles détiennent ;
- par ailleurs, les informations contenues dans ces documents administratifs sont librement utilisables par l'administration qui a reçu les documents, et cette communication ne peut pas faire l'objet d'une redevance quelconque lorsqu'elle a lieu au sein de l'État et ses établissements publics administratifs (EPA) ;
- les documents publics partagés entre administrations sont soumis au régime général des documents administratifs décrits ci-dessous, sauf dans le cas d'un document transmis à une administration française par une institution ou un organe de l'Union Européenne. Dans ce cas, c'est le règlement 1049/2001 du Parlement européen et du Conseil relatif à l'accès du public aux documents du Parlement européen, du Conseil et de la Commission qui s'applique, et pour l'application duquel la CADA n'est pas compétente[56].

S'agissant du droit à la communication des documents administratifs des personnes privées, la mise en ligne peut être effectuée à la demande d'une personne privée. L'article L. 311-9 du Code des relations entre le public et l'administration (CRPA) dispose que le demandeur peut accéder aux documents « *par publication des informations en ligne, à moins que les documents ne soient communicables qu'à l'intéressé* » et si les moyens techniques de l'administration concernée le permettent[57]. Le droit à la communication sert de fondement aux recours des personnes privées devant la Commission d'accès aux documents administratifs (CADA), notamment lorsqu'elles souhaitent avoir accès à des documents qui les concernent.

Enfin, l'obligation de publication par défaut des données et codes sources publics s'applique à certaines catégories de documents administratifs. Cette obligation, définie à l'article L. 312-1-1 du CRPA, concerne les administrations qui emploient plus de cinquante personnes en équivalent temps plein (ETP) et les collectivités territoriales de plus de 3 500 habitants. Ces documents sont :

- « les documents que [les administrations] communiquent [dans le cadre du droit à la communication des documents administratifs], et leurs mises à jour ;
- « les documents qui figurent dans le répertoire des informations publiques mentionné à l'article L. 322-6[58] et qui contient les principaux documents produits ou détenus par les administrations ;
- « les bases de données, mises à jour de façon régulière » que les administrations produisent ou reçoivent « et qui ne font pas l'objet d'une diffusion publique par ailleurs ;
- « les données, mises à jour de façon régulière, dont la publication présente un intérêt économique, social, sanitaire ou environnemental ».

[56] Avis 20183478 de la CADA relatif à l'Autorité de régulation des communications électroniques et des postes (ARCEP).

[57] En particulier, des frais de reproduction peuvent être mis à la charge du demandeur. Ces frais sont « le coût du support fourni au demandeur, le coût d'amortissement et de fonctionnement du matériel utilisé pour la reproduction du document ainsi que le coût d'affranchissement selon les modalités d'envoi postal choisies par le demandeur » (article R.311-11 du Code des relations entre le public et l'administration).

[58] Article L.322-6 du Code des relations entre le public et l'administration : « les administrations qui produisent ou détiennent des informations publiques tiennent à la disposition des usagers un répertoire des principaux documents dans lesquels ces informations figurent. Elles publient chaque année une version mise à jour de ce répertoire. »

Ainsi, dans une acception stricte des textes, les données sont régies par le principe d'ouverture par défaut, mais les codes sources continuent de relever du régime général d'accès aux documents administratifs, où l'initiative de l'ouverture revient à un requérant et non à l'administration en dehors de toute demande. Leur ouverture est donc faite dès lors qu'une demande est effectuée et acceptée par l'administration (après un éventuel avis de la CADA) pour un code source particulier. Cependant, ce droit à la communication des codes sources ne peut être mis en œuvre que si les citoyens ont connaissance de leur existence : le répertoire des informations publiques répond, en théorie, à ce besoin. Une acception plus large des textes, selon laquelle les codes rejoignent les données dans le principe d'ouverture par défaut, semble plus cohérente avec l'intention exprimée par le législateur en 2016 et avec la jurisprudence de la CADA[59]. De plus, l'interprétation extensive de certains acteurs de la notion de code source, comme une donnée, conduirait à leur publication par défaut dans la mesure où cela *« présente un intérêt économique, social, sanitaire ou environnemental »*.

Par ailleurs, la communication des documents administratifs ne s'applique qu'aux documents achevés ou non préparatoires à une décision administrative d'après l'article L.311-2 du CRPA. Dans le cas des codes sources, la jurisprudence considère qu'une version correspond à une phase d'achèvement, et que le code est donc communicable à chaque version[60]. D'après le guide établi par le département Etalab de la direction interministérielle du numérique (DINUM), un code source est achevé dès lors qu'il est effectivement utilisé dans le cadre de la mission de service public par l'administration en question[61].

Enfin, l'administration n'est pas tenue de mettre à disposition en cas de demandes abusives, celles-ci étant abusives « par leur nombre ou leur caractère répétitif ou systématique », d'après l'article L. 311-2 du CRPA, ou bien lorsqu'elles ont « pour objet de perturber le bon fonctionnement de l'administration sollicitée ou qui aurait pour effet de faire peser sur elle une charge disproportionnée au regard des moyens dont elle dispose », selon la jurisprudence du Conseil d'État[62]. Ainsi, une demande adressée au ministre de l'économie et des finances de mettre à disposition toutes les informations « communicables » des bases de données électroniques du ministère, sous une forme brute et dans un format réutilisable, relatives à l'application « Chorus », pour la période 2012 à 2017, a reçu un avis défavorable de la CADA le 18 avril 2019 car « par son volume, sa complétude et son hétérogénéité, le contenu de cette base de données ne se prête pas à la possibilité d'un examen au cas par cas de l'existence d'un intérêt protégé, qu'il soit absolu ou relatif. » La CADA ajoute que « le tri des documents versés dans l'application Chorus équivaudrait en réalité à la confection d'une nouvelle base documentaire, qui n'existe pas en l'état et ne pourrait être obtenue à ce jour par un traitement automatisé d'usage courant mais seulement au prix d'efforts disproportionnés au regard des moyens dont dispose l'administration. »

Un principe général de réutilisation, libre et gratuite, des informations publiques

Le droit d'accès au document est complété d'un droit de réutiliser les informations publiques contenues dans ce document. La réutilisation des informations publiques est, en principe, libre et gratuite. S'il est possible de mettre en place des redevances de réutilisation ou de mise à disposition, ces dernières sont toutefois strictement encadrées depuis la loi pour une République numérique de 2016.

---

[59] La CADA, dans son avis n°20161990 du 23 juin 2016 relatif à l'algorithme développé par le ministère de l'éducation nationale connu sous le nom d'admission post bac dit « APB », indique qu'un algorithme constitue un document administratif. Cette jurisprudence a notamment été reprise concernant les modèles de la direction générale du Trésor (avis n°20180276 du 28 juin 2018).

[60] Tribunal administratif de Paris, 18 février 2016, n°1508951/5-2

[61] Etalab guide sur le logiciel libre, 19 novembre 2019

[62] Conseil d'État, *Ministre de la culture c/ Société pour la protection des paysages et l'esthétique de la France,* n°420055, 14 novembre 2018.

La fin progressive des redevances

Le système de redevances, supposé permettre un financement des administrations et la possibilité d'une concurrence avec le secteur privé[63], a été remis en cause, d'abord partiellement par Lionel Jospin en 1997 qui a souhaité rendre l'accès aux *« données publiques essentielles »* gratuit[64], puis de manière globale à partir de 2011 avec la création d'Etalab[65], qui a mis en place le portail unique des informations publiques[66].

Le rapport au Premier ministre établi par Adnène Trojette en 2013[67] a proposé de limiter strictement le recours aux redevances par les administrations, puis le rapport au Premier ministre établi par Antoine Fouilleron en 2015[68] a affirmé le principe de gratuité des échanges entre les administrations, finalement acté dans la loi pour une République numérique en 2016.

La liberté de réutilisation des informations publiques est précisée à l'article L. 321-1 du CRPA : *« les informations publiques figurant dans des documents communiqués ou publiés par des administrations […] peuvent être utilisés par toute personne qui le souhaite à d'autres fins que celles de la mission de service public pour les besoins de laquelle les documents ont été produits ou reçus ».* Il s'agit donc d'une liberté de réutilisation de l'information publique, et non pas du document administratif en tant que tel.

Les administrations peuvent choisir différentes licences parmi une liste définie par décret. Les conditions fixées par les licences ne peuvent apporter de restrictions à la réutilisation que « *pour des motifs d'intérêt général et de façon proportionnée* » et « *ne peuvent avoir pour objet ou pour effet de restreindre la concurrence* »[69]. Lorsque les administrations ne choisissent pas de licence spécifique, les conditions de réutilisation sont régies par les articles L. 322-1 à L. 322-7 du CRPA, qui définissent les règles générales de réutilisation.

La liste des licences homologuées est la suivante[70] :

- pour les données : la licence ouverte de réutilisation des informations publiques, et l'*Open Database License* (ODbL) ;
- pour les codes sources : les licences dites « permissives » (licence Apache, licences Berkeley Software Distribution, licence CeCILL-B et la licence Massachussets Institute of Technology) ou les licences avec obligation de réciprocité (Mozilla Public License, les licences du projet GNU et la licence CeCILL).

[63] Circulaire du 14 février 1994 relative à la diffusion des données publiques

[64] Lionel Jospin, discours d'Hourtin, 25 août 1997 : « les données publiques essentielles doivent désormais pouvoir être accessibles à tous gratuitement sur Internet ».

[65] Décret n°2011-194 portant création d'une mission « *Etalab* » chargée de la création d'un portail unique ministériel des données

[66] Circulaire du 26 mai 2011 relative à la création du portail unique des informations publiques de l'État « data.gouv.fr » par la mission « *Etalab* » et l'application des dispositions régissant le droit de réutilisation des informations publiques

[67] Ouverture des données publiques, les exceptions au principe de gratuité sont-elles toutes légitimes ? Rapport au Premier ministre, M.A. Trojette, juillet 2013

[68] Les échanges de données réalisés à titre onéreux entre les administrations, rapport au Premier ministre, A. Fouilleron, novembre 2015

[69] Article L.323-2 du CRPA. A noter qu'en recherche ce sont les licences creative commons qui sont utilisées.

[70] Article D.323-2-1 du CRPA.

Les licences permissives (ou « *non-copyleft* ») n'imposent comme contrainte au réutilisateur que la mention de la paternité du code source qu'il utilise, au contraire des **licences avec obligation de réciprocité (dites « *copyleft* »)**, qui imposent au réutilisateur de diffuser le travail fait à partir du code source public sous les mêmes conditions que le code source public original. L'usage d'une licence à réciprocité s'inscrit parfois dans une logique de « bien commun », dans la mesure où elle demande aux réutilisateurs de restituer dans les mêmes conditions les améliorations qu'il aura pu apporter[71].

Il est possible de créer une licence particulière en faisant une demande d'homologation à la DINUM[72]. Cette procédure est obligatoire dans le cas de l'établissement d'une licence dans un cadre commercial.

La gratuité de la réutilisation des informations publiques est affirmée par l'article L.324-1 du CRPA : « *la réutilisation d'informations publiques est gratuite* ». Toutefois, la mise en place de redevances reste possible dans le cas où les administrations « *sont tenues de couvrir par des recettes propres une part substantielle des coûts liés à l'accomplissement de leur mission de service public* ».

Les redevances sont strictement encadrées :

- « le produit total de la redevance ne peut pas dépasser le montant total des coûts liés à la collecte, à la production, à la mise à disposition du public ou à la diffusion des […] informations publiques »[73] ;
- la numérisation des fonds et des collections des bibliothèques peut donner lieu à une redevance[74] ;
- « le montant des redevances […] est fixé selon des critères objectifs, transparents, vérifiables et non discriminatoires »[75] ;
- dans le cas d'informations publiques produites ou reçues par l'État, la liste des informations publiques concernées par des redevances est définie par décret[76] ;
- les établissements autorisés à établir des redevances pour certaines informations publiques[77] sont l'Institut national de l'information géographique et forestière (IGN), Météo-France et le Service hydrographique et océanographique de la marine (SHOM).

[71] Pour une analyse approfondie des licences, voir la saisine DAJ du ministère de l'économie, des finances et de la relance.

[72] Article D.323-2-2 du CRPA.

[73] Article L.324-1 du CRPA.

[74] Article L.324-2 du CRPA.

[75] Article L.324-3 du CRPA.

[76] Article L.324-5 du CRPA.

[77] Les informations publiques de ces établissements pouvant faire l'objet d'une redevance sont définies à l'article D324-5-1 du Code des relations entre le public et l'administration.

Les régimes de propriété intellectuelle limitent la réutilisation, dont certains sont des héritages historiques qui n'ont jamais été remis en question

Par ailleurs, les informations contenues dans **les documents sur lesquels existe un droit de propriété intellectuelle ne sont pas réutilisables librement**. Cette disposition vise à protéger les informations présentes dans les documents administratifs communiqués par les administrations mais qui n'ont pas été produites par l'administration. Seul un tiers peut se prévaloir de droit de propriété intellectuelle sur une information publique[78], au contraire des administrations pour qui les droits de propriété intellectuelle « *ne peuvent pas faire obstacle à la réutilisation du contenu des bases de données que ces administrations publient* »[79]. La CADA considère la propriété intellectuelle à la fois dans sa dimension patrimoniale et dans ses attributs d'ordre intellectuel et moral, depuis un arrêt n°375704 du Conseil d'État du 8 novembre 2018[80].

Il est possible que l'agent public ayant créé le document administratif en question fasse usage de son droit d'auteur et limite la diffusion ou la réutilisation de ce document. Dans ce cas, c'est le droit de la propriété intellectuelle des agents publics, défini notamment dans la loi n°2006-961 du 1er août 2006, qui s'applique. Aujourd'hui, il subsiste plusieurs régimes de propriété intellectuelle, comme celui des conservateurs du patrimoine.

Le droit moral de l'agent ne devrait pas faire obstacle à la communication du document en question par l'administration, dans la mesure où le document a été produit dans le cadre de l'exercice de ses fonctions. L'accès au document administratif n'est donc pas compromis par l'existence d'un droit de propriété intellectuelle. En revanche, sa réutilisation, notamment pour une réutilisation à des fins commerciales en dehors du service public, doit prendre en compte les droits patrimoniaux de l'agent public.

Il existe néanmoins des exceptions substantielles à ce régime général dans le cas de la production de logiciel et de l'application du droit *sui generis* du producteur de données. Dans le cas des logiciels, l'agent public ne possède pas de droits patrimoniaux, en application des articles L.121-7 et L.113-9 du Code de la propriété intellectuelle. Dans le cas des bases de données, le droit *sui generis* du producteur ne permet pas à l'agent public qui crée la base de données de faire valoir des droits pouvant faire obstacle à sa réutilisation.

Néanmoins, au vu de la diversité des cas envisagés par le code de la propriété intellectuelle, une analyse au cas par cas est recommandée afin de déterminer les règles applicables en matière de propriété intellectuelle des agents publics[81].

Enfin, cela n'exclut pas que l'agent puisse décider de mettre à disposition de lui-même sa production, sans tarification, comme dans le cas d'un rapporteur public du Conseil d'État qui s'était illustré par la mise à disposition gratuite de ses conclusions, quand la pratique est de les revendre à des cabinets d'études juridiques.

[78] Article L.321-2 du Code des relations entre le public et l'administration.

[79] Article L.321-3 du Code des relations entre le public et l'administration incluant une exception pour les bases de données produites ou reçues par les administrations mentionnées au premier alinéa de l'article L. 300-2 dans l'exercice d'une mission de service public à caractère industriel ou commercial soumise à la concurrence.

[80] Voir la page 22 du rapport d'activité de la CADA de 2018.

[81] Cette analyse est reprise de la réponse fournie par la direction des affaires juridiques des ministères économiques et financiers sur saisine de la mission, dont le détail est fourni en annexe au rapport.

1.2. Une ouverture encadrée par le respect du droit de protection des données personnelles et des secrets légaux

Le principe d'ouverture par défaut et de libre réutilisation des données et des codes sources publics est encadré par plusieurs dispositions protégeant les données personnelles et les secrets légaux.

Une ouverture encadrée par les principes d'occultation ou de non-communication des documents administratifs

Premièrement, plusieurs mentions doivent être occultées avant la publication d'un document administratif, définies dans les articles L. 311-5 et L. 311-6 du CRPA[82]. L'occultation des documents administratifs consiste dans le fait de masquer ou d'occulter les informations identifiées comme non communicables. L'administration n'est pas tenue de communiquer des documents dont l'occultation dénaturerait le sens ou si elle se révèle trop complexe[83]. Ces mentions sont celles dont la consultation ou la communication porterait atteinte :

- au **secret des délibérations du Gouvernement** et des autorités responsables relevant du pouvoir exécutif ;
- au **secret de la défense nationale** ;
- à **la conduite de la politique extérieure de la France** ;
- à **la sûreté de l'État, à la sécurité publique, à la sécurité des personnes ou à la sécurité des systèmes d'information des administrations** ;
- à la **monnaie** et au **crédit public** ;
- au **déroulement des procédures engagées devant les juridictions** ou d'opérations préliminaires à de telles procédures, sauf autorisation donnée par l'autorité compétence ;
- à la **recherche** et à la **prévention**, par les services compétents, **d'infractions** de toute nature ;
- ou, sous réserve de l'article L.124-4 du code de l'environnement, aux **autres secrets protégés par la loi**.

De plus, l'article L. 311-6 précise que « ne sont communicables qu'à l'intéressé les documents administratifs :

- 1° dont la communication porterait **atteinte à la protection de la vie privée**, au secret médical et au secret des affaires, lequel comprend le secret des procédés, des informations économiques et financières et des stratégies commerciales ou industrielles et est apprécié en tenant compte, le cas échéant, du fait que la mission de service public de l'administration mentionnée au premier alinéa de l'article L. 300-2 est soumise à la concurrence ;
- 2° portant une **appréciation ou un jugement de valeur sur une personne physique, nommément désignée ou facilement identifiable** ;
- 3° **faisant apparaître le comportement d'une personne, dès lors que la divulgation de ce comportement pourrait lui porter préjudice**. »

S'agissant plus particulièrement du **secret des affaires**, la CADA a explicité sa doctrine dans l'avis n°20190911 consécutif à une demande de conseil faite par la Haute autorité de santé à la CADA. Elle estime que le secret des affaires renvoie à la notion de secret industriel et commercial qui existait avant la loi n°2018-670 du 30 juillet 2018 relative à la protection du secret des affaires. Il faut donc qu'une information se rattache au secret des procédés, au secret des informations économiques ou financières ou au secret des stratégies commerciales et industrielles.

L'article L.151-1 du Code de commerce prévoit en outre trois conditions cumulatives auxquelles une information doit répondre pour être protégée par le secret des affaires :

[82] Articles L311-7 et L312-1-2 du Code des relations entre le public et l'administration.

[83] Guide pratique de la publication en ligne et de la réutilisation des données publiques (« open data »), CNIL (Commission nationale informatique et liberté) et CADA

- ne pas être connue du grand public et/ou du secteur concerné ;
- avoir une valeur commerciale, réelle ou potentielle, parce que secrète ;
- faire l'objet de mesures spécifiques destinées à la garder confidentielle.

La CADA considère dans le même avis que ce dernier critère n'a pas de portée et qu'il est satisfait dès lors que les deux premiers critères le sont.

Enfin, l'occultation n'est pas obligatoire pour des documents rentrant dans les catégories définies par l'article D.312-1-3 du Code des relations entre le public et l'administration (organigrammes de l'administration, résultats obtenus par les candidats aux examens et concours administratifs etc.), en particulier pour les documents relevant du champ des données personnelles. Toutefois, les catégories décrites sont peu explicites d'après certains acteurs, qui demandent une précision, voire une liste des documents en question. D'après la CNIL, la clarification de ces différentes dispositions pourrait permettre une meilleure application du cadre général de l'ouverture des documents administratifs.

La protection des données personnelles

Les données personnelles, définies à l'article 4 du règlement général pour la protection des données (RGPD) comme « *toute information se rapportant à une personne physique identifiée ou identifiable* », font l'objet d'une protection spécifique et ne sont pas concernées par l'ouverture par défaut et la libre réutilisation des données et codes sources publics.

Dans le cas des données personnelles qui relèvent de la vie privée, ces données ne sont pas communicables à des tiers, et *a fortiori* non publiables[84].

Dans le cas des données personnelles non couvertes par le secret de la vie privée, la publication en ligne est proscrite, sauf s'il est possible d'anonymiser le document administratif qui contient des données personnelles. L'anonymisation ne doit pas permettre d'identifier un individu dans la base, de relier entre eux des ensembles de données distincts concernant un même individu, et de déduire de l'information sur un individu[85].

La diffusion de données personnelles peut avoir lieu sans anonymisation seulement dans trois cas :

- si une **disposition législative ou réglementaire** l'autorise[86]. Par exemple, l'article L. 127-10 du Code de l'environnement permet la diffusion de certaines informations sur la base du découpage parcellaire et des adresses des parcelles au sein de bases de données de référence[87] ;
- si le **consentement des personnes** concernées a été recueilli[88] ;
- si les documents contenant des données personnelles sont **listés à l'article D. 213-1-3 du CRPA** et peuvent être rendus publics sans avoir été au préalable anonymisés (par exemple les permis de construire).

De plus, la réutilisation des informations contenues dans les données personnelles est encadrée. Cette réutilisation doit en particulier être licite, c'est-à-dire que toute réutilisation doit être fondée sur l'article 6-1 du RGPD qui exige le consentement des personnes pour un ou plusieurs traitements aux finalités définies et l'existence de fins légitimes poursuivis par le responsable du traitement. De plus, le responsable du traitement doit assurer la sécurisation des données.

[84] Article L..311-6 du Code des relations entre le public et l'administration.

[85] Page 14 du Guide pratique de la publication en ligne et de la réutilisation des données publiques, CNIL et CADA.

[86] Article L.312-1-2 du Code des relations entre le public et l'administration.

[87] C'est sur ce principe qu'a pu être diffusée la base de données DVF des demandes de valeurs foncières.

[88] Articles 4 et 7 du Règlement général pour la protection des données.

Par ailleurs, **l'article 9 du RGPD énumère les données personnelles ayant un caractère particulier**, à savoir celles qui révèlent *« l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique ».*

Le traitement de ces données est par principe interdit, sauf si cela entre dans les finalités listées par le même article. En particulier, en matière de santé, le traitement des données peut être autorisé s'il est nécessaire « *aux fins de la médecine préventive ou de la médecine du travail, de l'appréciation de la capacité de travail du travailleur, de diagnostics médicaux, de la prise en charge sanitaire ou sociale, ou de la gestion des systèmes et des services de soins de santé ou de protection sociale […]* ».

1.3. Des dispositions sectorielles spécifiques, qui réduisent la lisibilité du droit

Outre les principes généraux présentés, il existe de nombreuses dispositions législatives et réglementaires spécifiques. Ces dispositions correspondent, en général, à des dispositions sectorielles, qui font parfois exception aux dispositions générales d'ouverture des données et des codes sources public, et à celles de libre réutilisation des informations publiques. Sans viser l'exhaustivité des dispositions législatives particulières, deux exemples emblématiques peuvent être donnés, au travers des données de santé, et des données environnementales, qui comptent parmi les droits sectoriels les plus autonomes, l'un dans le sens d'une protection renforcée (santé), l'autre dans le sens d'une ouverture large et engagée de longue date (environnement).

Les données de santé : une sécurité juridique renforcée

Du fait de leur sensibilité, les données de santé font l'objet d'un régime juridique particulièrement protecteur, prévu par l'article 9 du RGPD. Les données de santé comprennent « *l'ensemble des données se rapportant à l'état de santé d'une personne concernée qui révèlent des informations sur l'état de santé physique ou mentale passé, présent ou futur de la personne concernée* »[89]. Ainsi, les données de santé peuvent se décliner en trois catégories :

- les données de santé par nature, comme par exemple les traitements réalisés ou les médicaments pris ;
- les données qui, croisées avec d'autres données, peuvent révéler l'état de santé d'une personne (par exemple, le croisement de la tension avec la mesure de l'effort) ;
- les données de santé par destination, suivant l'utilisation qui en est faite sur le plan médical.

Le RGPD prévoit des dispositions spécifiques pour l'utilisation des données de santé, étant entendu que cette utilisation n'est pas celle d'une ouverture complète (en *open data*). Les régimes d'autorisation et de déclaration varient également selon la nature de la donnée de santé. Dans certains cas, bien qu'il prévoie un régime d'autorisation préalable à tout usage des données (« traitement »), le RGPD a prévu un allègement des formalités préalables à l'autorisation, en se fondant sur le concept de l'*accountability*: le responsable de traitement doit être capable de démontrer que son usage des données est conforme au RGPD. Le responsable de traitement doit notamment mettre en place un registre des traitements, mener des analyses d'impact, veiller à encadrer l'information des personnes, formaliser les rôles et responsabilités du responsable de traitement, désigner un délégué à la protection des données lorsque cela est obligatoire, renseigner les actions menées en matière de sécurité, etc[90]. Cette procédure allégée s'applique par exemple aux traitements pour lesquels la personne a donné son consentement explicite, ou encore les traitements nécessaires à la sauvegarde de la vie humaine.

[89] Considérant 35, RGPD

[90] CNIL, quelles formalités pour les traitements de données de santé à caractère personnel ?

Toutefois, **le législateur français a retenu que cette procédure allégée ne s'appliquerait pas pour les traitements présentant une finalité d'intérêt public**[91], pour les « *traitements automatisés dont la finalité est ou devient la recherche ou les études dans le domaine de la santé ainsi que l'évaluation ou l'analyse des pratiques ou des activités de soins ou de prévention* », et enfin la demande d'avis concernant un acte réglementaire qui autoriserait un traitement de données de santé[92]. Le choix d'une procédure plus importante en matière de recherche a un impact important sur l'autorisation des projets de recherche, et *in fine*, sur la qualité et la rapidité même de la recherche, par rapport à ce qui est possible ailleurs dans l'Union européenne.

En 2019, la loi relative à l'organisation et à la transformation du système de santé[93] **a introduit des précisions en matière de données de santé par rapport au RGPD dans le système français.** Elle a notamment mis en place la plateforme des données de santé, ou *Health Data Hub*, qui a pour objectif de faciliter l'accès aux données de santé à différents acteurs professionnels, au premier rang desquels figurent les chercheurs (cf. partie 3). En stockant les données des producteurs partenaires sur son infrastructure, le *Health Data Hub* offre ainsi un cadre d'accès aux données, sans pour autant que cela ne dispense les institutions partenaires de mener par ailleurs leur politique d'ouverture des données qui peuvent être rendues publiques.

Par ailleurs, la plateforme des données de santé participe, avec la CNIL, à l'établissement des méthodologies de référence[94] homologuées et publiées par la CNIL et accompagne les projets qui lui sont portés en matière de recherche en santé[95], en devant toutefois présenter ces projets devant la CNIL pour son autorisation. Cette procédure induit un fort ralentissement du rythme des projets, et entraînerait à terme une diminution des projets portés.

Les données de l'environnement : un régime d'ouverture et précurseur, sous l'influence du droit européen et de la Charte de l'environnement

Le secteur de l'environnement a bénéficié d'une double influence favorable à l'ouverture de ses données : celle du droit européen, et notamment des directives 2003/4/CE de 2003 et INSPIRE de 2007, et celle de la Charte de l'environnement. La Charte prévoit ainsi à son article 7 que « *toute personne a le droit, dans les conditions et les limites définies par la loi, d'accéder à des informations relatives à l'environnement détenues par les autorités publiques* ». Dans ce cadre sectoriel, il s'agit d'un accès à l'information et non aux documents, ce qui est donc plus large que le simple accès aux documents administratifs. La notion d'information relative à l'environnement, précisée dans la directive 2003/4/CE concernant l'accès du public à l'information en matière d'environnement, est reprise dans les dispositions des articles L. 124 et suivants du code de l'environnement qui la transposent.

L'ordonnance n°2010-1232 codifiée dans le code de l'environnement a ainsi permis l'ouverture de données personnelles spécifiques, comme la parcelle ou l'adresse, ce qui a rendu possible la constitution puis l'ouverture de la base demandes de valeurs foncières (DVF) concernant le prix des transactions immobilières.

---

[91] Article 66 de la loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, modifié par la loi n°2019-774 du 24 juillet 2019.

[92] CNIL, cf. *supra*

[93] Loi n°2019-774 relative à l'organisation et à la transformation du système de santé

[94] Une méthodologie de référence constitue un mode d'emploi permettant aux acteurs de se mettre en conformité avec l'état actuel du droit. Par exemple, la méthode de référence MR-001 « encadre les traitements comprenant des données de santé et présentant un caractère d'intérêt public, réalisés dans le cadre de recherches nécessitant le recueil du consentement de la personne concernée ou de ses représentants légaux ».

[95] Article 41 de la loi n°2019-774 relative à l'organisation et à la transformation du système de santé

L'ordonnance de 2010 a également transposé la directive INSPIRE qui s'applique aux données géographiques numériques détenues par des autorités publiques. Pour ces données, elle impose la publication des métadonnées associées, l'usage de services en lignes pour l'accès aux données et aux métadonnées, une structuration harmonisée[96], enfin des mécanismes de coordination.

La loi pour la reconquête de la biodiversité de 2016[97] prévoit quant à elle l'ouverture des données brutes acquises lors des études d'impact en contribuant à l'inventaire du patrimoine naturel, ainsi que des données d'associations naturalistes participant à cet inventaire.

Si les secrets énoncés ci-dessus sont toujours présents, l'administration ne peut pas les invoquer pour refuser de communiquer une information environnementale sans considérer dans le même temps l'intérêt de la publication de la donnée environnementale (article L. 124-5 du code de l'environnement). Ainsi, l'administration possède le moyen légal de passer outre les différents secrets si elle l'estime opportun[98]. Dans le cadre des informations relatives à des émissions de substance dans l'environnement, l'administration ne peut invoquer la plupart des secrets existants pour justifier la non-publication de cette information[99]. Ainsi, la banque nationale des ventes des distributeurs de produits phytosanitaires est ouverte par le biais de l'article L.124-5 du code de l'environnement (BNV-d)

Les données de la recherche : un régime spécial dans un domaine alliant données et codes sources publics et privés

La recherche est un domaine à la frontière du public et du privé. Si une donnée est produite par un institut de l'enseignement supérieur et de la recherche, alors cette donnée est publique du fait du caractère administratif de l'institut. Dans ce cadre, le régime juridique décrit plus haut s'applique.

Dans le cas où une donnée est produite par un acteur privé ou ne relevant pas directement d'une administration, le critère du financement permet de déterminer si une donnée est publique ou non. Dans ce cas, la communication de la donnée suit les règles décrites plus haut, et sa réutilisation est libre si elles ont été publiées par l'organisme ou par le chercheur.

1.4. Un cadre européen en pleine transformation, avec un agenda ambitieux et en phase avec l'évolution française

Le droit européen et français est en constante évolution, notamment sur la question de l'ouverture et du partage des données ainsi que des codes sources. Ainsi, la directive (UE) 2019/1024 du 20 juin 2019[100] doit encore être transposée en droit français, d'ici au 17 juillet 2021. Cette directive s'inspire notamment de la loi pour une République numérique de 2016. Elle prévoit de plus d'élargir l'ouverture des données au sein de l'Union européenne, notamment en favorisant les dispositions pratiques permettant la réutilisation par tous, en particulier pour les données dynamiques. Elle définit un ensemble de « *données de forte valeur* ». L'identification de ces données est spécifiée à l'article 14 de la directive et repose sur « *leur aptitude potentielle » à :*

[96] Ce point a été souvent mentionné lors des auditions de la mission comme un facteur favorable à la standardisation de la donnée.

[97] Loi n°2016-1087 pour la reconquête de la biodiversité, de la nature et des paysages.

[98] CADA, conseil n°20192493 et avis n°20190373.

[99] Les conditions exactes du refus sont détaillées à l'article L.124-5 du Code de l'environnement. Notamment, un document ne peut porter atteinte à la « conduite de la politique extérieure de la France, à la sécurité publique ou à la défense nationale », « au déroulement des procédures juridictionnelles ou à la recherche d'infractions pouvant donner lieu à des infractions pénales », « à des droits de propriété intellectuelle ».

[100] *Ibid.*

- générer des avantages socio-économiques ou environnementaux importants et des services innovants ;
- bénéficier à un grand nombre d'utilisateurs, notamment des PME ;
- contribuer à générer des recettes ;
- et être associées à d'autres ensembles de données ».

Pour chaque série de données de forte valeur, la Commission procède à une analyse d'impact et consulte les experts adéquats, avant de déclarer une donnée comme étant une donnée de forte valeur par un acte d'exécution. La procédure des actes d'exécution sur les séries de données de grande valeur commencera au premier trimestre 2021[101].

De plus, un cadre législatif générique pour la gouvernance des espaces européens communs des données (*Data Governance Act*) est prévu pour le quatrième trimestre 2020[102]. Ce cadre traitera notamment :

- de la gouvernance au niveau de l'Union européenne et dans les États membres avec comme but de favoriser l'interopérabilité des données ;
- des données sensibles, détenues par le secteur public, mais non couvertes par la directive des données ouvertes, et de leur utilisation dans des projets scientifiques ;
- de l'« altruisme en matière de données », pour permettre aux particuliers d'autoriser l'utilisation de leurs données personnelles dans le cadre du RGPD.

Le *Data Governance Act* devrait être suivi d'un *Data Act* dans la deuxième partie de l'année 2021.

De plus, un changement de paradigme du droit européen en matière de données personnelles est en cours de réflexion dans le cadre de la révision de la directive donnée ouverte de 2018[103]. En effet, l'émergence de nouvelles technologies permettant de traiter des données personnelles tout en respectant l'anonymat des individus crée des incertitudes sur la frontière entre données personnelles et données non personnelles. La Commission conduit une réflexion pour placer l'individu au cœur du dispositif, afin qu'il puisse choisir quand, à qui et sous quelles conditions il met à disposition ses propres données personnelles.

Les évolutions prévues par le *Digital Services Act*, priorité de l'exécutif européen, et par le *Data Act* sont notamment abordées dans la partie 5 du rapport.

[101] Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au Comité des régions, *Une stratégie européenne pour les données*, 19 février 2020.

[102] *Ibid*.

[103] Audition de la DG Connect de la Commission européenne par la mission.

2. Une dynamique d'ouverture des données et codes sources publics à relancer

2.1. La France a acquis une position favorable dans les classements internationaux mais ne progresse plus depuis 2017

Une dynamique qui a porté la France parmi les meilleurs

Depuis 2017, la France fait partie des trois premiers pays des deux classements internationaux de référence en matière d'ouverture des données, l'*OURdata Index* de l'OCDE (deuxième place) et l'*Open Data Maturity* du Portail européen des données (troisième place)[104]. Dans le classement OCDE, la France se situe derrière la Corée du Sud, avec un indice global de 0,89 (l'indice allant de 0 à 1). Dans le classement européen, elle figure derrière l'Irlande et l'Espagne. Parmi les forces du système français, l'accessibilité de la donnée est un domaine dans lequel la France obtient un bon score dans le classement de l'OCDE[105], grâce au portail *data.gouv.fr*, qualifié comme un des plus aboutis à l'international, en permettant aux utilisateurs d'ajouter leurs propres jeux de données et de contribuer à l'amélioration de la qualité de la donnée.

La France figure également en bonne place dans deux autres classements de référence de l'ouverture des données, qui exploitent d'autres sources d'informations. Elle est en quatrième position, derrière le Canada, le Royaume Uni et l'Australie, dans l'Open Data Barometer de la Word Wide Web Foundation, qui enrichit les données gouvernementales avec des informations externes. L'Open Data Barometer évalue la publication de certaines catégories de données, observe de manière plus globale la maturité des politiques d'open data, leur mise en œuvre et leur impact économique, social et politique. La France est également en quatrième position également, derrière Taïwan, l'Australie et la Grande-Bretagne, dans le Global Open Data Index, qui utilise uniquement des données « crowdsourcées » et évalue l'état de diffusion de certaines catégories de données considérées à fort potentiel d'impact.

Dans le domaine de l'ouverture des codes sources publics, la France est également un des pays les plus avancés en Europe, et, plus largement, pour la place des logiciels libres dans le secteur public. L'observatoire de la Commission européenne (*Open Source Observatory*) salue notamment l'action de la direction du numérique (DINUM) et de l'association des développeurs et des utilisateurs de logiciels libres pour les administrations et les collectivités locales (ADULLACT) dans ce domaine[106]. Le projet Openfisca a par exemple reçu en 2019 le premier prix *Sharing & Reuse* décerné par la Commission européenne, dans le domaine des ouvertures de codes sources les plus innovantes.

Enfin, la politique d'ouverture des données contribue à la position de la France dans certains classements internationaux de compétitivité, comme le classement *Doing Business* établi par la Banque mondiale depuis 2003 pour mesurer le caractère favorable de l'environnement règlementaire d'un pays pour les affaires. À titre d'exemple, la disponibilité et la transparence de l'information en matière de propriété est un indicateur suivi par la Banque mondiale, pouvant représenter jusqu'à deux points de classement supplémentaire sur 100, si d'une part un système d'information géographique entièrement numérisé est mis à disposition, avec les informations du cadastre (1 point) et si d'autre part le registre de propriété immobilière et la cartographie sont liés entre eux (1 point également). L'ouverture de la base demandes de valeurs foncières (DVF) en 2019 a ainsi permis de faire progresser la France dans ce classement.

---

[104] Les classements retenus ici sont ceux qui abordent le plus spécifiquement les actions en matière d'ouverture des données et des codes sources. Les résultats ne sont donc évidemment pas nécessairement homogènes avec le classement de la Commission européenne mesurant l'innovation (*European Innovation Scoreboard*), où la France a le rang de 15ème pays sur 32.

[105] OCDE, Ourdata Index 2019, France.

[106] Commission européenne, Open Source Software Country Intelligence Report, France, mars 2020.

Des résultats inégaux, qui progressent peu depuis 2017

L'essentiel des progrès réalisés par la France dans les classements internationaux est intervenu avant 2017, les places et positions n'ayant pas évolué depuis cette date (stable pour les deux classements de référence, renouvelés depuis 2017, l'OCDE et le portail européen des données). Les résultats des trois sous indicateurs utilisés par l'OCDE progressent également peu : la disponibilité de la donnée est certes passée d'une note de 0,28 à 0,30 entre 2017 et 2019, mais l'accessibilité reste notée à 0,31, sans progrès depuis 2017.

Les classements mettent par ailleurs en évidence plusieurs marges de progression pour la France. Dans son classement, le score attribué par l'OCDE à la France en matière de soutien public à la réutilisation de la donnée est relativement plus faible que pour les autres pays, notamment par rapport à la Corée du Sud : si la France progresse depuis 2017 grâce à des moyens engagés au sein du secteur public, cette action pourrait être améliorée, selon l'OCDE, par un meilleur suivi des réutilisations au sein du secteur public.

En outre, si la France est toujours placée au-dessus de la moyenne des pratiques dans les classements de l'OCDE et du Portail européen des données, cette avance relative est plus faible dans le domaine de la qualité de la donnée, où la France n'est que septième dans le classement du Portail européen des données et où ses résultats se situent plutôt dans les standards observés chez ses partenaires européens[107].

Des engagements pris mais inachevés pour le partenariat pour un gouvernement ouvert

La France a pris des engagements en matière d'ouverture des données, dans le cadre de la démarche du partenariat pour un gouvernement ouvert (PGO, *Open Government Partnership*). Cette initiative multilatérale, lancée en 2011 et que la France a rejointe en avril 2014, regroupe 79 pays et des centaines d'organisations de la société civile. Les pays qui rejoignent le PGO s'engagent à respecter les principes de la Déclaration du gouvernement ouvert de septembre 2011 : transparence de l'action publique, notamment *via* l'ouverture des données publiques ; participation des citoyens à l'élaboration et à l'évaluation des politiques publiques ; intégrité de l'action publique et des agents publics ; utilisation des nouvelles technologies en faveur de l'ouverture et de la redevabilité. Les pays membres doivent élaborer tous les deux ans un plan d'action national, en concertation avec la société civile.

Le plan adopté par la France pour 2018-2020[108] fixe ainsi 21 engagements dont 8 sont étroitement liés aux enjeux d'ouverture des données et des codes sources :

- l'engagement 4, « enrichir le service public de la donnée : vers une nouvelle liste de données de référence », dont deux des trois actions sont considérées comme réalisées par la DITP au 1er octobre 2020 ;
- l'engagement 5, « désigner des administrateurs ministériels des données et accompagner la mise en œuvre du principe d'ouverture par défaut », encore en cours ;
- l'engagement 6, « renforcer la transparence des algorithmes et des codes sources publics », dont deux des sept actions sont encore en cours ;

[107] La qualité est définie selon quatre critères par le Portail européen des données : i) le suivi et les mesures en vigueur pour améliorer la qualité des métadonnées et le niveau de conformité en matière d'information correcte sur les licences (la France atteint un niveau de 97 %) ; ii) la conformité des métadonnées à la spécification DCAT-AP (niveau de 74 %) ; iii) son actualisation et sa complétude (80 %) ; iv) la mesure dans laquelle les données fournies sur les portails sont disponibles sous licence ouverte, dans un format lisible pour une machine, en utilisant un URI, et liées à d'autres données pertinentes (62 % pour la France, ce dernier critère étant le moins réalisé pour l'ensemble des pays, avec une moyenne de 53 %).

[108] Plan d'action de la France 2018-2020, 3 avril 2018. Disponible sur le site de l'OGP (https://www.opengovpartnership.org/documents/france-action-plan-2018-2020/)

- l'engagement 7, « accompagner les territoires dans la mise en œuvre du principe d'ouverture des données par défaut », dont seul une des quatre actions est réalisée (cf. *infra* sur l'ouverture des données des collectivités) ;
- l'engagement 8, « créer un laboratoire d'intelligence artificielle ouvert pour l'État », réalisé ;
- l'engagement 9, « ouvrir l'administration à de nouvelles compétences et accompagner les initiatives d'innovation ouverte au sein de l'État », avec notamment le programme d'entrepreneurs d'intérêt général (EIG) pour favoriser les usages de la donnée, dont quatre des dix actions sont encore en cours ;
- l'engagement 11, « améliorer la fluidité des données au sein de l'État avec FranceConnect plateforme », encore en cours ;
- l'engagement 18, « construire un écosystème de la science ouverte », dont neuf actions sont réalisées mais une en cours et une autre jugée difficile à réaliser.

Le bilan des objectifs du plan OGP 2018-2020 est très inégal selon les actions envisagées et la France accuse un retard important dans la réalisation de plusieurs d'entre elles. Cette évaluation est par ailleurs fragile méthodologiquement, la réalisation de chaque engagement étant fortement dépendante d'un retour déclaratif de la part des acteurs chargés de porter l'engagement, devant déclarer si l'engagement est réalisé ou non (plutôt qu'un taux de réalisation). À date, la mission a pu constater, en s'appuyant sur les échanges avec la direction interministérielle de la transformation publique (DITP) et avec les acteurs concernés comme les administrateurs ministériels de la donnée (AMD), que seuls trois des huit engagements en matière d'ouverture des données et des codes sources sont près de leur réalisation complète, à la veille du nouveau plan d'engagements. Certaines actions sont par ailleurs jugées difficiles à réaliser par les acteurs chargés de les mettre en œuvre, comme l'action consistant à « *recommander l'adoption d'une politique de données ouvertes associées aux articles et le développement des data papers, dans le cadre du soutien public aux revues* », portée par le ministère de l'enseignement supérieur, de la recherche et de l'innovation.

2.2. La dynamique d'ouverture a nettement ralenti depuis 2017 et seule une minorité d'acteurs publics se conforme à la loi pour une République numérique

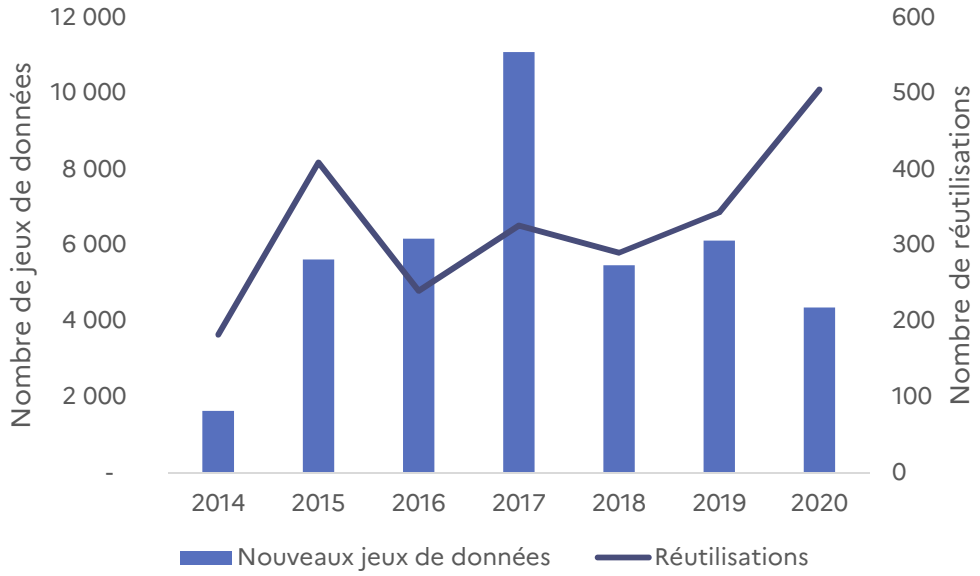
Le principe d'ouverture par défaut des données et codes sources publics prévu par la loi pour une République numérique a produit de premiers effets significatifs, mais n'est pas pleinement appliqué aujourd'hui. Ce constat est visible aussi bien dans l'analyse de l'activité des plateformes de mise à disposition de données et de codes sources publics, notamment la plateforme publique *data.gouv.fr* et le site Github, que dans les réactions des acteurs publics rencontrés par la mission mais aussi dans les réponses qu'ils ont apportées aux questions de la mission, lorsqu'ils se sont sentis concernés par la mission.

Une dynamique d'ouverture des données qui s'essouffle

Il n'est pas possible de dresser un bilan quantitatif exhaustif de l'ouverture. Il n'existe pas de recensement exhaustif à jour des données et des codes sources publics qui permettrait de mesurer le taux d'ouverture de ces derniers. Les dictionnaires et cartographies disponibles ne sont jamais exhaustifs et ne permettent pas non plus de conduire cette analyse au niveau d'un ministère ou d'un secteur par exemple. Cependant, une majeure partie de la dynamique d'ouverture peut être observée à travers l'analyse de l'activité de la plateforme *data.gouv.fr*, en raison du caractère centralisateur, quoique non exhaustif, de la plupart des données publiques. Ce choix introduit cependant un biais, dans la mesure où le portail *data.gouv.fr* tend à être délaissé au profit de portails locaux. Concernant les codes sources, la mission a retenu une analyse quantitative de l'activité des plateformes de dépôts, et a croisé cette analyse avec celle des réponses apportées par les acteurs publics ayant répondu au questionnaire de la mission (essentiellement des administrations centrales et des opérateurs, ce qui crée un biais de sélection).

L'ouverture des données a connu une dynamique forte de 2014 à 2017, avec un effet de la loi pour une République numérique visible en 2017, mais cette tendance est nettement retombée depuis pour ce qui est de la mise en ligne de nouveaux jeux de données. Le nombre de jeux de données créés par an est ainsi passé de 1 637 en 2014 à 4 362 en 2019, après un pic de 11 085 en 2017, tandis que le nombre de réutilisations est passé de 182 à 505 sur la même période.

Nouvelles données et réutilisations par année sur *data.gouv.fr* (2014-2020)



Source : Mission (analyse des données de *data.gouv.fr*, novembre 2020)

Il est difficile d'analyser ces statistiques, étant donné que le système de mesure de *data.gouv.fr* ne permet pas d'appréhender réellement les évolutions du nombre de dépôts et de réutilisations.

En effet, il est possible d'interpréter cette baisse du nombre de réutilisations comme la conséquence d'une « API-fication » croissante des données publiques, qui ne sont parfois plus accessibles et réutilisables dans un format libre, mais mises à disposition sous le contrôle du producteur de la donnée par le biais d'une API (cf. partie 3), ou bien comme une utilisation réduite de la plateforme *data.gouv.fr* au profit de portails locaux. C'est probablement le cas pour l'information géographique, le développement du portail spécialisé *geo.data.gouv.fr* ayant permis à cette communauté, contributrice majeure de l'ouverture, de trouver un lieu d'exploration et d'analyse des données en dehors de *data.gouv.fr*.

Depuis 2018, les collectivités représentent une part grandissante des jeux de données publiés chaque année sur *data.gouv.fr*[109], tandis que la part des administrations centrales continue de reculer, et celle des administrations déconcentrées, majoritaire entre 2015 et 2017, diminue fortement. La portée réelle de ces évolutions statistiques doit cependant être relativisée, compte tenu d'un effet de nombre, les collectivités étant plus nombreuses que les administrations centrales, et le nombre de jeux de données publiés n'ayant que peu à voir avec la dynamique d'ensemble d'une catégorie d'acteurs publics.

[109] Source : Datactivist (catalogue *data.gouv.fr*, Wikidata, recodage manuel, 24 septembre 2020).

Les données sont parfois ouvertes sans être rendues accessibles

Au-delà du volume de données, la qualité et l'accessibilité de l'offre présente également un bilan en demi-teinte. L'*open data* ne consiste pas seulement à ouvrir et à mettre à disposition des bases de données, mais à accompagner cette publication des informations et des outils indispensables à l'appropriation et à la bonne réutilisation de la donnée. L'accessibilité constitue aussi un enjeu de transparence et de pédagogie, permettant à des usagers non experts de prendre connaissance de la donnée disponible et de se familiariser avec son interprétation. Le portail *data.gouv.fr* remplit de fait une fonction d'exploration de la donnée qui n'est pas négligeable : 19 % des usagers interrogés ne cherchaient rien sur la plateforme en s'y connectant[110], ce qui souligne l'importance de l'activité éditoriale menée par Etalab sur *data.gouv.fr*.

Ce défaut d'accessibilité est le plus souvent lié à un défaut de qualité ou d'accompagnement de la publication du côté du producteur. De ce point de vue, l'*open data* initié en 2016 a besoin d'un nouvel élan, et d'une nouvelle maturité (cf. partie 3). Parmi les marges de progression figurent notamment la fraîcheur des données (52 % des jeux de données n'ont pas été mis à jour depuis plus d'un an[111]), la disponibilité de leurs séries temporelles, le catalogage et la capacité à avoir une vision panoptique de l'offre de données, en dépit de la multiplication des portails. L'insuffisance de la documentation est également un défaut majeur de l'*open data* tel que pratiqué depuis 2014 : 22 % des jeux de données ont une description de moins de 180 caractères (soit la taille d'un SMS) et 59 % font moins de 1 000 caractères (soit une demi-page)[112].

Mais certaines données peuvent aussi n'être pas exploitables en raison d'une volonté manifeste du producteur de limiter la réutilisation. Dans ces cas, la base est officiellement ouverte mais en pratique inexploitable, par exemple dans le cas d'une base de données impossible à interpréter, en l'absence d'identification des variables et de documentation suffisante. Plusieurs recours devant la CADA puis portés au tribunal administratif, dans le cas où le producteur de la donnée n'a pas souhaité communiquer les données, ont ainsi abouti à des transmissions de données tronquées ou indéchiffrables, selon les informations communiquées par plusieurs journalistes spécialistes de la donnée à la mission. Ces cas de publications intentionnellement incomplètes sont *a priori* limités et peuvent être résolus par une nouvelle intervention de la CADA puis du tribunal administratif, qui peuvent reconnaître dans ce cas le refus de communication[113].

Des administrations centrales très en retard et qui se sentent inégalement concernées

Les administrations centrales sont une faible minorité à avoir engagé la publication de leurs données. Les statistiques de la plateforme *data.gouv.fr* montrent que 12,6 % des jeux de données publiés le sont par des administrations centrales[114]. Au-delà de cet indice, ce sont surtout les auditions de la mission et les informations collectées auprès des administrations centrales[115] qui révèlent un manque profond d'appropriation de la politique de la donnée. Plusieurs administrations ont ainsi considéré ne pas être « concernées » par les sollicitations de la mission, et ont tardé à apporter une réponse malgré les relances de la mission. Une direction générale a par exemple répondu qu'« *[elle] ne produi[sait] pas de donnée* » car elle n'était pas un service statistique ministériel. Ces réactions sont révélatrices de la place que beaucoup de directions métiers réservent à la donnée dans le pilotage de leurs politiques.

---

[110] Enquête usagers conduite par Etalab de juin à septembre 2020.

[111] Source : Datactivist (catalogue *data.gouv.fr*, Wikidata, recodage manuel, 24 septembre 2020).

[112] Source : Datactivist (catalogue *data.gouv.fr*, Wikidata, recodage manuel, 24 septembre 2020).

[113] Audition de la CADA par la mission.

[114] Source : Datactivist (catalogue *data.gouv.fr*, Wikidata, recodage manuel, 24 septembre 2020).

[115] La mission a transmis un questionnaire au mois de septembre et d'octobre 2020 à plusieurs administrations centrales, opérateurs de l'État, notamment dans les principaux ministères concernés par la lettre de mission

Du reste, la crise sanitaire de la Covid19 a agi comme un révélateur du manque patent de données, pourtant nécessaire à la gestion de crise. Si la crise a eu pour effet d'accélérer la collecte et la remontée de données, il apparaît nécessaire de transformer l'essai pour replacer la donnée au cœur des politiques publiques pilotées par les directions, et non pas comme un sujet réservé aux services statistiques ministériels ou aux directions du numérique ministérielles.

Certains ministères sont plus avancés que d'autres, en particulier le ministère de la transition écologique et les ministères économiques et financiers. Cependant, au sein même de ces ministères, certaines directions centrales, y compris celles contribuant à des fonctions stratégiques majeures de la politique de l'État et exploitant de nombreuses données susceptibles d'être mises à disposition, continuent de ne pas considérer que la mise à disposition de données ou de codes sources fait partie de leurs missions ou qu'elle n'en constitue pas une priorité. Les données de gestion des administrations sont rarement envisagées comme un objet intéressant pour la société.

Le ministère de l'écologie fait également figure d'exception par la mise à disposition de nombreuses données de ses services déconcentrés, en particulier des directions départementales des territoires et de la mer (DDT-M). Pour ces données, il s'agit d'un moissonnage des catalogues Géo-IDE, plateforme développée par le ministère de l'écologie et le ministère de l'agriculture, et qui permet notamment le catalogage des données. Il s'agit principalement de données des plans locaux d'urbanisme (PLU) ou de plans d'occupation des sols (POS). Cette ouverture est directement liée à l'application de la directive européenne INSPIRE, qui s'applique notamment au droit des sols. Cependant, la mise à jour de ces données des services déconcentrés est en net recul depuis 2018, la part des jeux de données publiés sur *data.gouv.fr* ayant fortement reculé[116]. Outre l'essoufflement de la dynamique d'ouverture mentionné plus haut, deux facteurs techniques peuvent expliquer cette évolution plus spécifique : l'arrêt de la maintenance du service *geo.data.gouv.fr* qui moissonnait les infrastructures de données géographiques pour *data.gouv.fr* ; la montée en charge du géoportail de l'urbanisme où sont désormais stockés la majorité des PLU et POS, permettant aux DDT-M de ne plus stocker systématiquement ces données.

En outre, les administrations centrales mettent peu à jour leurs données, une fois publiées. Sur la plateforme *data.gouv.fr*, 60 % des jeux de données des administrations centrales n'ont pas été mis à jour depuis plus d'un an (alors qu'il s'agit de 52 % des jeux de données en moyenne, tous producteurs confondus), et 35 % ne l'ont pas été depuis plus de quatre ans[117].

Ce retard de l'administration centrale est largement dû à un manque de portage politique et administratif, à la fois au niveau des objectifs fixés par les ministres et de l'action du réseau des administrateurs ministériels de la donnée. Laissée à l'initiative des administrations, la politique de l'ouverture des données publiques ne parvient pas à prendre son essor, faute d'être comprise dans ses finalités. Même les administrations qui sont engagées dans la démarche, comme la DGFiP avec la base DVF, ne semblent pas avoir cherché à analyser les réutilisations et à comprendre les effets de cette publication : ainsi, après y avoir participé de manière sporadique, la DGFiP n'est plus présente régulièrement dans les deux instances d'animation de la communauté des réutilisateurs (le groupement national de la base DVF, GnDVF, et le laboratoire d'initiatives foncières et territoriales innovantes, le LIFTI).

Quant au ministère des solidarités et de la santé, il dispose d'une masse de données importante contenant pour majeure partie des données à caractère personnel, dont certaines sensibles, telles que les données de santé. Si une bonne partie d'entre elles relevant de l'*opendata* est d'ores et déjà accessible, des travaux ont été menés pour définir un cadre de partage des données de santé au travers de la mise en place du *Health Data Hub* (cf. partie 3).

(ministères sociaux, ministères économiques et financiers, ministère de la transition écologique, ministère de l'enseignement supérieur et de la recherche, ministère de l'éducation nationale, ministère de l'intérieur) ainsi qu'aux administrateurs ministériels de la donnée des ministères pour lesquels une collecte d'informations plus complète n'était pas possible (affaires étrangères, agriculture, culture, défense, justice, notamment).

[116] En 2020, 18 % des jeux de données des administrations déconcentrées ont été publiés depuis moins de 3 ans (source : Datactivist).

[117] Source : Datactivist (catalogue *data.gouv.fr*, Wikidata, recodage manuel, 24 septembre 2020).

Parmi les opérateurs de l'État, la situation est très hétérogène selon les secteurs, certains présentant parmi les situations les plus avancées de mise à disposition des données, comme Pôle Emploi ou les opérateurs de l'environnement, et d'autres étant encore à l'écart de la démarche. Dans le domaine de la santé, par exemple, plusieurs opérateurs ont adopté des stratégies et des gouvernances propres, à l'image de Santé publique France ayant créé une direction de la donnée, un comité spécialisé et des procédures internes pour la mise à disposition des données en *open data* et à destination des chercheurs.

Au sein des collectivités, une dynamique portée par celles qui bénéficient d'un projet et qui en ont la capacité

Les collectivités territoriales de plus de 3 500 habitants et employant plus de 50 agents (en équivalent temps plein) sont soumises aux obligations d'ouverture des données et des codes sources énoncées par la loi pour une République numérique. En pratique, cette obligation concerne environ 4 600 collectivités locales (communes, EPCI, départements et régions)[118].

Plusieurs collectivités ont souligné l'importance de la place de la donnée dans le pilotage des politiques locales. Ainsi, un élu souligne *qu'« il faut penser la question de la donnée comme un facteur de développement territorial et pas nécessairement comme un acte de modernisation nationale ».* Néanmoins, la mise en œuvre des obligations prévues par la loi pour une République numérique montre des disparités importantes entre les collectivités territoriales, qui s'expliquent principalement par les moyens dont elles disposent.

Pour mettre en œuvre cette obligation, les collectivités peuvent être accompagnées par l'association Opendata France : créée le 9 octobre 2013 à Toulouse, elle regroupe et soutient les collectivités engagées activement dans une démarche d'ouverture des données publiques et favorise toutes les démarches entreprises par ces collectivités dans le but de la promotion de l'open data. Pour remplir sa mission, Opendata France est financée principalement par des subventions et, pour une part croissante, par les adhésions versées par les collectivités territoriales. Ainsi, les recettes de l'association étaient composées, à fin 2019, de 67 % de subventions d'exploitation pour un montant de 182 500 € (dont 71 % versées par l'État – au travers d'Etalab – et 18 % par la Banque des territoires de la Caisse des dépôts et consignations) et de 32 % de cotisations des collectivités.

Plus précisément, Opendata France a pu bénéficier de subventions de la part de l'État dans une logique de « projet » avec des objectifs à atteindre : à savoir, un budget de 200 000 €, au titre des mesures d'accompagnement de la loi pour une République numérique, puis un budget de 160 000 €, affecté à la conception et la réalisation d'un outil de validation des données publique (Validata), et enfin, un budget de 120 000 € pour le concours à l'entretien des dispositifs de soutien des collectivités. L'association regrette ainsi l'interruption des soutiens financiers de l'État en 2020, et son effet sur l'évolution des ressources et des projets à destination des collectivités, en particulier des plus petites. Une contribution financière de la part des associations des collectivités pourrait aussi répondre aux besoins.

Quant à l'Agence nationale de la cohésion des territoires (ANCT), elle n'a pas encore investi cette politique, considérant que la politique d'ouverture des données et des codes sources relève de la mission d'Etalab. Créée par la loi du 22 juillet 2019 et mise en place au 1er janvier 2020, l'ANCT a pour mission d'accompagner les collectivités territoriales dans la mise en œuvre de leurs projets de politique publique. En revanche, l'ANCT souhaite davantage se positionner sur une offre de services aux collectivités territoriales, visant à développer la culture de la donnée, en l'orientant sur le service aux usagers et le développement économique. Ainsi la start-up d'Etat « bases adresses locales » incubée au sein de l'incubateur des territoires de l'ANCT propose un service aux collectivités afin de faire un lien entre leurs bases adresses locales et la base adresse nationale figurant parmi les données de référence du service public de la donnée. Elle utilise pour cela les ressources développées au sein d'Etalab.

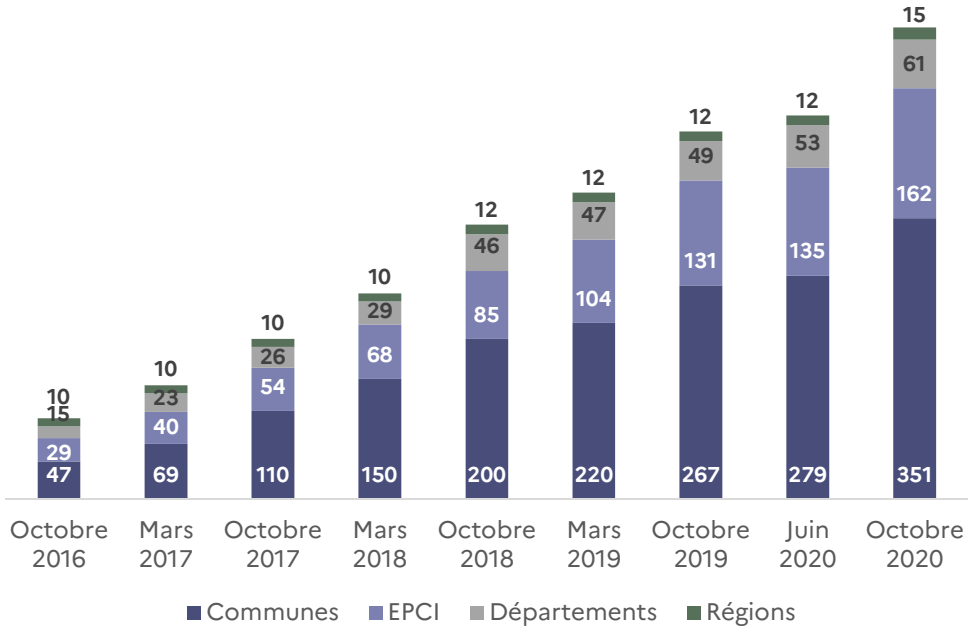
S'agissant plus particulièrement de la problématique des données personnelles, la CNIL a publié un certain nombre de fiches pratiques à destination des collectivités territoriales afin de faciliter la mise en œuvre de cette politique.

[118] Source : Observatoire Opendata des territoires

L'observatoire Opendata des territoires, co-financé par la Caisse des dépôts et consignations, Etalab (jusqu'en 2019) et l'association Opendata France, évalue l'ouverture des données dans les collectivités territoriales sur les dimensions quantitatives et qualitatives. L'association publie ainsi chaque année une analyse de l'état de l'ouverture des données dans les territoires qui s'appuie sur les chiffres et les indicateurs, collectés et élaborés en temps réel, et publiés sur le site de l'observatoire. Les résultats publiés en octobre 2020 sur le site d'Opendata France[119] montrent une vraie dynamique dans la mise en œuvre de la politique d'ouverture des données.

En octobre 2020, 589 collectivités territoriales publiaient des données en *open data*. En ajoutant 175 structures publiques ou privées, partenaires des collectivités et produisant des données dans le cadre d'une mission de service public, on compte 764 acteurs territoriaux engagés dans cette dynamique. Le graphique ci-dessous montre l'évolution du nombre de collectivités territoriales engagées dans une démarche d'*open data*, selon le type de collectivité (régions, départements, EPCI et communes).

Évolution du nombre de collectivités engagées dans une démarche d'*open data*



Source : Observatoire Opendata des territoires, octobre 2020

La dynamique d'ouverture des données est cependant différente selon le type et la taille des collectivités territoriales : plus la taille de la collectivité est petite, moins elle est engagée dans une démarche d'ouverture des données, ce qui s'explique principalement par les moyens nécessaires à la mise en œuvre d'une telle politique.

L'ensemble des régions proposent désormais des jeux de données en *open data*, même si cela ne doit pas masquer des disparités dans la quantité de jeux de données mise à dispositions, « de quelques unités à plusieurs centaines de jeux de données » selon Opendata France. **Les départements quant à eux poursuivent leur dynamique d'ouverture des jeux de données** : 60% d'entre eux sont ainsi engagés dans cette démarche.

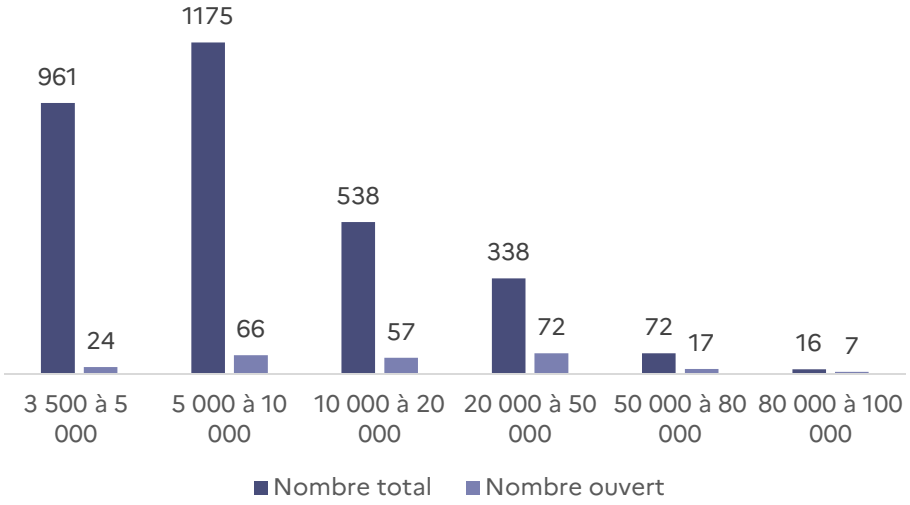
S'agissant des communes et des EPCI, près de 90 % des collectivités ne respectent pas leurs obligations (cf. graphique) :

- plus de 50 % des collectivités de plus de 100 000 habitants sont ouvertes (ce pourcentage intègre les métropoles ouvertes à 82 % et les plus grosses intercommunalités) ;

[119] https://www.opendatafrance.net/2020/10/15/observatoire-open-data-des-territoires-edition-oct-2020/

- **en dessous de 100 000 habitants, moins de 8 % collectivités publient des données.** En particulier, seules 22 % des collectivités de taille entre 20 000 et 80 000 (soit 89 sur 410) sont ouvertes ;
- s'agissant des communes et EPCI de moins de 3 500 habitants, 86 collectivités sont comptabilisées, ce qui témoigne soit d'une volonté locale très importante compte tenu de l'absence d'obligation légale les concernant, soit d'un accompagnement systématique par une structure de plus haut niveau (métropole, SMI ou département).

Répartition des communes de moins de 100 000 habitants mettant en œuvre une politique d'*open data* (en octobre 2020)



Source : Observatoire Opendata des territoires, octobre 2020

In fine, selon les personnes rencontrées par la mission, ces résultats mitigés pour les collectivités d'une taille plus réduite s'expliquent autant par un manque d'acculturation numérique, la question de l'*open data* étant parfois vécue comme une obligation légale qui « tombe du ciel », que par un véritable besoin d'un accompagnement et de mutualisation des moyens.

Le manque de mutualisation se retrouve dans le dénombrement des plateformes de publication. Ainsi, l'observatoire dénombrait 180 plateformes de publication territoriales en octobre 2020 dont 6 % sur un portail web « classique », 29 % sur une plateforme d'information géographique, et 64 % sur un portail *open data* dédié ou mutualisé. La multiplication des portails reflète aussi l'enjeu de visibilité et de communication qu'ils représentent pour la collectivité productrice des données. La plateforme *data.gouv.fr* reste la principale plateforme d'hébergement avec 41 % des collectivités, suivie des plateformes sous technologie propriétaire comme OpenDataSoft (28 %) ou libre comme CKAN (14 %).

Par ailleurs, la majorité (64 %) des acteurs publient leurs données sur un portail propre, témoignant du faible effort de mutualisation des moyens des collectivités territoriales néanmoins en augmentation par rapport à 2019. Ainsi, 36 % des collectivités ont adopté des plateformes mutualisées (29 %) ou en partenariat avec des services déconcentrés de l'État (7 %), permettant de réduire les coûts et les délais tout en favorisant la visibilité et l'interopérabilité des données qu'elles publient.

Plusieurs régions se sont cependant lancées dans une démarche de mutualisation des ressources sur leur territoire ou de rassemblement des acteurs autour de la donnée : on peut notamment citer les démarches des régions Occitanie, Bretagne et Normandie.

La région Occitanie fait preuve d'un grand dynamisme dans la mise en œuvre de la politique d'ouverture des données, tirée par deux grandes métropoles (Toulouse et Montpellier) et de trois départements (Haute-Garonne, Hérault, Hautes-Pyrénées) qui accompagnent l'ouverture des données de petites collectivités dans leur périmètre de compétence. Ainsi, un système de subsidiarité à trois étages a été mis en place pour que les petites communes parviennent à mettre en œuvre une politique d'*open data* : la région soutient les départements, le département soutient les communes membres en publiant leurs données et l'EPCI (Toulouse par exemple) publie les données des communes, membres de l'EPCI. En effet, l'enjeu est aussi de ne pas s'arrêter aux frontières administratives afin d'avoir la couverture la plus complète d'un territoire : il s'agit d'avoir une couverture des territoires urbains, péri-urbains et ruraux, permettant de développer des modèles économiques utilisant ces données.

La région Normandie constitue un autre exemple intéressant de mutualisation au travers du Datalab Normandie qui rassemble, au sein d'un consortium, un écosystème d'acteurs, aussi bien publics que privés, pour utiliser de la donnée à travers une plateforme d'échange et de partage. Ce programme d'action bénéficie d'un financement européen (2,5M€ issus du FEDER). Différentes actions et expérimentations ont été lancées telles que la création d'un « data emploi formation » pour faciliter le pilotage sur l'appareil de formation qui est une compétence de la région, et des actions avec l'agence de l'orientation pour la promotion des métiers de la donnée.

La région Bretagne promeut une approche partenariale à l'échelle régionale entre acteurs publics (GéoBretagne, Service Public Métropolitain de la Donnée de Rennes, DATARMOR-Pôle de Calcul et de Données pour la Mer, Portails Open Data, EskemmData…) pour contribuer collectivement au service public de la donnée et s'engager dans une démarche de régulation de l'écosystème du territoire au travers des actions suivantes[120] :

- « définir les cas d'usages communs prioritaires et les données de référence associées ;
- « proposer, en complément, des services d'analyse, de prospective et de croisement de données ;
- « investir dans des architectures techniques partagées proposant : la mise à disposition de données fondées sur des standards en API et en téléchargement, des mécanismes de participation des utilisateurs à la montée en qualité des données et dotées de mécanismes d'identification, de sécurisation et de contrôle d'accès pour certains usages ;
- « intégrer à l'échelle régionale de nouvelles formes de régulation de l'économie et d'aménagement du territoire par la donnée. »

Il convient par ailleurs de souligner la montée en puissance des portails nationaux thématiques, alimentés par les acteurs territoriaux qui publient soit localement, soit sur un site national mutualisé, des données normalisées dans l'objectif d'un référencement maximal. À titre d'exemple, le portail *transport.data.gouv.fr* agrège de nombreuses données de mobilité, produites par les collectivités et leurs délégataires (autorité organisatrice de la mobilité, AOM) : près de 218 jeux de données sur les 337 AOM françaises, soit 64 %, étaient référencées en octobre 2020 pour les données de transport en commun. Au 4 décembre 2020, 10 159 collectivités ont déposé leurs documents d'urbanisme (PLUi, PLU, cartes communales) sur le Géoportail de l'urbanisme mis en place par l'ordonnance n°2013-1184 actée par la loi ALUR.

[120] Feuille de route de la Région Bretagne en faveur de stratégies numériques « S'engager en faveur de stratégies numériques responsables pour le territoire breton », février 2020.

2.3. L'ouverture des codes sources, une démarche encore confidentielle et qui manque de visibilité sur les réutilisations

L'ouverture des codes sources, un mouvement encore confidentiel dans l'État, malgré un progrès depuis 2015

Plus encore que pour les données, l'analyse de la dynamique d'ouverture des codes sources publics est difficile à mesurer, en l'absence d'un recensement exhaustif et d'une métrique bien établie de leurs réutilisations. La plateforme de recensement existante, *code.etalab.gouv.fr*, bénéficie d'une moindre attractivité que celle de *data.gouv.fr*, en l'absence de services pour répondre aux besoins des développeurs (hébergement des codes sources, suivi des versions, suivi des retours utilisateurs) et aucun mécanisme n'existe aujourd'hui pour faire connaître de manière certaine les codes ouverts à la DINUM, chargée de leur valorisation. Ainsi, la plateforme privée github.com héberge 61% des codes sources publics. code.etalab.gouv.fr joue un rôle de portail d'accès à tous les dépôts de code source pour l'ensemble des plateformes, privées (github.com, gitlab.com), associatives (framagit.org, gitlab.adullact.net et gitlab.ow2.org) ou publiques (21 instances du logiciel libre GitLab, la plupart installées dans l'enseignement supérieur et la recherche, à l'exception notable du Grand Lyon, qui a déployé forge.grandlyon.com)[121].

La plateforme *code.etalab.gouv.fr* recense au 8 décembre 2020 un total de 5 679 dépôts, dont 29 % liés à la recherche (1 644 dépôts). Parmi les principaux contributeurs, outre *beta.gouv.fr* et Etalab, qui figurent en tête, on peut mentionner le Médialab de Sciences Po (176 dépôts), les applications métiers LUTECE destinées au secteur public (131 dépôts), la Fabrique numérique des ministères sociaux (118 dépôts), le portail Esup (107 dépôts).

L'ouverture des codes publics a nettement progressé depuis 2015, mais reste confidentielle, concernant une minorité d'acteurs publics. La mission a en effet complété l'analyse de la plateforme *code.etalab.gouv.fr* des informations collectées auprès des administrations centrales et des opérateurs par le biais de son questionnaire, permettant d'obtenir un premier indice de la dynamique créée par la loi pour une République numérique de 2016, même si l'échantillon retenu est loin d'être représentatif. Sur un échantillon de dix-neuf acteurs[122], le nombre total de codes sources publiés est passé de 5 en 2015 à 66 en 2020 (dont 31 pour l'INSEE et 10 pour Fun-Mooc). Il est intéressant de noter que l'INSEE attribue son choix d'une « *philosophie explicitement Open Source* » depuis 2015 à « *un mouvement qui se consolide peu à peu à l'échelle du système statistique européen* ».

De manière plus significative, sur ce même échantillon, le nombre d'acteurs à n'avoir publié aucun code était de 16 en 2015, mais n'était plus que de 8 en 2020. Il est cependant impossible de généraliser ce constat pour l'ensemble du secteur public car cet échantillon comporte un biais majeur de sélection, les acteurs ayant répondu l'ayant fait généralement pour valoriser leur démarche dans ce cadre.

Seulement 31% des 5655 dépôts ont une licence identifiée par GitHub ou par GitLab. Même si des dépôts peuvent avoir une licence qui reste non identifiée automatiquement, cela indique néanmoins qu'une majorité de dépôts sont publiés sans licence. Pour ceux qui sont publiés avec licence, **la licence la plus utilisée est la licence MIT (30,2 % des codes du site)**[123], suivie par la licence GNU Affero General Public License v3.0 (16,5%). Les licences Creative commons représentent une minorité, avec 0,5 % des codes pour la Creative Commons Attributions Share Alike 4.0 International, et 0,3 % pour la Creative Commons Attribution 4.0 International.

---

[121] Sur les 5655 dépôts, 4072 sont hébergés sur github.com (61%) et 1583 sur gitlab.com ou des instances GitLab (39%). Source : code.etalab.gouv.fr. framagit.org est un service opéré par l'association Framasoft reposant sur une instance GitLab. Github est le service le plus utilisé dans le monde, opéré par Microsoft qui l'a racheté en 2018.

[122] La mission a reçu la réponse à un questionnaire de la part de 48 directions d'administration centrale et opérateurs mais la question des codes sources a été une de celles qui ont été le moins renseignée par les acteurs, soit qu'elle ait été laissée sans réponse, soit que le nombre de dépôts n'ait pas été restitué dans le temps, ne permettant pas une analyse de l'évolution dans le temps.

[123] Statistiques du site *code.etalab.gouv.fr*. Ce constat est confirmé par les informations transmises par les administrations sollicitées par la mission.

S'il n'est pas possible de mesurer globalement et avec exactitude la réutilisation des codes sources, cela reste possible pour une partie d'entre eux, ceux qui correspondent à des bibliothèques logicielles (ou *libraries* en anglais). Ainsi, par exemple, pour les données concernant le découpage administratif français (communes, arrondissements, départements…), on mesurait en novembre 2020 environ 200 téléchargements en moyenne par semaine.

Comme pour la donnée, l'ouverture d'un code ne garantit pas à elle seule sa potentielle réutilisation, comme en témoigne l'exemple du code de l'impôt sur le revenu de la DGFiP ou celui des modèles de la direction générale du Trésor (cf. encadré). Les codes sont parfois mis à disposition dans un langage difficilement accessible et nécessitent d'être transposés dans un langage plus connu (Scilab, R) pour pouvoir s'améliorer au contact des retours des utilisateurs.

L'effort de mise à disposition du code peut être jugé disproportionné et justifier une absence de communication à date, selon la jurisprudence de la CADA. Dans le cas des codes sources employés par la Caisse nationale des allocations familiales (CNAF) par exemple, la CADA a indiqué dans un avis n° 20181891 du 18 juillet 2019 que, « *compte tenu de l'ancienneté et de la complexité du système d'information actuel, l'extraction des codes sources n'était pas techniquement possible sans effort disproportionné* ». La CADA indique cependant qu'elle prend acte des travaux de redéfinition du modèle de gestion des prestations qu'entreprend la CNAF, qui doivent permettre de répondre à la demande.

L'ouverture des modèles économiques de la direction générale du Trésor

La direction générale du Trésor a mis en ligne en septembre 2018 les trois modèles économiques qu'elle utilise pour ses analyses (Saphir, Mésange et Opale), à la suite d'une saisine en 2017 de l'association « Ouvre-boîte » et après un avis de la CADA du 19 avril 2018. De l'avis de la direction, cette ouverture constitue « une évolution substantielle consécutive à la loi de 2016 ».

Comme la plupart des acteurs publics, la direction générale du Trésor a fait le choix de publier les codes sources en téléchargement direct, mais aussi sur deux plateformes d'hébergement gratuites (Framagit et GitHub) pour conserver une traçabilité des utilisations et des contributions. La direction juge « l'écho médiatique relativement faible », surtout repris sur des sites spécialisés.

La direction a accompagné l'ouverture des codes d'une publication expliquant la démarche d'évaluation macroéconomique des politiques publiques[124]. Les modèles nécessitent aussi des compétences spécifiques pour être exploités (connaissance du cadre de la comptabilité nationale pour Opale et Mésange, connaissance des aides sociales ou des barèmes des impôts pour Saphir). Le véritable frein à une démarche d'enrichissement du code et de réutilisations réside en fait dans le langage utilisé dans les codes. La direction du Trésor a ainsi engagé un travail de transposition des modèles dans un langage plus connu (Scilab, R).

[124] Direction générale du Trésor, Trésor Éco n°252, « Le recours à la modélisation macroéconomique dans l'évaluation des politiques publiques ».

Les informations communiquées à la mission montrent que les acteurs n'ont pour l'immense majorité pas de stratégie pour décider de l'ouverture de leurs codes sources, y compris ceux qui ont une stratégie claire d'ouverture de leurs données. Le choix de l'ouverture est plutôt le résultat d'opportunités ou de sollicitations externes (comme une action associative dans le cas des modèles du Trésor). En outre, une stratégie en matière de codes sources ne pose pas uniquement la question de leur ouverture, mais parfois celle de leur maintenance et de la gouvernance. Seules quelques institutions font désormais le choix de concevoir l'ouverture *ab initio*, comme l'INSEE, qui dit concevoir « nativement » cette ouverture, dès la conception ou la refonte du modèle économique, et non plus seulement *a posteriori*[125].

Plusieurs institutions font valoir dans leur réponse à la mission que l'ouverture des codes sources est un facteur de vulnérabilité des systèmes d'information (cf. partie 1, titre 3). Pourtant, deux acteurs aux activités particulièrement sensibles figurent parmi les plus engagés dans la démarche, et parmi ceux dont l'activité suscite le plus d'intérêt : le Commissariat à l'énergie atomique et aux énergies alternatives (CEA) et l'Agence nationale de la sécurité des systèmes d'information (ANSSI). Le site *code.etalab.gouv.fr* permet en effet d'avoir un indice de l'intérêt pour les utilisateurs d'un code source public, qui peuvent indiquer cet intérêt sous forme d'étoile (équivalent d'un favori ou d'un marque-page)[126]. Selon ce critère, le CEA et l'ANSSI sont les deux organisations productrices qui recueillent le plus d'intérêt, avec 4 757 et 3 522 « étoiles » respectivement (pour l'ANSSI, la plateforme dédiée OpenCTI figure également parmi les plus suivies, avec 1 476 étoiles).

Par ailleurs, l'ouverture du code peut être freinée par une peur d'être jugé dans son action. Cela vaut d'abord lorsque l'ouverture du code révèle avec lui un algorithme utilisé par la puissance publique, qui reflète un choix, qui peut être contesté et remis en cause. C'est également vrai sous un angle technique, les choix d'écriture du code pouvant être questionnés lors de l'ouverture et donner le sentiment à la structure d'être « fragilisée ». Un développeur a ainsi répondu à une enquête conduite par Etalab en 2020 qu'il avait publié « *une version beta très maladroite (mais qui fonctionne)* » et avait obtenu l'accord de sa direction pour la publier sur GitHub « *afin de tester ce genre de démarche* », mais n'avait pas pu la valoriser au-delà, faute de temps et de soutien. De manière plus révélatrice encore, l'absence de compréhension des enjeux par la hiérarchie des développeurs est révélée par une autre citation de cette enquête : ***« ce partage est vu par ma hiérarchie comme un potentiel point noir qu'on pourrait nous reprocher alors que j'y vois des cerveaux prêts à m'aider lorsque je ne suis pas sûr de ce que je fais. »***

L'ouverture des codes sources, un levier de développement dont s'emparent certaines collectivités

Certaines collectivités territoriales ont manifestement bien compris l'intérêt de la démarche d'ouverture des codes et d'utilisation des logiciels libres, et les gains à attendre d'une mutualisation du développement des outils informatiques, par l'ouverture des codes sources et le recours à des logiciels libres. Le mode de fonctionnement des collectivités, nombreuses et partageant des problématiques de mise en œuvre des politiques publiques très similaires, est probablement une explication à cet engagement, même s'il n'est pas partagé par l'ensemble d'entre elles.

[125] L'INSEE fait référence à sa stratégie publique « INSEE 2025 », qui ne mentionne toutefois pas explicitement l'ouverture des modèles et des codes sources et la stratégie dans ce domaine (« Pour remplir au mieux ses missions, l'Insee doit cultiver l'ouverture, aussi bien en direction des utilisateurs que de la communauté universitaire et scientifique. Dans ses orientations stratégiques à horizon 2025, l'institut veillera également à renforcer sa contribution dans l'enseignement et la recherche en matière de production et d'analyse statistiques. »)

[126] Les étoiles indiquées sur code.etalab.gouv.fr sont uniquement celles relevés via l'API de GitHub. L'intérêt mesuré n'est donc pas global sur l'ensemble des dépôts. Il faudra pour cela une métrique à part, développée aux seules fins de code.etalab.gouv.fr.

Comme pour l'*open data*, c'est une association qui joue un rôle majeur dans la promotion de cette démarche : l'ADULLACT, association des développeurs et utilisateurs de logiciels libres pour les administrations et les collectivités territoriales, fondée en 2002, qui rassemble 289 adhérents représentant 15 000 collectivités territoriales, administrations publiques et centres hospitaliers[127]. En 2019, 7 régions sur 15 étaient adhérentes à l'association, et 35 départements. 41 % des adhérents sont des communes.

L'ADULLACT met notamment à disposition une forge, c'est-à-dire un site de développement coopératif, dont l'objectif est de centraliser les développements du secteur public. La forge a notamment été financée par le Fonds européen de développement régional (FEDER). Pour mémoire, un tel outil n'existe pas au niveau national spécifiquement pour les administrations d'État. L'ADULLACT a également développé depuis 2016 un *« comptoir du libre »*, qui est une *« place de marché destinée à faciliter la recherche et l'adoption d'un logiciel libre »*, mettant en relation collectivités utilisatrices et prestataires.

Comme au niveau de l'État, les collectivités non engagées dans la démarche font souvent valoir des enjeux de sécurité ou bien l'éventuel débat que pourrait susciter la transparence sur leurs outils. Une collectivité interrogée sur la publication de son algorithme fondant l'attribution des places en crèches invoque la sensibilité du sujet pour ne pas le publier et le fait que cela nécessiterait une clarification des critères d'attribution de la part des élus et une acculturation, à la fois des élus mais aussi des agents administratifs pour qu'ils comprennent le gain qu'ils peuvent en retirer. Elle souligne également la nécessité de pouvoir adapter les outils à chaque territoire.

À cet égard, la publication récente du code source « Investissement Social dans l'Accueil du Jeune Enfant » (ISAJE) par la CNAF constitue une démarche intéressante allant dans le sens de la mutualisation des outils pour les communes. Ce code source a été développé dans le cadre d'une recherche[128] sur l'impact des crèches sur les compétences sociocognitives des enfants et utilisé par une communauté de communes (Valence Romans Agglo) pour l'affectation de ses places en crèches. Le code source utilisé par l'algorithme a été mis à disposition de la commune et en ligne. Une procédure a également été construite afin de permettre aux parents de comprendre les décisions d'attribution des places en crèche.

La démarche de science ouverte de l'enseignement supérieur et de la recherche

L'objectif de promouvoir l'ouverture des données et des codes sources dans l'enseignement supérieur et la recherche a été clairement affiché dans le Plan national pour la science ouverte présenté en juillet 2018 par la ministre de l'enseignement et de la recherche (cf. partie 1, titre 2).

Dans cette continuité, on peut citer plusieurs initiatives. L'agence de mutualisation des universités et des établissements (AMUE) est en train d'évoluer vers le logiciel libre traduisant son rapprochement avec le consortium Cocktail. Plusieurs actions ont été conduites en ce sens : des actions de sensibilisation de la communauté de l'enseignement supérieur et de la recherche au travers de publications et d'un séminaire (avec l'organisation d'une journée dédiée au sujet Open Data pour l'ensemble de la communauté Enseignement Supérieur et Recherche en novembre 2019).

Le centre national de la recherche scientifique (CNRS) s'est doté en 2019 d'une feuille de route pour la science ouverte s'appuyant sur des actions concrètes structurées autour de quatre grands objectifs (100 % de la production scientifique en accès ouvert, développement d'une culture de la gestion et du partage des données, développement d'infrastructure pour la fouille et l'analyse des contenus et la transformation des modalités d'évaluation des chercheurs).

[127] ADULLACT, Dossier de presse 2019.

[128] Projet de recherche sur l'impact d'un accès en EAJE sur le développement des enfants et les conditions de vie des familles, Arthur Heim (Chef de projet à la CNAF et doctorant à l'école d'économie de Paris) et Julien Combe (Professeur assistant à École Polytechnique), juin 2020

Du côté de l'institut national de recherche en sciences et technologies du numériques (INRIA), l'équipe InriaSoft a été mise en place pour diffuser les logiciels libres. Son rôle est d'augmenter l'impact de la production logicielle en favorisant son utilisation par des partenaires externes, académiques ou industriels. Ses missions consistent à :

- organiser la construction de consortiums autour des logiciels issus de la recherche afin de structurer la coopération entre les chercheurs et les utilisateurs ;
- organiser le développement logiciel afin de favoriser la participation des utilisateurs et la prise en compte de leurs besoins ;
- assurer la pérennité des logiciels afin de garantir aux utilisateurs leur maintenance sur le long terme ;
- favoriser l'émergence de start-up dédiées à l'exploitation d'un logiciel et des services spécifiques qui en résultent.

Par ailleurs, l'INRIA a développé une offre en matière de formation continue sur les technologies numériques dédiée aux logiciels libres au sein d'Inria Academy. L'institut s'est donné pour objectif d'accompagner le développement numérique, en partageant avec le plus grand nombre les logiciels libres.

Dans le domaine de la santé, l'institut national de la santé et de la recherche médicale (Inserm) a positionné la science ouverte comme une priorité de son plan stratégique. En effet, l'institut considère que l'ouverture des données peut influer sur la visibilité de ses activités et de ses publications. Par exemple, le portail Orphanet et la base partagée Orphadata qui est labellisée comme base de référence en *open data* par l'infrastructure Elixir à l'échelle européenne contribue largement à la reconnaissance de l'Inserm en Europe dans le domaine des maladies rares. Dans sa réponse au questionnaire de la mission, l'Inserm souligne ainsi que *« le partage des données est indéniablement un levier du partage de la connaissance et apporteur de valeur notamment par la transversalité qu'il peut induire ainsi que la reproductibilité, et donc la capacité de validation des résultats de recherche »*.

À l'Inserm, le partage de codes source se fait très largement dans une logique de co-développement des outils qui s'inscrit au sein d'une communauté de recherche dès lors que tous les acteurs jouent leur rôle de contributeur aux codes sources. Cela concerne la plupart des équipes de recherche académiques et certains industriels. Les acteurs qui ont une logique de « capture » des codes sources ou de non-respect d'un usage partagé ne peuvent pas bénéficier de ce système : cela peut être le cas de certains industriels qui ont une politique de propriété intellectuelle basée sur le secret ou d'acteurs ayant une politique de sécurisation des outils particulière, ce qui peut être le cas de certains acteurs de la santé.

Le commissariat à l'énergie atomique et aux énergies renouvelables (CEA), l'agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (ANSES) et d'autres organismes de recherche ayant répondu au questionnaire de la mission se positionnent de même sur la stratégie de science ouverte, selon le principe européen « *aussi ouvert que possible, aussi fermé que nécessaire* ».

3. Renforcer le portage politique et la gouvernance

3.1. La nécessité d'un portage politique fort

Tirer les leçons des expériences du Royaume-Uni et de l'Estonie

Les expériences internationales soulignent la nécessité d'un portage politique fort, ainsi que la fragilité des acquis des politiques de numérisation de l'État et d'ouverture des données et codes sources publics.

Ainsi, l'Estonie qui fait partie des trois nations numériques les plus avancées d'après l'enquête sur les E-gouvernements réalisée par l'ONU en 2020 en se reposant notamment sur le X-road, une infrastructure permettant l'échange de données entre les administrations et les citoyens et au sein même des administrations, n'a pas engagé une démarche soutenue d'ouverture des données et des codes sources publics, du fait notamment d'un manque de portage politique. Dans ce domaine, le pays se classe 24e en 2019 dans l'indicateur *OURdata* réalisé par l'OCDE. Si une politique d'ouverture a bien été engagée en 2012 au moment où l'Estonie a rejoint le Partenariat mondial pour un Gouvernement Ouvert, l'absence d'utilité d'une politique d'ouverture des données perçue par la classe politique n'a pas permis de soutenir cette volonté d'ouverture des données.

Au Royaume-Uni, alors qu'une politique volontariste d'ouverture des données avait été mise en place à la suite du *Freedom of Information Act (2000)*, lui permettant d'atteindre la 2e place, devant la France 4e, à l'indice mondial *Global Open Data Index* de 2017, le Royaume-Uni a récemment vu sa position dans les classements se dégrader. Ainsi, le classement de la Commission européenne le classe 11e en 2018 (contre 4e en 2015)[129] et l'*OURdata Index 2019* place le Royaume-Uni à la 20e position (contre 4e en 2017)[130]. La Commission européenne a ainsi noté qu'il n'y avait pas de stratégie prédéfinie pour maintenir à jour les données ouvertes, et que les utilisateurs avaient des difficultés à comprendre ou même à prendre connaissance des licences pour les données ouvertes. La dégradation de la place du Royaume-Uni dans le classement de l'OCDE est liée au 3e pilier de l'indice qui concerne le « soutien du gouvernement à la réutilisation des données ». Le rapport de l'OCDE relève que la politique d'ouverture des données n'a pas été maintenue dans l'agenda politique et que les efforts pour associer les réutilisateurs dans la politique d'ouverture des données ont été réduits.

Une avance dans la politique de gouvernement numérique dans le cas de l'Estonie ou dans la politique d'ouverture des données dans le cas du Royaume-Uni n'a pas suffi, en l'absence de portage politique fort, à développer une politique d'ouverture des données et des codes sources publics efficace.

[129] *Open Data Maturity in Europe*, Report 2018, Commission européenne.

[130] OCDE *OURdata Index 2019*

Suivre la politique publique de la donnée au niveau interministériel

Le portage politique de la politique d'ouverture et de circulation des données et des codes sources publics n'est pas clairement défini, et son suivi budgétaire n'est pas à la hauteur des enjeux.

La politique d'ouverture et de circulation des données a évolué dans les décrets d'attribution des différents ministères. En 2018, le décret d'attribution du secrétaire d'État chargé du numérique[131] mentionnait la politique d'ouverture et de circulation des données. Cependant, celle-ci ne figure plus dans le décret de 2020 du secrétaire d'État chargé de la transition numérique et des communications électroniques[132], mais a été inscrite dans celui de la ministre de la transformation et de la fonction publiques[133]. L'ouverture des données et des codes sources publics étant un sujet interministériel ayant trait à la transformation publique, à la transition numérique, à l'économie numérique et à la transparence de l'action publique, la politique d'ouverture des données et des codes sources relève à la fois du secrétaire d'État chargé de la transition numérique et des communications électroniques et de la ministre de la transformation et de la fonction publiques. Néanmoins, la politique relative à l'ouverture et à la circulation des données et des codes sources produits par l'État et les administrations n'est mentionnée dans aucun décret d'attribution, alors même que c'est un facteur important de modernisation de l'action publique.

Cette dichotomie dans les décrets d'attribution se retrouve au sein des administrations en charge de la politique d'ouverture des données et des codes sources.

D'une part, la direction interministérielle du numérique (DINUM), anciennement la DINSIC (direction interministérielle du numérique et des systèmes d'information et de communication), est rattachée au secrétariat général du gouvernement, sous l'autorité de la ministre de la transformation et de la fonction publiques[134]. Parmi ses missions, la DINUM est chargée de promouvoir l'action de l'État en matière d'ouverture des données dans un périmètre incluant les administrations et ses opérateurs sous tutelle. Pour ce faire, elle participe à l'administration du portail *data.gouv.fr* et concourt à la mise en œuvre du service public des données de référence. C'est à ce titre que le directeur interministériel du numérique occupe les fonctions d'administrateur général des données.

D'autre part, la direction interministérielle de la transformation publique (DITP)[135], est placée sous l'autorité de la ministre de la transformation et de la fonction publiques et est rattachée au secrétaire général du gouvernement. Elle a pour objectifs principaux l'amélioration de « *l'action des administrations au profit des usagers* » et l'animation des « *travaux de modernisation de la gestion publique* »[136]. Pour remplir ces missions, la DITP fait appel aux services de la DINUM. Dans ce cadre, la DITP a pris en charge la coordination de l'action de l'État à partir de 2018 dans le cadre du *Partenariat pour un gouvernement ouvert* lancé en 2011 - mission auparavant assurée par la DINUM.

Il convient de noter que la DITP est secrétaire du comité interministériel de la transformation publique (CITP). Ce comité, présidé par le premier ministre, rend compte des engagements prioritaires du Gouvernement en matière d'organisation administrative, largement issus des contributions du Grand débat national, parmi lesquels figure l'objectif d'*« une administration plus efficace capable de tirer le meilleur parti des outils numériques au service des usagers »*. Dans ce cadre la politique d'ouverture des données et des codes sources fait régulièrement l'objet d'un point d'avancement.

---

[131] Décret n°2018-906 du 24 octobre 2018 relatif aux attributions déléguées au secrétaire d'État auprès du ministre de l'économie et des finances et du ministre de l'action et des comptes publics, chargé du numérique

[132] Décret n°2020-1045 du 14 août 2020 relatif aux attributions du secrétaire d'État auprès du ministre de l'économie, des finances et de la relance et de la ministre de la cohésion des territoires et des relations avec les collectivités territoriales

[133] Décret n°2020-882 du 15 juillet 2020 relatif aux attributions du ministre de la transformation et de la fonction publiques

[134] Décret n°2020-882 relatif aux attributions du ministre de la transformation et de la fonction publiques

[135] Décret n°2017-1584 relatif à la direction interministérielle de la transformation publique et à la direction interministérielle du numérique et du système d'information et de communication de l'État

[136] Décret n°2015-1165 du 21 septembre 2015 relatif à la direction interministérielle de la transformation publique

Compte tenu des enjeux liés à la politique publique de la donnée (cf. partie 1), la mission propose d'assurer un portage politique des recommandations issues du présent rapport au niveau du Premier ministre, notamment par l'inscription à l'ordre du jour des comités interministériels présidés par le Premier ministre.

Par ailleurs, une circulaire pourrait venir préciser les principes de la politique publique de la donnée sur les sujets ayant trait à la gouvernance, les missions et les responsables dans les administrations, l'interopérabilité des systèmes d'information, les standards de qualité de la donnée, les guides juridiques, etc.

> **Recommandation** : Assurer un portage politique au niveau du Premier ministre des enjeux de la donnée et des codes source. Inscrire, notamment par leur inscription à l'ordre du jour des comités interministériels présidés par le Premier ministre, le suivi et la mise en œuvre de cette politique. Édicter une circulaire établissant les principes de la politique publique de la donnée (gouvernance, missions et responsables dans les administrations, interopérabilité, qualité, guides juridiques)

Réinvestir le rôle d'administrateur général des données, et élargir son périmètre à l'ouverture des algorithmes et de codes sources

L'administrateur général des données (AGD) a connu un positionnement variable au sein de l'État. Le titre d'AGD a été créé par le décret 2014-1050 du 16 septembre 2014 : il a pour rôle d'organiser l'ouverture et la circulation des données publiques, « *notamment aux fins d'évaluation des politiques publiques, d'amélioration et de transparence de l'action publique et de stimulation de la recherche et de l'innovation* ». Il était à l'origine placé sous l'autorité du secrétariat général à la modernisation de l'action publique et a été rattaché en 2017 au directeur interministériel du numérique et du système d'information et de communication de l'État[137]. Le rôle de l'AGD a ensuite été confié en 2019 au DINUM[138], alors que le titre d'administrateur général des données a disparu du fait de l'abrogation du décret de 2014 dans le même temps.

In fine, la fonction d'AGD est aujourd'hui une mission, parmi d'autres, que doit remplir le DINUM. En effet, les fonctions du DINUM, définies dans le décret de 2019, sont nombreuses et variées. Au-delà de sa mission de coordination et promotion de l'action de « *l'État et des organismes placés sous sa tutelle en matière d'inventaire, de gouvernance, de production, de circulation, d'exploitation et d'ouverture des données, et notamment des codes sources* », la DINUM est en charge de :

- l'élaboration de la stratégie numérique de l'État ;
- la contribution à la transformation numérique des politiques publiques ;
- la promotion de la dématérialisation des formalités administratives ;
- la définition des principales règles de construction et d'urbanisation des systèmes d'information publics ;
- la promotion de l'innovation, de l'expérimentation, « des méthodes de travail ouvertes, agiles et itératives », en utilisant des synergies possibles avec la société civile ;
- soutien du développement des compétences de l'État dans le domaine du numérique, notamment en lien avec les aspects ressource humaine et l'appropriation des méthodes du numérique par tous les fonctionnaires ;
- organisation et pilotage des démarches de mutualisation entre différentes administrations ;
- construction du réseau interministériel d'État ;
- élaboration et mise à disposition de ressources numériques partagées ;

[137] Décret n°2017-1584 relatif à la direction interministérielle de la transformation publique et à la direction interministérielle du numérique et du système d'information et de communication de l'État

[138] Décret n°2019-1088 relatif au système d'information et de communication de l'État et à la direction interministérielle du numérique

- exercer un rôle de conseil ou d'expertise sur les systèmes d'information d'une administration qui le demande ;
- conception et direction des projets numériques d'intérêt collectif.

Le décret d'octobre 2019 reflète ainsi une grande partie des étapes qui se rattachent au cycle de vie de la donnée, la politique d'ouverture des données et des codes sources publics n'étant qu'une mission parmi d'autres pour la DINUM.

Dans le cadre de sa fonction d'AGD, la DINUM peut être saisie *« par une personne morale de droit public de toute question portant sur [la] circulation [des données] ».* Elle administre le portail data.gouv.fr, participe à la mise en œuvre du service public des données de référence. En pratique, c'est le département Etalab qui remplit ce rôle au sein de la DINUM.

Il ressort des entretiens menés avec le DINUM, qu'il souhaite positionner sa direction davantage dans un rôle d'accompagnement de la politique d'ouverture menée par l'ensemble des ministères, plutôt que dans un rôle proactif de cette politique. Du reste, il ne souscrit pas à l'idée d'une « politique publique de la donnée » en tant que politique en soi, mais souhaite privilégier l'idée de la donnée au service des politiques publiques. Si la mission partage l'idée selon laquelle la donnée doit être replacée au cœur de la conception et du pilotage des politiques publiques menées par les administrations, les services déconcentrés et les collectivités - incluant non seulement les directions du numérique et les services statistiques ministériels mais aussi et surtout les directions métiers de l'administration centrale – et que l'ouverture des données et des codes sources ne doit pas être un objectif déconnecté des autres politiques publiques, il importe aussi de redonner un portage politique et administratif transversal à une telle démarche compte tenu des enjeux majeurs auxquelles une telle politique doit répondre, sans délaisser le besoin de portage de cette politique au sein de chaque ministère. La mission considère par ailleurs que l'utilité et la finalité de l'ouverture des données et des codes source ne peuvent être préjugées par les administrations productrices, et qu'il est donc nécessaire d'être à l'écoute des besoins de la société civile et des acteurs économiques.

Cette conception du rôle que doit jouer la DINUM dans la politique publique de la donnée se retrouve dans l'animation du réseau des administrateurs ministériels des données (AMD), mission relevant de l'AGD. Or, cette mission, assumée en pratique par Etalab, apparaît délaissée : les réunions d'information ont été interrompues en 2018, et sont trimestrielles depuis 2020 seulement. Par ailleurs, elles sont jugées trop descendantes par les AMD et pas suffisamment horizontales. Néanmoins, la DINUM a entrepris des actions, en lien avec les ministères, pour repositionner les AMD récemment nommés au sein des ministères, afin qu'ils puissent disposer de moyens d'intervention ; cette démarche initiée tardivement doit cependant encore porter ses fruits (cf. infra).

Afin de remédier à ce déficit de portage interministériel de la politique publique de la donnée, la mission propose de nommer un administrateur général de la donnée, des algorithmes et des codes sources (AGDAC) exerçant cette mission à plein temps, positionné comme adjoint auprès du DINUM. Compte tenu des enjeux rattachés à cette mission, une nomination de l'AGDAC par le premier ministre permettrait de renforcer sa visibilité au plan interministériel.

L'AGDAC aurait ainsi pour mission de définir la stratégie nationale d'ouverture des données et des codes sources publics, et de s'assurer de la cohérence de la doctrine appliquée au sein des administrations (périmètre d'ouverture, gestion de la sensibilité des données, choix de la profondeur temporelle). À titre d'exemple, l'ouverture des données et des codes sources publics repose sur des décisions au cas par cas des administrations sans qu'il n'existe de véritable doctrine définie au niveau interministériel. Par exemple, la manière dont la DGFiP a sélectionné les champs qui seraient ouverts dans la base DVF par rapport à la base anciennement accessible de manière limitée (Patrim), s'est fait en interne à la direction, sans harmonisation ministérielle ou interministérielle dans la manière d'apprécier la pertinence du choix de ces champs. Les données relatives au régime d'imposition (article du code général des impôts - CGI), le code SAGES du service responsable et l'identifiant du local ont ainsi été exclu de l'ouverture, sans qu'il soit permis de penser qu'ils permettent une réidentification plus forte que ne le permettent déjà les données ouvertes aujourd'hui dans la base. Ces champs contribueraient pourtant à la fiabilité de l'exploitation de la base, notamment l'article du CGI qui permet de déterminer la nature de la transaction.

Recommandation : Nommer un administrateur général de la donnée, des algorithmes et des codes sources (AGDAC), missionné par le Premier ministre, auprès du DINUM, ayant pour mission à temps plein de piloter la stratégie nationale d'ouverture de la donnée et des codes sources, en s'appuyant sur les administrateurs ministériels des données, des algorithmes et des codes source (AMDAC)

***In fine*, la prise en charge de cette nouvelle mission que constitue pour les acteurs publics la production et la valorisation des données mobilise des ressources, sans que ce travail et cet investissement ne soit toujours visible et reconnu.** Le déficit de reconnaissance de ces travaux nécessaires à la production de données en quantité et en qualité, amène ainsi la mission à préconiser une labellisation des services publics producteurs de données, proposition qui a notamment été formulée dans le cadre de la consultation publique. Ce label pourrait être établi selon un cahier des charges précisant des critères de qualité et d'accessibilité de données qui garantissent un réel investissement et une réelle prise en compte des besoins de la société civile et de l'économie. Ce label pourrait ainsi être rapproché de la démarche du service public de la donnée (SPD), notamment si les jeux de données faisant partie de ce SPD venaient à être élargis.

Recommandation : Créer un label de service producteur de la donnée pour reconnaître les efforts investis dans la donnée, par exemple dans le cadre du service public de la donnée

Par ailleurs, la transformation du titre d'AGD en AGDAC vise à renforcer l'ouverture et la mutualisation des codes sources produits au sein de l'État. En effet, les gains associés à cette démarche permettraient une meilleure utilisation des fonds publics. Un centre de ressource sur les briques « *open source* » utilisées par les administrations permettrait ainsi d'avoir des informations sur chacune des briques « open source » nouvellement utilisée. De plus, il n'existe pas de forge souveraine pouvant accueillir les codes sources produits par l'administration[139]. La mission recommande donc que l'AGDAC puisse piloter une stratégie visant pratiquer la politique d'ouverture des codes sources publics, et favoriser leur réutilisation au sein de l'État ainsi qu'une animation interministérielle rassemblant les développeurs de l'État au travers d'un « Open Source Program Office » (OSPO). L'OSPO pourrait être constitué par exemple d'un responsable et de deux à trois chargés de mission (3 à 4 ETP), dimensionnement qui semble nécessaire à la variété des missions qui doivent être prises en charge en matière d'ouverture et de partage des codes sources et des logiciels libres développés au sein du secteur public et plus particulièrement de l'État.

[139] Les deux acteurs principaux du marché des forges, GitLab et GitHub, sont tous deux états-uniens, et ce dernier acteur, largement dominant, est critiqué par l'association APRIL[139]. Lorsqu'un ministère souhaite héberger un code source qu'il produit sur une forge hébergée en France, il est aujourd'hui incité à en déployer une ou à se tourner vers l'ADULLACT.

Recommandation : Créer un « Open Source Program Office » (OSPO) ou une mission logiciels libres au sein de TECH.GOUV, chargée d'aider l'administration à ouvrir et à réutiliser les codes sources publics, d'identifier les enjeux de mutualisation et de créer des liens avec les communautés open source existantes et d'accompagner les talents français dans ce domaine

Par ailleurs, si la DINUM peut être saisie par une personne morale de droit public de toute question portant sur la circulation des données, elle ne prévoit pas de modalités permettant de mieux prendre en compte les problématiques émanant de la société civile. Certes, il est possible de s'adresser directement à l'administration détentrice des données et des codes sources pour l'ouverture de ses données et codes sources et d'engager un dialogue avec elle. En cas de refus de l'administration, il est ensuite possible de s'adresser à la CADA pour avis, avant d'engager éventuellement une action contentieuse devant le tribunal administratif.

Cependant, il n'existe pas de cadre de discussion intermédiaire permettant de débattre avec la société civile et avec les porteurs de politiques publiques sur l'opportunité d'ouvrir tel ou tel jeu de données ou code source, en lien avec la doctrine définie au niveau interministériel. Par ailleurs, la mission proposer de confier un rôle de sanction à la CADA, faisant substantiellement évoluer ses missions (cf. paragraphe 3.2). C'est pourquoi la mission propose d'associer la société civile, par les consultations citoyennes et le Forum ouvert du Partenariat pour un gouvernement ouvert, afin d'identifier et recueillir, les souhaits en matière d'ouverture de jeux de données et de codes sources, et les relayer à l'AGDAC afin que celui-ci puisse être en mesure de garantir une cohérence dans la mise en œuvre de la politique publique d'ouverture selon la doctrine définie au niveau ministériel et discutée au sein du réseau des AGD.

Recommandation : Associer la société civile, par les consultations citoyennes et le Forum du Partenariat pour un gouvernement ouvert, à l'identification des jeux de données et des codes sources à ouvrir

Rétablir le suivi de la politique d'ouverture des données et des codes sources

Le programme budgétaire n°129 relatif à la coordination du travail gouvernemental est articulé autour de sept objectifs parmi lesquels figurent les objectifs consistant à « améliorer l'information du citoyen sur les actions du Gouvernement » ou encore « Accompagner les administrations dans leur transformation et la simplification de leurs relations avec les usagers ». Ce dernier objectif est mesuré au travers de l'indicateur « ouverture et diffusion des données publiques », lui-même décliné au travers de sous-indicateurs qui concernent le site *data.gouv.fr*, à savoir : nombre de ressources en open data (site *data.gouv.fr*), nombre de contributeurs actifs (site *data.gouv.fr*), nombre de réutilisations (site *data.gouv.fr*).

Dans le cadre du projet de loi de finances pour 2021, ces sous-indicateurs ont évolué et suivent désormais l'objectif d'ouverture et de diffusion des données publiques au travers du nombre d'API référencés sur *api.gouv.fr* et de l'indice de satisfaction des usages telle qu'issue de l'Observatoire de la dématérialisation de la qualité.

À compter de 2021, la mise en œuvre de la politique d'ouverture des données publiques sera évaluée sous l'angle de la simplification des démarches en ligne. En effet, la mesure du nombre d'API vise à suivre la stratégie d'accélération de la mise en œuvre du principe *« dites-le nous une fois », « point d'entrée des APIs du service public, mises à la disposition des collectivités, des ministères et des entreprises pour construire des services informatiques au service de tous »*. Quant à l'indice de satisfaction des usagers de l'Observatoire de la dématérialisation de la qualité, celui-ci permet de rendre compte des résultats obtenus à partir des réponses des usagers à la question « Comment s'est passée cette démarche pour vous ? » à la fin d'une démarche via le bouton « Je donne mon avis ».

Si ces indicateurs présentent un intérêt, il est particulièrement regrettable que l'ouverture des jeux de données et des réutilisations ne fassent plus l'objet d'un suivi dans le cadre des projets de loi de finances. En outre, il serait intéressant de suivre l'ouverture des algorithmes et des codes sources au travers d'indicateurs dédiés.

Aussi, la mission préconise de maintenir dans chaque projet de loi l'examen d'un volet relatif à la politique publique de la donnée, permettant de traiter à la fois la problématique de la collecte des données pour les besoins de pilotage de la politique publique, mais aussi de partage et de publication des données et des codes sources.

> **Recommandation :** Structurer le pilotage et le suivi de la politique d'ouverture des données et des codes sources au niveau interministériel (indicateurs de performance, insertion dans les études d'impact des projets de loi)

Repositionner les administrateurs ministériels des données et étendre leur périmètre aux algorithmes et aux codes sources

Créé en 2016 pour relayer l'action du DINUM dans son rôle d'administrateur général des données, le réseau des AMD est encore incomplet en 2020. Plusieurs périmètres ministériels sont imparfaitement couverts ou seulement sur le point de l'être, comme la culture, les sports, l'enseignement supérieur, l'outre-mer, le travail, l'emploi et l'insertion.

La fonction d'AMD est peu visible au sein de l'administration et de l'extérieur. Ainsi, plusieurs réponses au questionnaire adressé par la mission, font état d'une confusion entre l'AMD et le délégué à la protection des données personnelles. De plus, les relations entre les AMD et les délégués à la protection des données personnelles sont peu fréquentes. Enfin, la distinction entre les AMD et les SSM (services statistiques ministériels) n'est pas bien comprise de certains acteurs extérieurs. Enfin, seuls 2 des 9 AMD interrogés disposent de référent pour la donnée au sein des opérateurs, ce qui dénote leur manque de visibilité.

Le périmètre des missions des AMD et les compétences des AMD ne sont pas clairement définis, ce qui induit une forte hétérogénéité**.** Il n'y a jamais eu de texte réglementaire pour encadrer la fonction, alors qu'il était envisagé par la DINSIC. Historiquement, les AMD n'avaient pas vocation à être en charge au niveau du ministère de la politique d'ouverture des données et des codes sources publics, dans la mesure où Etalab possédait un référent *open data* dans chaque ministère. Toutefois, la fin du dispositif des référents *open data* a orienté naturellement la politique d'ouverture des données et des codes sources publics vers l'AMD, sans que cela soit officiel. Ainsi, certains AMD considèrent que la vocation des AMD est avant tout interne à l'administration. Le sujet de l'ouverture des codes sources est considéré par une minorité d'entre eux comme relevant de leurs missions.

Dès lors, le rôle de l'AMD - renommés « administrateurs ministériels de la donnée, des algorithmes et des codes sources » (AMDAC) - devrait être redéfini autour de trois fonctions principales :

- décliner la politique d'ouverture des données et des codes sources publics au niveau ministériel dans les lettres de mission des directions d'administrations centrales et dans les contrats d'objectifs et de moyens des opérateurs sous tutelle ;
- identifier et négocier l'ouverture ou le partage des jeux de données et des codes sources d'intérêt général détenus par des acteurs privés (cf. partie 5) ;
- faciliter le partage des données et des codes sources entre administrations.

Par ailleurs, **le recrutement des AMDAC constitue un point d'attention majeur pour s'assurer de la bonne adéquation entre le profil et les compétences attendues pour la réalisation de cette mission**. Les profils professionnels des AMD sont ainsi variés : certains ont participé à la transformation numérique au sein de leur ministère et ont repris à ce titre le poste d'AMD, tandis que d'autres ont eu une carrière avec davantage de fonctions d'encadrement et approchent les sujets de la donnée par ce prisme. Il est à noter que la DINUM n'a été associée à presque aucun recrutement d'AMD sur les 9 AMD en poste, mis à part l'AMD du Quai d'Orsay qui occupe par ailleurs le poste de DNUM (directeur du numérique). Dans ce cadre, la mission recommande de s'assurer que l'AMDAC détient les compétences techniques et juridiques requises en matière de gestion des données et des codes sources et qu'il a une bonne connaissance du secteur, de l'organisation du ministère et des opérateurs sous tutelle. Le principe d'une lettre de mission signée par le ministre ou les ministres et adressée à l'AMDAC pourrait figurer dans le décret d'organisation de chaque DG accueillant l'AMDAC. Cette lettre de mission comporterait un socle défini par l'AGDAC et une partie définie par le ministère. Le recrutement aurait lieu sur la base de cette lettre de mission et d'un profil de poste proposé par le ministère, sur lequel l'AGDAC donne un avis.

Par ailleurs, l'AMDAC pourrait avoir la latitude d'adapter annuellement sa feuille de route pour l'année suivante sous forme d'une lettre d'objectifs, faisant l'objet d'une consultation des directions du ministère, puis d'une signature par le ministre. Réciproquement, l'AMDAC pourrait être consulté chaque année pour proposer un ou des objectifs liés à la donnée, intégrés de façon transverse aux objectifs annuels de chaque direction et faisant l'objet d'une évaluation annuelle des directeurs.

Enfin, les AMD possèdent des moyens limités et extrêmement variables**.** En moyenne, les AMD possèdent un budget de 300 000 € par an, avec des équipes représentant 2,5 ETP. Ces moyennes cachent cependant des situations diverses : l'AMD du ministère de l'intérieur a ainsi vu ses moyens considérablement augmentés en 2020 lorsqu'il est devenu sous-directeur ce qui lui permet d'allouer des moyens de la sous-direction à son rôle d'AMD, tandis que la DGFiP dispose d'un demi ETP pour remplir la fonction d'AMD. La mission recommande de s'assurer de la cohérence des moyens de l'AMDAC avec ses missions, afin qu'il dispose des moyens humains et matériels, et de temps, suffisants, mais aussi qu'il soit associé en amont des projets conduisant à la production de données et de codes sources.

De plus, des disparités importantes existent entre les ministères suivant le type de données produites. Ainsi, dans les ministères où le sujet des données personnelles est prégnant (notamment dans le domaine des politiques sociales, de la santé, de l'intérieur, de la justice), l'ouverture des données est rendue plus difficile pour deux raisons : d'une part, l'AMD considère que l'ouverture des données est un objectif secondaire par rapport à la circulation des données à l'intérieur du ministère du fait de l'impossibilité de les ouvrir, et d'autre part, le positionnement du délégué à la protection des données, chargé de veiller au respect du cadre juridique entourant les données personnelles, est susceptible de freiner l'action de l'AMD en matière d'ouverture des données. Ainsi, la mission recommande d'instaurer des formations conjointes entre AMDAC et délégués à la protection des données, de sorte à favoriser l'acculturation mutuelle aux fonctions exercées par ces deux types de profil.

Recommandation : Élargir et renforcer la fonction d'administrateur ministériel des données, des algorithmes et des codes sources (AMDAC) :

- en redéfinissant leurs missions dans une fiche de poste type ;
- en dotant les AMDAC d'une lettre de mission signée par les ministres concernés après consultation des directions générales et de l'AGDAC
- en s'assurant que l'AMDAC a des moyens d'intervention suffisants ;
- en systématisant des formations conjointes entre AMDAC et délégués à la protection des données.

Soutenir les collectivités locales dans la mise en œuvre de leur politique d'ouverture des données et des codes sources

Les collectivités locales de plus petite taille peinent à mettre en œuvre une politique d'ouverture des données et des codes sources (cf. paragraphe 2.2). Plusieurs régions ont néanmoins développé des initiatives intéressantes permettant de mettre à disposition leurs moyens à destination des collectivités territoriales de leur région.

Du reste, il est à noter que la loi NOTRe de 2015 confie aux régions « *la coordination, au moyen d'une plateforme de services numériques qu'elle anime, de l'acquisition et de la mise à jour des données géographiques de référence nécessaires à la description détaillée de son territoire ainsi qu'à l'observation et à l'évaluation de ses politiques territoriales, données dont elle favorise l'accès et la réutilisation* ». Ainsi, à chaque région est associée une ou deux plateformes[140] selon que l'État et la région se sont entendus ou non sur l'infrastructure, y compris sur la gouvernance. Ces plateformes sont principalement construites sur la base de logiciels libres (Prodige, GeOrchestra) qui répondent aux principes de la directive INSPIRE (cf. partie 2).

Néanmoins, plusieurs collectivités territoriales rencontrées par la mission ont souligné l'importance du soutien de l'État dans cette démarche, qui bénéficie à tous. Dès lors, la mission propose de confier à l'ANCT une mission d'accompagnement des collectivités territoriales dans la publication des données et des codes sources via des programmes cofinancés entre État et région.

> **Recommandation** : Confier à l'ANCT une mission d'accompagnement des collectivités territoriales dans la publication des données et des codes sources via des programmes co-financés entre État et région

Encourager la démarche de science ouverte par des incitations

Dans une note de novembre 2019[141], le Comité pour la Science Ouverte (CSO) rappelait l'importance des codes sources dans la recherche d'aujourd'hui et préconisait plusieurs axes de travail, notamment sur l'archivage et le référencement, sur le système de citation/réputation, sur la valorisation des productions logicielles, sur la pérennisation du patrimoine logiciel issu de la recherche, et sur la mutualisation des ressources.

Pour développer la culture de l'ouverture des données et des codes sources et convaincre les communautés et les établissements du supérieur de mettre en place cette politique, deux principaux leviers peuvent être évalués : l'évaluation et le financement.

S'agissant de l'évaluation, la démarche de science ouverte pourrait être davantage suivie dans les indicateurs utilisés par le Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur (HCERES), comme c'est le cas pour les laboratoires d'informatique. L'ouverture des données comme des codes sources doivent faire partie du « cahier des charges » sur lequel un laboratoire est évalué et ce, quel que soit son champ disciplinaire.

En outre, l'ouverture des données et des codes sources serait nettement plus généralisée si elle devenait un critère essentiel pour le financement des projets de recherche par l'ANR. Les critères définis par l'ANR sont de nature à avoir une influence forte sur la politique des laboratoires et peuvent inciter les partenaires industriels de ces projets à accepter la démarche. Cette politique ne peut cependant se limiter à une approche nationale. Au niveau européen, la France devrait favoriser l'intégration de ces démarches dans la définition, le pilotage et le financement des projets de recherche.

[140] Voir cette cartographie de 2018 publiée par l'AFIGEO : http://www.afigeo.asso.fr/la-vie-en-region/catalogue-des-idg.html mais elle n'est pas exhaustive (exemple en Auvergne-Rhône-Alpes où l'Etat déploie sa propre plateforme DatARA) et a depuis évolué.

[141] Comité pour la Science Ouverte, Note d'opportunité sur la valorisation des logiciels issus de la recherche, Groupe projet « Logiciels libres et *open source* », novembre 2019.

À cet égard, les organismes de transfert de technologie ne doivent pas constituer un frein au développement de la science ouverte. Ainsi, les collaborations sur des projets de recherche ne devraient pas être ralenties par l'absence de maîtrise, de la part des cellules de valorisation des instituts de recherche, des processus de mise à disposition et de programmation informatique.

> **Recommandation :** Prendre davantage en compte les démarches d'open source et d'open data pour le rayonnement de la recherche française dans les évaluations et le financement des projets

3.2. Un cadre de régulation à repenser

CADA et CNIL, deux autorités garantes de l'équilibre entre la transparence de la vie publique d'une part et protection des données personnelles d'autre part

La CADA, créée par la loi n°78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal, a pour but de garantir au public un accès aux documents administratifs face à l'administration. La loi qui a créé cette commission est contemporaine de la **loi n°78-17 du 6 janvier 1978 relative l'informatique, aux fichiers et aux libertés, et donc de la CNIL**.

Ces deux institutions, autorités administratives indépendantes, jouent un rôle majeur dans la régulation de l'ouverture des données, la première au titre de l'application des dispositions du code des relations entre le public et l'administration, et la deuxième au titre de l'application du règlement général de protection des données (RGPD) :

- d'après l'article L.340-1 du Code des relations entre le public et l'administration, la CADA est plus particulièrement en charge de veiller au respect de la liberté d'accès aux documents administratifs et aux archives publiques et à la légalité des réutilisations des informations publiques telles que prévues par le code des relations entre le public et l'administration ;
- la CNIL, quant à elle, exerce ses missions conformément à la loi Informatique et Libertés du 6 janvier 1978 modifiée par la loi du 20 juin 2018, à la suite du règlement général sur la protection des données personnelles (RGPD) entré en application le 25 mai 2018. Elle assure ainsi le rôle de régulateur des données personnelles au travers d'une mission générale d'information des personnes sur leurs droits, d'une mission d'accompagnement et de conseil des différents acteurs, mais aussi au travers d'une mission de contrôle et de sanction en cas de non-respect du droit.

Les deux autorités ont donc des missions obéissant à des logiques très différentes. A l'occasion de l'examen du projet de loi pour une République numérique, la question d'un rapprochement entre la CNIL et la CADA avait fait l'objet d'une analyse particulière par le Conseil d'État en mars 2016[142]. Celui-ci soulignait que *« le problème d'un recoupement des attributions des deux autorités ne se pose donc que dans le document administratif auquel l'accès est demandé ou dont la réutilisation est souhaitée contient des données à caractère personnel, ce qui reste, pour l'instant marginal ».* Il concluait ainsi : *« il est apparu dans les entretiens menés avec leurs dirigeants, [que] rien ne devrait faire obstacle à ce que, sur le sujet le plus nouveau, l'anonymisation des données à caractère personnel dans les documents et les bases de données mis en ligne au nom de l'*open data*, la CNIL ait le rôle pilote de définir les référentiels ou méthodologies dont la CADA ferait application dans les cas individuels ».*

In fine, le rapprochement entre la CNIL et la CADA s'est traduit, en matière de gouvernance, par une participation croisée dans les deux collèges des autorités. En outre, la loi prévoit également que **les deux institutions peuvent se réunir** « *dans un collège unique, sur l'initiative conjointe de leurs présidents, lorsqu'un sujet d'intérêt commun le justifie* »[143].

[142] Jean Massot, Président de section du Conseil d'Etat et ancien représentant de la CNIL à la CADA, rapport sur les incidences sur la CNIL et la CADA du projet de loi pour une République numérique, mars 2016

[143] Articles 26, 27 et 28 de la loi pour une République numérique.

Dans sa réponse au questionnaire de la mission, la CNIL parle ainsi de co-régulation de l'*open data*, étant entendu que celle-ci n'intervient que sur des problématiques faisant intervenir des données personnelles. **Un protocole de collaboration CADA/CNIL permettant aux administrations d'identifier l'autorité compétente pour traiter une problématique a ainsi été établi dans le cadre du guide pratique de la publication en ligne et de la réutilisation des données publiques (cf. *infra*).**

Renforcer l'effectivité de la loi pour une République numérique en confiant un pouvoir de sanction à la CADA

L'analyse précitée du Conseil d'État, à l'occasion de l'examen du projet de loi pour une République numérique, soulignait le contraste dans les modes d'intervention de la CADA d'une part, et de la CNIL d'autre part.

La CADA peut être saisie pour des demandes d'avis par une personne physique ou morale, étant entendu que « *la saisine pour avis est un préalable obligatoire à l'exercice d'un recours contentieux* »[144]. Elle peut être « *saisie par une personne à qui est opposé un refus de communication ou un refus de publication de documents administratifs* […], *un refus de consultation ou de communication des documents d'archives publiques*[145]*,* […] *ou une décision défavorable en matière de réutilisation d'informations publiques* » (article L.342-1 du Code des relations entre le public et l'administration). Ainsi, l'action de la CADA concerne essentiellement des demandes individuelles des usagers du service public souhaitant accéder aux données détenues par ces services, dépassant la problématique de l'ouverture des données et des codes sources.

Par ailleurs, la commission peut être saisie par une administration pour prononcer, le cas échéant, une sanction dans le cadre d'une réutilisation illégale des informations publiques. Ce cadre permet aux administrations d'avoir un contrôle effectif, par rapport à la loi, de la réutilisation des informations publiques contenues dans les documents administratifs qu'elles publient ou communiquent.

La portée des avis de la CADA dans le cadre de la demande de communication d'un document administratif est par nature limitée. En effet, l'administration peut se contenter de ne pas répondre aux demandes pour signifier son refus[146], que ce soit en amont ou en aval de l'avis de la CADA. Un avis favorable de la CADA à l'ouverture d'un document administratif ne rend pas obligatoire la publication de ce document par l'administration en question. Les administrations doivent simplement informer la CADA des suites qu'elles entendent donner à l'avis de la CADA. Néanmoins, le taux d'information sur les suites données aux avis favorables rendus par la Commission est en diminution constante depuis 2015, passant de 71,66% à 53,81%[147]. Ce taux ne reflète pas la réalité du suivi des avis favorables de la CADA, dans la mesure où de nombreuses administrations ne tiennent pas la CADA informée de la suite qu'elles donnent à ses avis favorables, sans que cela signifie qu'elles n'ont pas ouvert les documents administratifs concernés.

D'après l'Ouvre-boîte, association dont l'objet est d'« obtenir l'accès et la publication effective des documents administratifs, et plus particulièrement des données, bases de données et codes sources, conformément aux textes en vigueur », 77% des demandes d'accès ont nécessité une saisine de la CADA (sur un total de 98 demandes depuis 2017) et 12% de ces demandes ont été satisfaites grâce à la CADA mais 41% ont nécessité de porter la demande devant le tribunal administratif (et 7% des demandes ont été satisfaites grâce à cette démarche contentieuse).

[144] Article L.342-1 du Code des relations entre le public et l'administration

[145] À l'exception des « minutes et répertoires des officiers publics ou ministériels et les registres de convention notariés de pacte civil de solidarité » (article L.211-4 c) du Code du patrimoine), ainsi que les actes et documents produits ou reçus par les assemblées parlementaires (article L.342-1 du Code des relations entre le public et l'administration).

[146] Article R.311-12 et R.311-13 et article L.311-14 du Code des relations entre le public et l'administration

[147] Source : rapport d'activité de la CADA 2019

En outre, si l'avis de la CADA demande l'ouverture des documents administratifs concernés, les administrations peuvent ouvrir les documents en question mais dans un format les rendant difficilement utilisables voire en ne les rendant que partiellement disponibles aux demandeurs. Par exemple, la Fédération Française de Football (FFF), face à une demande de transmission d'une base de données concernant les incidents au cours des matchs (demande approuvée par la CADA puis par le tribunal administratif), a décidé de transmettre une base de données sans explication, avec des colonnes mal nommées et des données sans profondeur temporelle.

Par ailleurs, de manière générale, l'administration n'a pas à communiquer les documents administratifs lorsque les demandes ne sont pas précises. Or, la précision des demandes découle aussi de la transparence de l'administration, et en particulier des documents qu'elle déclare détenir dans le Répertoire d'informations publiques notamment. Ainsi, il arrive que des demandes soient jugées irrecevables par la CADA, faute de précision, lorsqu'elles portent sur des informations publiques dont le document porteur n'est pas clairement identifié dans la demande, faute que ce document soit connu du public.

La mission propose de renforcer l'effectivité de la loi pour une République numérique en faisant de la CADA une juridiction de premier niveau spécialisée et en lui donnant un pouvoir de sanction sur les demandes d'ouverture de données et de codes sources. Néanmoins, il est à noter que la CADA dispose de 16 ETPT en 2019 et a reçu 6 786 saisines la même année (dont une seule concerne une saisine pour sanction dans le cadre d'une réutilisation illégale des informations publiques). En outre, au 31 décembre 2019, la CADA accusait un retard de 19 mois dans le traitement des dossiers[148], qu'elle explique par une augmentation des demandes sans une augmentation des moyens de la commission proportionnelle au nombre de demandes. Si le premier confinement a permis de réduire considérablement le stock des demandes et de revenir à des délais plus raisonnables, des réformes de procédure vont être mises en place par la commission afin d'accélérer le traitement des demandes, en particulier le traitement des demandes par le seul président, en application de l'article L.341-5-1 du Code des relations entre le public et l'administration. Aujourd'hui, 30% des demandes sont ainsi traitées par le seul président.

Aussi, la réforme envisagée, au-delà des moyens supplémentaires qu'elle engendrerait, nécessite d'alléger l'activité de la CADA. La mission a ainsi identifié deux pistes possibles :

- supprimer le recours administratif préalable obligatoire pour les documents qui sont communicables de manière évidente, quitte à l'inscrire clairement dans la loi (listes électorales, les permis de construire, les dossiers médicaux, les dossiers des fonctionnaires, etc.) : en effet, l'avis de la CADA ne présente aucune valeur ajoutée dans ces situations et a tendance à déresponsabiliser les acteurs publics qui s'abritent derrière un avis de la CADA ;
- renforcer l'animation des 1600 personnes responsables de l'accès aux documents administratifs (PRADA) dont le rôle consiste à « *réceptionner les demandes d'accès aux documents administratifs et de licence de réutilisation des informations publiques ainsi que les éventuelles réclamations et de veiller à leur instruction* » et à « *assurer la liaison entre l'autorité auprès de laquelle elle est désignée et la commission d'accès aux documents administratifs* »[149] : si un ETP est nécessaire pour remplir cette mission, l'animation de ce réseau permettrait de responsabiliser les acteurs publics dans la communication des documents administratifs pour des demandes simples et désengorgerait les flux de saisines auprès de la CADA.

Ces mesures permettraient ainsi à la CADA de repositionner son activité sur l'instruction de dossiers plus complexes, et en particulier sur l'instruction des demandes de sanction des administrations refusant de rendre publics des jeux de données ou des codes sources.

[148] Des délais précis sont définis par l'article R.343-3 du Code des relations entre le public et l'administration : « la commission notifie son avis à l'intéressé et à l'administration mise en cause, dans un délai d'un mois à compter de l'enregistrement de la demande au secrétariat. Cette administration informe la commission dans le délai d'un mois qui suit la réception de cet avis, de la suite qu'elle entend donner à la demande ».

[149] Article R.330-4 du Code des relations entre le public et l'administration

Il est à noter, en effet, que la CADA a reçu 44 saisines en dehors des demandes simples (contre une centaine environ en 2018 / 2019). Ces dossiers concernent soit la mise en ligne d'un document, qui ne pose, en règle générale, pas de difficulté particulière, soit la mise en ligne des bases de données de l'administration. Ce sont ces dernières (représentant environ 40 % de ces saisines) qui nécessitent des connaissances techniques et une vision exhaustive du contenu et de l'architecture des bases de données, pour pouvoir apprécier si la mise en ligne n'est, techniquement ou compte tenu de la charge qu'impliquerait le retraitement des secrets protégés et des données à caractère personnel, pas possible. L'instruction de ces dossiers nécessite un temps d'échanges et une expertise plus approfondie qu'actuellement**. En première analyse, deux emplois supplémentaires seraient ainsi nécessaires à la commission.**

Recommandation : Faire évoluer le droit d'accès aux documents administratifs pour renforcer l'effectivité de la loi en confiant un pouvoir de sanction à la CADA en cas de non-respect des dispositions du CRPA relatives à la communication et à la publication des données et documents et pour alléger l'activité de la CADA sur les saisines simples, et pour fluidifier la gestion des dossiers récurrents devant la CADA

Rééquilibrer le rôle de la CNIL en faveur de l'accompagnement des producteurs et des usagers de la donnée

Dans la réponse au questionnaire de la mission (cf. annexe du rapport), la CNIL détaille les actions d'accompagnement des différents acteurs dans la mise en œuvre de la politique d'ouverture des données. En particulier, CADA et CNIL ont travaillé ensemble à la publication d'un guide pratique de la publication en ligne et de la réutilisation des données publiques, en exposant le cadre juridique ainsi que des points de doctrine. Une fiche pratique relative à l'anonymisation des documents administratifs y est également annexée. L'élaboration de ce guide s'est inscrite dans une démarche collaborative et a fait l'objet d'une consultation publique effectuée auprès de l'ensemble des acteurs de l'open data (collectivités publiques et réutilisateurs). D'autres publications, abordant des questions aussi bien sectorielles (par exemple sur les marchés publics) que transverses (par exemple sur les licences de réutilisation), doivent paraître. Ce type de démarche a permis de réduire, selon la CNIL, les sollicitations des délégués à la protection des données personnelles. Néanmoins, les auditions ont également fait apparaître que ces outils ne sont pas encore suffisamment connus des acteurs.

Dans sa réponse au questionnaire, la CNIL souligne également l'effort de pédagogie nécessaire qu'il reste à faire auprès des administrations, pour les sensibiliser notamment aux enjeux de l'anonymisation des données et de leur réutilisation et préconise ainsi des formations conjointes des administrateurs des données et des délégués à la protection des données. Plus globalement, l'effort de pédagogie sur la mise en œuvre du RGPD – qui introduit un changement de paradigme - doit encore être poursuivi, la CNIL regrettant que les acteurs, aussi bien privés que publics, s'« autocensurent » dans l'ouverture et le partage des données alors que le règlement édicte des principes de travail mais pas des principes d'abstention.

Par ailleurs, la CNIL est régulièrement saisie par des chercheurs ou des sociétés privées de sollicitations pour valider des solutions particulières d'anonymisation ; elle procède à des analyses au cas par cas et se prononce au regard des critères établis par le groupe des CNIL européennes[150]. La CNIL indique ainsi ne pas s'être emparée des dispositions de la loi pour une République numérique « de certifier ou d'homologuer des méthodologies générales aux fins de certification par des tiers agréés, de processus » compte tenu de la complexité de ces procédures et de leur non réplicabilité.

Malgré ces efforts, les acteurs auditionnés par la mission regrettent le caractère peu opérationnel des outils mis à leur disposition par la CADA et la CNIL et attendent un accompagnement et des conseils plus concrets pour la mise en œuvre de leur politique d'*open data* mais aussi pour l'accès aux données ou aux codes sources.

[150] Voir l'avis 2005/2014 du G29.

Plusieurs personnes auditionnées par la mission décrivent ainsi un véritable parcours du combattant lorsqu'il s'agit d'obtenir auprès de la CNIL une autorisation de traitement de données de santé à des fins de recherche, d'autant que les chercheurs ne sont pas formés pour mener des études d'impact. Cela engendre ainsi une charge de travail très importante et, par voie de conséquence, des délais conséquents pour constituer des dossiers avant de pouvoir engager les travaux de recherche.

Par ailleurs, dans certains cas, les engagements demandés dans ce cadre, par exemple sur la sécurité informatique de leur système d'information, dépassent parfois les paramètres sur lesquels le chercheur a réellement prise. Cet aspect très formel des demandes peut avoir un effet contre-productif dans la mesure où cela conduit à remplir le dossier de manière mécanique[151], là où la CNIL devrait amener le chercheur à réfléchir à la meilleure manière d'inscrire ses travaux dans le cadre de la protection des données personnelles.

Dans un secteur où la concurrence se joue à l'échelon international, la lourdeur des démarches administratives couplées à l'interprétation stricte du RGPD, visant certes un objectif essentiel – celui de la protection des données personnelles – amènent aussi les chercheurs et les entreprises à se tourner vers l'étranger, notamment vers les pays anglo-saxons.

Globalement, les différents témoignages, aussi bien sur le manque d'opérationnalité des outils pour mener une politique d'opendata que sur le manque d'accompagnement des acteurs dans leur démarche d'accès aux données, révèlent des fortes attentes vis-à-vis des autorités de régulation, mais aussi d'acteurs tels que le *Health Data Hub* dans le cas de la santé pour faciliter l'ouverture et l'accès aux données (cf. partie 3).

Acculturer la CADA et la CNIL aux risques des systèmes d'information et aux nouveaux usages de la donnée

Les nouveaux usages de la donnée nécessitent une adaptation du cadre juridique et européen (cf. partie 3), mais aussi une acculturation des autorités en charge d'assurer la régulation de la politique publique de la donnée. Cela passe par le recrutement de compétences scientifiques, plus particulièrement d'ingénieurs et de data-scientists à la fois au niveau des collèges de la CADA et de la CNIL mais aussi dans les équipes en charge d'instruire les dossiers pour le collège.

[151] Un chercheur a ainsi indiqué à la mission : « *Du coup, on fait un copier-coller d'un ancien dossier car on ne sait pas trop.* »

Les collèges des deux autorités ont une composition similaire :

- **la CADA est composée de onze membres (article L.341-1 du Code des relations entre le public et l'administration)** : un membre du Conseil d'État, un membre de la Cour de cassation, de la Cour des Comptes, un député, un sénateur, un élu des collectivités territoriales, un professeur de l'enseignement supérieur, une personnalité qualifiée en matière d'archives, le président de la CNIL (ou son représentant), une personnalité qualifiée en matière de concurrence et de prix, une personnalité qualifiée en matière de diffusion publique d'informations ;
- **la CNIL est composée de 18 membres (article 9 de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés)** : deux députés et deux sénateurs, deux membres du Conseil économique, social et environnemental, six représentants des hautes juridictions (deux conseillers d'État, deux conseillers à la Cour de cassation, deux conseillers à la Cour des comptes), cinq personnalités qualifiées désignées par le Président de l'Assemblée nationale (une personnalité), le Président du Sénat (une personnalité), en Conseil des ministres (trois personnalités), le Président de la CADA (Commission d'accès aux documents administratifs).

La mission considère que la composition de ces collèges pourrait être revue pour tenir davantage compte des besoins d'expertise en matière d'usage des données et de sécurité des systèmes d'information. Cette recommandation est en lien avec la participation proposée de l'ANSSI à la politique d'ouverture des données et des codes sources (cf. *infra*).

> **Recommandation :** Prévoir dans les collèges de la CNIL et de la CADA deux personnalités qualifiées compétentes, l'une en matière de sécurité des systèmes d'information et l'autre sur les nouveaux usages de la donnée

Ce besoin d'expertise doit aussi se traduire, au-delà de la composition des collèges, dans les équipes qui composent les services de la CNIL et de la CADA. En particulier, la CNIL est composée de 215 agents, dont 48 % de postes occupés par des juristes, 22 % par des assistants juridiques, 19 % par des ingénieurs et auditeurs. Compte tenu des besoins d'accompagnement et de compréhension des enjeux liés aux nouveaux usages de la donnée, il apparaît indispensable de diversifier les compétences et les profils au sein de l'autorité.

À cet égard, la mise en place du laboratoire d'innovation numérique de la CNIL (LINC) au sein de la DTI (Direction des technologies et de l'innovation) est un dispositif qui participe à cet objectif. Il a pour objet de mener des réflexions, d'informer et de partager sur les tendances émergentes d'usage du numérique et des données, ainsi que de conduire des projets d'expérimentation et de prototypage d'outils, de services ou de concepts autour des données. Cependant, cette démarche apparaît encore confidentielle et mériterait d'être amplifiée.

Calibrer les moyens de la CNIL à la hauteur des enjeux

Compte tenu des enjeux auxquels les deux autorités doivent répondre et des attentes, aussi bien de la part des pouvoirs publics mais aussi des citoyens, des entreprises et des chercheurs, le renforcement des moyens de la CNIL et de la CADA apparaît nécessaire.

La loi de finances pour 2020 a accordé 10 postes supplémentaires à la CNIL, portant ses effectifs de 215 en 2019 à 225. Le projet de loi de finances pour 2021, encore en débat au Parlement au moment de la rédaction de ce rapport, a prévu d'octroyer 20 postes supplémentaires à la CNIL, pour un total atteignant 245 postes fin 2021.

Dans sa réponse au questionnaire, la CNIL envisage un renforcement de ses moyens à hauteur de 300 agents en 2022. Elle part en effet du constat, partagé par la mission, que les effectifs de la CNIL demeurent en-deçà du niveau minimal requis pour absorber l'ensemble des missions qui lui ont été confiées par le législateur. En particulier, l'autorité est consciente que « *la capacité de conseil aux acteurs publics et privés n'est pas adaptée au besoin exprimé par les administrations et les entreprises* ». Elle admet également que « *ces moyens limités entravent la capacité de l'institution à se projeter sur des sujets nouveaux comme l'intelligence artificielle avec les mêmes moyens que son homologue britannique* ». Elle reconnaît enfin que « *les ressources dédiées à la cybersécurité doivent aussi être renforcées, en lien notamment avec les nouvelles obligations découlant du RGPD en matière de notifications de violations de données.* »

En comparaison internationale, une CNIL à 300 agents représenterait la moitié des effectifs de l'*Information Commissioner's Office*, l'autorité britannique de protection des données (qui assure également les missions équivalentes de la CADA française). La France a un agent CNIL pour 300 000 habitants. C'est le 3ème ratio le plus bas en Europe. À titre de comparaison, l'Allemagne a 1 agent pour 80 000 habitants, les Pays-Bas ou le Royaume-Uni ont 1 agent pour 90 000 habitants ; la Pologne a un agent pour 150 000 habitants. Si la France devait se doter des mêmes moyens que le Royaume-Uni pour les besoins de régulation des données, les effectifs CADA et CNIL devraient réunir environ 735 ETP.

Sans aller jusqu'à proposer de rejoindre le standard britannique, qui nécessiterait de porter les effectifs combinés de la CADA et de la CNIL à 735 ETP, la mission considère que l'évaluation des moyens supplémentaires nécessaires réalisée par la CNIL (300 ETP en 2022) est justifiée par l'évolution de ses missions, voire sous-calibrée par rapport aux besoins, en particulier pour permettre un renforcement de son rôle d'accompagnement et pour renforcer ses compétences en ingénierie de la donnée et des systèmes d'information.

Au global, une augmentation du plafond d'emplois de la CNIL permettrait de tirer les conséquences de la hausse de l'activité, structurellement dynamique, pour augmenter et rétablir un niveau de service suffisant (notamment en matière de délais), et de rééquilibrer et de diversifier les compétences de la CNIL, en augmentant la part des ingénieurs et des auditeurs dans le total des effectifs (par exemple à 30 % contre 19 % aujourd'hui).

En tout état de cause**, la mission souhaite insister sur le fait que l'augmentation des effectifs de la CNIL doit venir renforcer le besoin d'accompagnement et de conseil des acteurs** qui souhaitent, soit mettre les données à disposition du public ou les partager dans le cadre d'un tiers de confiance (cf. partie 3), soit accéder à ces données dans le respect du RGPD **– et non pas seulement pour accroître les effectifs des équipes en charge des contrôles.**

En outre, il est à noter que les indicateurs de performance annexés au projet de loi de finance sont davantage focalisés sur la mission de contrôle et de sanction de la CNIL (et plus généralement des huit autorités administratives indépendantes financées par le programme 308 « Protection des droits et libertés »), à savoir :

- un premier objectif « Défendre et protéger efficacement les droits et les libertés », mesuré au travers du nombre de dossiers et de réclamations traités par an et par un ETP d'agent traitant, du délai moyen d'instruction des dossiers, du nombre de lieux contrôlés, du taux d'effectivité du suivi des prises de position des AAI ;
- un deuxième objectif « Éclairer la décision politique en offrant une expertise reconnue » mesuré au travers d'un indicateur sur le développement « d'une expertise reconnue permettant d'éclairer avec réactivité la décision politique ou le débat public »
- Un troisième objectif « Optimiser la gestion des fonctions support ».

Aussi, une réflexion pourrait être engagée afin de réviser ces indicateurs de performance afin de mieux mesurer, d'un point de vue qualitatif, l'activité de ces autorités par exemple en mesurant le taux de satisfaction des conseils et de l'accompagnement fourni. Cela permettrait, dans le cas de la CNIL, de mieux rendre compte de la diversité des modes d'intervention de l'autorité.

> **Recommandation** : Évaluer les besoins en ressources humaines de la CNIL pour renforcer son rôle de conseil et d'accompagnement et assortir l'augmentation des moyens correspondant d'un suivi au travers d'indicateurs de performance sur la satisfaction des usagers (dans le cadre du PLF)

Impliquer l'ANSSI, un acteur dans la régulation de la politique publique de la donnée

L'agence nationale de la sécurité des systèmes d'information (ANSSI), créée par le décret n°2009-834, est l'autorité nationale en matière de sécurité des systèmes d'information. Outre ses missions relatives à la sécurité des systèmes d'informations et à la sécurité des dispositifs et des services offerts par des prestataires nécessaires à la protection des systèmes d'information, elle apporte son soutien, dans le cadre de l'article 5 du décret n°2009-834 (dans sa version à jour du 23 novembre 2020) à la DINUM pour ce qui concerne la sécurité du réseau interministériel de l'État.

L'ANSSI a développé une doctrine particulière sur les risques liés à l'ouverture des données publics et des codes sources publics.

La sécurité des systèmes d'information des administrations, et plus largement la crainte liée aux questions de sécurité, est souvent utilisée comme argument par les administrations pour justifier un refus de communication d'une base de données publiques ou d'un code source public (cf. partie 1, paragraphe 3). Dans ce cas, l'appréciation du risque de sécurité se fait par la seule administration, sans consultation préalable de l'ANSSI, ce qui résulte en l'application d'un principe de précaution souvent sans appréciation du risque réel. De même, l'ANSSI n'est pas consultée dans le cadre des avis de la CADA et des décisions de la CNIL même lorsque l'administration refuse la publication des données ou des codes sources publics pour des raisons de sécurité.

Recommandation : Associer l'ANSSI à la mise en œuvre de la politique d'ouverture des données et des codes sources afin d'assurer que cette politique n'entre pas en contradiction avec les impératifs de sécurité des systèmes d'information :

- prévoir que la CADA et la CNIL puissent saisir l'ANSSI pour avis quand il y a un doute sérieux en matière de sécurité des systèmes d'information ;

- prévoir la possibilité, pour l'AGDAC de solliciter l'ANSSI pour un audit de bibliothèques et de logiciels libres sensibles

CAS D'USAGE – Infogreffe et les données de la justice commerciale

Les données de la justice commerciale, notamment les informations disponibles sur la plateforme Infogreffe et les décisions des tribunaux de commerce, sont parmi celles dont les acteurs rencontrés par la mission ont le plus demandé l'ouverture, soit en *open data*, soit de manière partagée, à la fois de la part des acteurs publics et des acteurs privés.

Une ouverture partielle depuis 2015

Infogreffe est un groupement d'intérêt économique (GIE) qui rassemble l'ensemble des greffiers des tribunaux de commerce de France, qui centralise les informations collectées par les greffes des tribunaux de commerce (fiche d'identité des entreprises, le K-BIS ; comptes annuels ; statuts et actes de sociétés ; historique des événements significatifs comme une procédure collective, entre autres).

Certaines données d'Infogreffe ont fait l'objet d'une ouverture en *open data*. C'est le cas des données du Registre national du commerce et des sociétés (RCNS), mis à disposition par l'INPI (Institut national de la propriété intellectuelle), chargé de la tenue du registre depuis sa création en 1954 (l'INPI archive les données des greffiers, à l'origine pour garantir la pérennité des actes en cas de sinistres dans les tribunaux). Après une mise à disposition sur demande du public, mais qui n'interdisait pas la pratique de redevances, prévue par un décret de 2014, la loi du 6 août 2015 pour la croissance, l'activité et l'égalité des chances économiques (dite « loi Macron ») a prévu l'ouverture gratuite des données du RCNS.

Les données mises à disposition par l'INPI, *via* une API ou un FTP, ne recoupent que partiellement les données disponibles à titre payant et par délivrance de documents individuels sur Infogreffe : les actes et les comptes annuels sont mis à disposition par l'INPI et Infogreffe, mais les jugements (cf. *infra*), les privilèges et nantissements, les extraits K-BIS, ou encore les procédures collectives ne font pas partie des données mises à disposition par l'INPI, dans la mesure où elles ne font pas partie du RCNS. De plus, l'INPI et Infogreffe n'ouvrent que les données non confidentielles des entreprises. Ne sont donc pas présentes les données concernant les comptes sociaux d'entrepreneurs individuels qui ont demandé l'option de confidentialité.

L'ouverture des données des greffiers des tribunaux de commerce, et en particulier des comptes annuels, a des conséquences sur le modèle économique de la profession. La diffusion des comptes annuels des entreprises représente environ un tiers du chiffre d'affaires d'Infogreffe pour un chiffre d'affaires total d'environ 65-70 M€, selon le Conseil national des greffiers des tribunaux de commerce (CNGTC).

La communauté France Digitale a indiqué que les données relatives aux entreprises mentionnées par la directive européenne de 2019 comme faisant partie des « jeux de données de forte valeur » pouvaient s'entendre comme incluant celle des greffiers des tribunaux de commerce. Ce périmètre des données de forte valeur n'est pas encore précisé, à date, par la Commission européenne.

Un *open data* des décisions des tribunaux de commerce encore loin d'être atteint

L'*open data* des décisions de justice, dont celles de la justice commerciale, est un processus qui a cours depuis plus de quatre ans et n'est pas effectif à ce jour. Aujourd'hui, la mise à disposition d'un seul jugement coûte 4,93 €. De plus, pour accéder à une décision, il est nécessaire de disposer de la juridiction ayant rendu la décision et du numéro RG de la décision. Il n'existe donc pas d'accès simple au contentieux.

Décidé en 2016 par la loi pour une République numérique, il a été programmé en 2019 par la loi de programmation 2018-2022 et de réforme pour la justice, après une mission conduite par le professeur Loïc Cadiet, concluant en 2017 au besoin d'échelonner dans le temps l'ouverture des décisions par niveau d'instance pour tenir compte des enjeux techniques de mise en œuvre de l'*open data*.

Parmi ces enjeux techniques, la capacité d'anonymiser les décisions de justice avec une garantie d'efficacité suffisante est le problème le plus aigu. Le décret n° 2020-797 du 29 juin 2020 confère la responsabilité de l'anonymisation au juge de l'espèce, mais la mise en œuvre et le calendrier d'ouverture doivent être définis par un arrêté du garde des Sceaux, qui n'a pas été pris à date.

Au début du mois de décembre 2020, selon les informations recueillies par la mission auprès du CNGTC et de la Cour de cassation, la mise à disposition des décisions des tribunaux de commerce ne semblait pas pouvoir être réalisée avant la mi-2022.

La Cour de cassation, responsable du projet d'*open data* des décisions de la justice judiciaire, prévoyait une ouverture à l'automne 2021 des arrêts de la Cour de cassation (environ 15 000 décisions par an), puis au premier trimestre 2022 des décisions des cours d'appel (environ 230 000 à 240 000 décisions par an, le stock n'étant pas considéré à ce stade), mais ne pouvait se prononcer sur la date d'ouverture des décisions de premier niveau, compte tenu du fait qu'il n'existe pas de processus d'exportation des décisions des tribunaux de commerce vers la Cour de cassation.

Les greffiers des tribunaux de commerce disposent bien de numérisations des décisions, mais seulement dans un format PDF, dont l'exploitation et l'anonymisation est donc particulièrement complexe. En outre, le nombre de décisions total n'est pas lui-même précisément connu, mais estimé entre 800 000 et 2 millions de décisions. À la fin de l'année 2020, la Cour de cassation et le CNGTC n'avaient pas été en mesure de collaborer au-delà de l'expérimentation d'un traitement sur une dizaine de décisions.

CAS D'USAGE – La base SIRENE

Une des bases les plus réutilisées de l'*open data*

Le répertoire des entreprises et des établissements, dénommé SIRENE et créé par le décret modifié n°72-314 du 14 mars 1973, enregistre l'état civil de toutes les entreprises et leurs établissements[152]. Géré par l'INSEE et mis à disposition notamment sur *data.gouv.fr*, il est considéré par l'État comme une base pivot dans le cadre de la stratégie pour un État-plateforme. La base s'articule autour du numéro SIREN, qui est un numéro à neuf chiffres permettant d'identifier les entreprises, et du numéro SIRET permettant d'identifier les établissements d'une entreprise (les neuf premiers numéros correspondent au numéro SIREN d'identification unique de l'entreprise, et les cinq suivants permettent d'identifier l'établissement).

La base SIRENE a été ouverte en janvier 2017 après l'organisation d'un hackathon, et constitue aujourd'hui une des bases de données les plus consultées et réutilisées. Avant la crise de la Covid19 et la mise en ligne des données épidémiologiques, elle était la troisième base la plus consultée sur le site *data.gouv.fr*, après le répertoire national des associations (RNA) notamment.

Les utilisateurs de la base SIRENE sont aussi bien des administrations (le ministère de l'éducation supérieure, de la recherche et de l'innovation par exemple) que des entreprises privées (Ellisphere ou inqom par exemple). Les *legaltechs* utilisent notamment la base SIRENE afin de mettre en perspective certaines informations juridiques, et proposer, par exemple, de tirer un panorama de contentieux d'une entreprise. Ce résultat peut être obtenu en croisant la base SIRENE avec d'autres bases. Ainsi, la société Doctrine produit un panorama de contentieux des entreprises en croisant les données de la base SIRENE avec les décisions de justice, les annuaires des avocats et le registre national du commerce et des sociétés.

Le nombre de réutilisateurs mensuels réguliers est passé de 500 en moyenne avant 2017 à 4 400 depuis l'ouverture, progression concomitante au développement de l'offre par API (231 millions de requêtes sur dix mois en 2020) et *via* des sélections sur différentes critères opérables sur le site SIRENE.fr (118 000 listes éditées sur 10 mois en 2020).

[152] Quelle que soit leur forme juridique, quel que soit leur secteur d'activité (industriels, commerçants, artisans, professions libérales, agriculteurs, collectivités territoriales, banques, assurances, associations…) et situés en France. Il enregistre l'état civil des organismes publics ou privés et des entreprises étrangères qui ont une représentation ou une activité en France.

Des modes de diffusion discutés par les réutilisateurs

L'ouverture de la base SIRENE avait parmi ses objectifs, celui d'augmenter sa qualité *via* les retours des réutilisateurs de la base. Toutefois, les retours observés par l'INSEE demeurent le fait d'anciens rediffuseurs de la base SIRENE[153]. De plus, l'INSEE n'est pas compétent pour décider des modifications des variables qui relèvent du traitement des formalités d'entreprises : les modifications doivent être déclarées par les entreprises par le biais de formalités déposées dans les CFE[154] [155]. Dans l'ensemble, la qualité générale de la base SIRENE semble aujourd'hui reconnue au sein de la communauté, d'après les auditions conduites par la mission.

Cependant, depuis l'ouverture, les réutilisateurs déplorent plusieurs aspects dans le mode de diffusion de la base et la relation avec l'INSEE. En particulier, plusieurs réutilisateurs déplorent un manque de communication avec l'INSEE et regrettent notamment de ne plus avoir accès à un correspondant à l'INSEE qui leur permettait de faire remonter leurs besoins dans l'ancien système.

Par ailleurs, les modifications substantielles du dispositif de diffusion pour basculer de l'ancien dispositif Syracuse vers la nouvelle offre organisée autour de l'API SIRENE, ont pu créer des difficultés pour certains réutilisateurs. Les deux dispositifs ont été maintenus en parallèle de juin 2018 à avril 2019 pour laisser le temps aux réutilisateurs d'adapter leurs systèmes d'information et d'autres évolutions sont intervenues ultérieurement pour offrir des services complémentaires. L'INSEE est à présent dans une phase de consolidation de l'offre, et d'augmentation de la disponibilité de l'API. Par ailleurs, l'INSEE considère que des activités commerciales de services spécifiques assurés pour un nombre limité d'utilisateurs est incompatible avec le principe d'égalité d'accès aux données gratuites.

Certains utilisateurs regrettent enfin, comme pour les données d'Infogreffe, un manque d'exhaustivité de l'information contenue dans la base, concernant des entreprises individuelles ne souhaitant pas que leurs données soient diffusées pour des motifs de prospection. L'INSEE ne diffuse pas ces données au secteur privé et les réserve aux administrations en ayant l'usage, jugeant que leur accessibilité était contraire aux règles du RGPD.

On peut enfin citer plusieurs marges d'amélioration dans l'exploitation de la base, comme l'interopérabilité avec la base RNA (l'identifiant RNA n'étant pas systématiquement renseigné dans la base), et le référentiel d'adresse pour l'interopérabilité avec des bases comme la base adresse nationale (BAN) ou le code officiel géographique (COG). Sur ce dernier point, l'INSEE souhaiterait au-delà du cas de SIRENE, avoir une base adresse unique, et la DINUM essaie de fédérer l'ensemble des acteurs (IGN, DGFiP et INSEE). Actuellement, l'INSEE évalue l'intérêt d'utiliser la BAN pour SIRENE.

[153] Informations communiquées à la mission par l'INSEE.

[154] Les CFE sont régis par les articles R.123-1 à R.123-30 du Code de commerce. Ils sont chargés de récolter les données sur les entreprises et de les transmettre aux organismes destinataires dont l'INSEE. Ressources : liste des CFE, page INSEE sur SIRENE et entreprise.gouv.fr.

[155] Informations communiquées à la mission par l'INSEE.

Partie 3
Pour une donnée ouverte à tous les usages

Les éléments analysés dans cette partie ne concernent pas exclusivement l'intelligence artificielle (IA), qui peut être définie comme l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence, même s'ils sont particulièrement prégnants pour cette dimension de la valorisation de la donnée.

Après une première partie rappelant les enjeux de l'ouverture des données et des codes sources, et une deuxième partie consacrée aux problématiques de l'*open data*, cette partie élargit la réflexion aux enjeux de valorisation de la donnée dans un cadre moins grand public : celui du partage entendu comme la situation où l'utilisateur possède une copie physique des données sur son serveur et celui de l'accès à la donnée entendu comme le cas où il ne peut l'exploiter que par un accès au serveur du producteur des données, sans en garder une copie physique. Pour autant, aussi bien les questions de qualité de la donnée (évoquées dans le titre 1 ci-après) que de plateformes sectorielles ou intersectorielles de données (développées dans le titre 2) sont des éléments structurants pour la stratégie de l'*open data*.

En outre, cette deuxième partie met l'accent sur intelligence artificielle car elle est un des usages les plus prometteurs de la donnée, que cette dernière soit en accès libre, partagée dans un ensemble fermé d'acteurs (publics ou privés) ou soumise à des conditions d'accès strictement définies. Pour autant, les sujets du partage et de l'accès aux données dépassent largement le domaine de l'IA.

1. Pour une donnée plus accessible et de meilleure qualité

1.1. Des données en quantité, pour développer les usages

Les données, qu'elles soient publiques ou privées, sont un facteur essentiel de production des nouvelles générations de système d'IA. Même s'il ne faut pas sur-simplifier, le paradigme contemporain le plus actuel de l'IA repose sur l'extraction de connaissance à partir des données, et sur l'automatisation de cette extraction (vision par ordinateur, traitement automatique du langage naturel, autres types de données et signaux). Certes, pour progresser encore, la représentation des connaissances et l'automatisation de la logique resteront importantes, mais on ne peut figurer en bonne place dans la compétition technologique de l'IA sans les données.

Des données en quantité pour développer les usages

Grâce à la puissance des ordinateurs, il est possible de traiter un nombre considérable de données dans des délais très courts. C'est ce qui explique l'explosion de l'IA via des techniques d'apprentissage (dites de *« machine learning »*). Avec des algorithmes simples basés sur des réseaux de neurones au sens informatique du terme, le logiciel se perfectionne grâce à des jeux de données qui lui permettent d'affiner ses paramètres. Plus il « apprend » de ces données, plus il s'améliore. C'est le cas pour la reconnaissance faciale ou les outils d'aide à la décision.

Pour donner une idée du volume de données nécessaire au développement de l'IA, on peut citer l'exemple du télescope Hubble, qui transmet chaque jour 120 Go ($1{,}2 \times 10^{14}$ octets) d'images que les astronomes du monde entier ont utilisées pour rédiger plus de 17.000 publications scientifiques.

Un autre exemple illustrant l'enjeu de massification des données est le projet « signaux faibles ». Ce projet permet, grâce à un outil de *data science* et d'intelligence artificielle, de détecter les entreprises fragilisées afin de mieux les accompagner. Il s'appuie, entre autres, sur des données issues, au niveau local, de l'union de recouvrement des cotisations de sécurité sociale et d'allocations familiales (URSSAF), des directions régionales des entreprises, de la concurrence, de la consommation, du travail et de l'emploi (DIRECCTE), de la base de données de la société Altares, de la base SIRENE, de Diane (outil d'analyses économiques de la société Ellisphere) et de la Banque de France. Cela a permis de générer une base de données considérable, avec des mises à jour quasi quotidiennes, sur lesquels peuvent être faites des statistiques prévisionnelles grâce à une IA de type *« machine learning »*. On voit l'importance du projet quand on sait que, chaque année, ce sont entre 50.000 et 60.000 entreprises qui se trouvent en situation de défaillance.

Dans le domaine de la santé, c'est par le traitement et le croisement d'un grand volume de données de qualité, que les recherches à plus grand impact peuvent être menées, par exemple pour améliorer le dépistage et le diagnostic d'une maladie, analyser les effets secondaires des traitements, faire évoluer les essais cliniques. C'est dans cette perspective qu'a été créé la plateforme des données de santé (ou Health Data Hub) avec pour objectif de faciliter l'accès au maximum de sources de données médicales possibles.

Les données publiques et d'intérêt général peuvent donc être mises au service de la data science et de l'IA, qu'il s'agisse de projets publics, privés ou mixtes. Une grande partie de l'IA se développe sur des ressources ouvertes, des jeux de données d'entraînement, des modèles pré-entraînés et des logiciels libres.

Des ressources à développer pour asseoir une autonomie stratégique

De nombreux jeux de données, modèles et logiciels sont développés par les anglo-saxons mais la France dispose de certaines ressources qui pourraient être davantage valorisées pour développer l'IA francophone, notamment la librairie Scikit Learn (cf. partie 1), OSCAR (Inria) et les modèles ouverts CamemBERT et FlauBERT.Le Lab IA d'Etalab a mené également un projet pilote pour créer le premier jeu de données de questions réponses francophones (Projet PIAF - Pour une IA francophone)[156]. PIAF a permis avec des ressources modestes (de l'ordre de 300 000 €) de créer une plateforme d'annotation ouverte, un jeu de données de 10 000 questions réponses « *crowdsourcé »* et ouvert[157]. Ce jeu de données est utilisable par chacun pour développer des algorithmes de questions réponses. Une entreprise, Illuin Tech, a mené un projet similaire en parallèle dénommée FQUAD[158].

Pour développer l'autonomie stratégique en matière d'IA, il est nécessaire de donner de la visibilité aux ressources existantes et d'encourager le développement de nouveaux jeux de données d'entraînements et de modèles pré-entraînés ouverts.

[156] https://piaf.etalab.studio/francophonie-ia/

[157] https://www.data.gouv.fr/fr/datasets/piaf-le-dataset-francophone-de-questions-reponses/

[158] https://fquad.illuin.tech/

1.2. Des données de qualité au service des différents usages

Des données devant respecter les principes généraux FAIR

La qualité fait d'abord référence à des données respectant les principes FAIR[159] (Findable, Accessible, Interoperable, Reusable) qui fonctionnent comme une ligne directrice pour ceux qui veulent atteindre cet objectif. La Commission Européenne a repris ces principes et demande à tous les projets financés par le programme H2020 de les suivre (cf. encadré).

Les principes FAIR

Les caractéristiques des données « *Findable* » :

- les données et les métadonnées sont identifiées par un identifiant global unique et pérenne ;
- les métadonnées décrivant les données sont riches ;
- les données et les métadonnées sont enregistrées et indexées dans un dispositif permettant de les rechercher ;
- les métadonnées spécifient l'identifiant de la donnée.

Les caractéristiques des données « *Accessible* » :

- les données et les métadonnées sont accessibles par leur identifiant via un protocole de communication standardisé.
- les métadonnées sont accessibles même quand les données ne le sont plus.

Les caractéristiques des données « *Interoperable* » :

- les données et les métadonnées utilisent un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances ;
- les données et les métadonnées utilisent des vocabulaires qui respectent les principes FAIR ;
- les données et les métadonnées incluent des liens vers d'autres (méta)données ;

Les caractéristiques des données « *Reusable* » :

- les données et les métadonnées ont des attributs multiples et pertinents ;
- les données et les métadonnées sont mises à disposition selon une licence explicite et accessible ;
- les données et les métadonnées sont associées à leur provenance ;
- les données et les métadonnées correspondent aux standards des communautés indiquées.

Source : Projet GO-FAIR.

Le projet européen GO FAIR[160] a pour objectif d'ouvrir progressivement les données de la recherche en suivant les principes FAIR au sein des institutions scientifiques et académiques dans tous les domaines de la recherche et au-delà des frontières nationales. GO FAIR est une initiative des Pays-Bas, de l'Allemagne et de la France qui propose le développement d'un environnement international de recherche enrichi par les données.

Une qualité à adapter aux usages de la donnée

Une donnée, pour être vraiment réutilisable, doit avoir des qualités d'éligibilité, c'est-à-dire que sa structure doit être compatible avec l'application qui va servir à son utilisation, et des qualités d'efficacité.

[159] https://www.pasteur.fr/fr/file/19235/download, https://www6.inrae.fr/datapartage/Produire-des-donnees-FAIR

[160] https://www.go-fair.org/fair-principles/

L'efficacité d'une donnée pour une réutilisation dépendent de critères qualité intrinsèques dont la pertinence varie en fonction des finalités de la réutilisation : la qualité d'une donnée n'est pas un absolu, elle se conçoit au regard de la finalité des traitements qui la consomment. Il n'existe ainsi pas d'unique liste de critères de qualité, mais une multitude, en fonction de la nature des données et des traitements.

Une typologie de ces critères existe cependant. Le « cadre commun d'architecture des référentiels de données » publié en 2013 par la direction interministérielle des systèmes d'information et de communication (DINSIC)[161] propose de les réunir au sein de trois critères :

- des critères intrinsèques : unicité de représentation, complétude (exhaustivité ou vis-à-vis du processus de collecte), exactitude, conformité, intégrité des relations entre objets décrits, cohérence ;
- des critères de service : accessibilité, cohérence dans le temps, actualité (fraicheur des données), pertinence ;
- des critères de sécurité : disponibilité, intégrité (pas d'altération), confidentialité, traçabilité (apporter la preuve d'un traitement), lisibilité

Certains critères peuvent parfois s'opposer. Par exemple la recherche d'une trop grande robustesse peut se traduire par des temps d'investigation et de traitement trop longs qui nuisent à la fraicheur. Une trop grande finesse de détail (résolution d'image, grain territorial, etc.) peut exiger de très longs traitements pour une finalité qui n'en a pas besoin.

L'amélioration de la qualité d'un jeu de donnée a un coût. L'investissement nécessite de connaître *a priori* les bénéfices à en attendre, par la connaissance des cas d'usage, quand cela est possible. De nombreuses administrations rencontrées publient leurs données en l'état (comme le dispose la loi pour une République numérique) sans s'inquiéter des cas d'usage et donc des critères qualités requis afin que cette donnée soit réappropriée. Des projets d'amélioration des données dans le but de fédérer les usages existent néanmoins. Par exemple, le projet de représentation parcellaire cadastrale unique (RPCU) de l'institut national de l'information géographique et forestière (IGN) et de la DGFIP qui vise à donner au plan cadastral une exactitude géométrique qu'il ne peut garantir actuellement, celui-ci étant conçu pour garantir une intégrité des surfaces des parcelles suffisantes pour les besoins de la DGFIP mais pas une localisation précise. Une inexactitude de plusieurs dizaines de mètres peut être constatée sur certaines parcelles. Cette inexactitude, quand elle se produit, peut rendre l'utilisation du plan cadastral trompeur, dans le cas d'espèce si l'on souhaite superposer l'assiette d'une servitude d'utilité publique sur le dessin d'une parcelle d'habitation[162].

Aujourd'hui, le manque de qualité est souvent une pauvreté d'information, ce qui nuit à la richesse de la donnée pour certains usages, notamment en matière d'IA. Plusieurs administrations ayant répondu au questionnaire de la mission font valoir que ce défaut de qualité est lié à l'application des secrets légaux et de l'anonymisation. Les processus d'anonymisation et de protection des secrets peuvent en effet parfois dégrader la qualité, ce qui plaide pour distinguer la mise à disposition des données en *open data*, où cette anonymisation et cette protection ne peuvent être remis en cause, et la mise à disposition dans des cadres sécurisés et sans ouverture publique des données. Cet accès sécurisé est nécessaire à l'IA, en cas d'utilisation de données sensibles : toute mise en œuvre d'applications, y compris à base d'IA, tire avantage de la connaissance fine de la qualité des données qu'elle consomme et les principes FAIR poussent à les renseigner correctement dans les métadonnées.

[161] Cadre commun d'architecture des référentiels de données, complément n°2 au cadre commun d'urbanisation du système d'information de l'État version 1.0, Direction interministérielle des systèmes d'information et de communication (DISIC), version n°1.0 du 18 décembre 2013.

[162] Voir l'exemple cité à l'annexe 1 du rapport au Gouvernement de Valéria Faure-Muntian, *Les données géographiques souveraines*, daté de juillet 2018.

Dans les domaines pauvres en données ou dans ceux dont les données sont soumises à la protection des divers secrets (données personnelles, secrets des affaires, etc.), posséder des données, même de mauvaise qualité, présente malgré tout un intérêt y compris pour alimenter l'IA. Par exemple, le service d'administration nationale des données et référentiels sur l'eau (SANDRE) diffuse des rapports de contrôle de certaines bases de données du système d'information de l'eau (SIE) à partir de ses propres outils de contrôle des données tels que l'intégrité ou la conformité[163]. Ces rapports permettent aux éventuels réutilisateurs de prendre connaissance des anomalies constatées.

Un besoin de normalisation à des fins d'interopérabilité des données

L'interopérabilité constitue une caractéristique essentielle pour une démarche de rapprochement, d'appariement, d'enrichissement de données et donc pour alimenter l'IA. Elle recouvre la capacité à agréger des données issues de sources différentes et nécessite que les données convergent sur des structurations compatibles, en particulier à travers :

- une sémantique et une syntaxe partagée ;
- une structuration commune pour des données semblables ;
- des dictionnaires et registres pour remplir les champs des données et des métadonnées ;
- une même projection cartographique pour les données géographiques ;
- un même carroyage ou des carroyages compatibles pour des données agrégées ou résultantes d'opérations statistiques.

Le processus de standardisation qui définit ces éléments est au cœur de l'écosystème des données car il en conditionne le potentiel de valorisation. C'est un processus qui doit recueillir le consensus pour qu'un standard soit adopté par l'ensemble des producteurs, *a minima* d'un même secteur, à moins de leur être imposé par un texte règlementaire (c'est le cas du standard qui encadre les données publiées par les collectivités sur le Géoportail de l'urbanisme) ou par l'usage lorsqu'une plateforme publique ou privée détient un monopole.

Le constat fait par la mission montre que les marges de progression sont importantes.

D'une part, la standardisation de la donnée n'est pas qu'une affaire d'experts, réservée aux spécialistes de la donnée et aux métiers utilisateurs, mais appelle aussi une vision de la politique publique au service de laquelle la donnée est utilisée, pour sa partie sémantique. L'exemple de l'évaluation de l'artificialisation des sols le montre bien : cette évaluation a longtemps été freinée par l'absence d'une définition partagée entre tous les acteurs de ce qu'est une surface artificialisée. L'observatoire de l'artificialisation[164] est une action récente mise en œuvre par la DGALN avec le CEREMA, l'IGN et l'INRAE en 2019, prévue par le plan biodiversité, qui *« vise à documenter les données utiles au suivi de l'artificialisation des sols et de la consommation d'espace ».*

D'autre part, le référencement, la création et la validation de schéma de données est capital lorsque plusieurs producteurs de données produisent des jeux de données sur un même sujet, afin que ces jeux de données puissent être facilement croisés (par exemple des données de marchés publics, des lieux de stationnement ou encore des bases adresses locales). Ces schémas visent à décrire de manière précise et univoque les différents champs qui composent un jeu de données et les valeurs possibles. Il n'est pas possible de dresser un bilan de la situation en matière de schémas de données car cela supposerait une analyse secteur par secteur. Pour certains systèmes d'information, la problématique du référencement est d'ores et déjà prise en compte comme le SIE (Système d'information sur l'eau) où le SANDRE[165] est établit et met à disposition le référentiel des données sur l'eau constitué de spécifications techniques et de listes de codes libres d'utilisation et décrivant les modalités d'échange des données sur l'eau à l'échelle de la France.

[163] http://mdm.sandre.eaufrance.fr/geo/rapportsv3

[164] https://artificialisation.biodiversitetousvivants.fr/sites/artificialisation/files/inline-files/observatoir_artificialisation_flyer.pdf

[165] https://www.sandre.eaufrance.fr/missions-et-organisation-du-sandre

Au sein du conseil national pour l'information géographique (CNIG), la conception d'un standard fait l'objet d'un groupe de travail ouvert. Le résultat fait l'objet d'un appel à commentaires public, chaque commentaire est ensuite discuté et commenté. Le document final est validé en commission. Le processus qui doit garantir le plus grand consensus se déroule sur plusieurs mois.

Au sein de l'infrastructure mise en œuvre par la DINUM, *schema.data.gouv.fr* [166] a pour objectif d'être une plateforme de partage et d'échange sur le référencement, la création et la validation de schéma de données. Il est possible d'y consulter l'ensemble des schémas de données référencés, de proposer un nouveau schéma de données, de tester la conformité d'un jeu de données à un schéma de données (Validata) et de créer un jeu de données conformément à un schéma (CSV-GG). Des travaux sont en cours pour intégrer au mieux ces outils à la plateforme *data.gouv.fr*.

Un catalogage des données à consolider

Par ailleurs, la mission fait le constat que le catalogage, au niveau agrégé, des données disponibles en *open data* est particulièrement problématique. En effet, à chaque infrastructure appartient un catalogue des données hébergées. Celui-ci est produit à partir des métadonnées accompagnant les jeux de données. Or si l'on observe l'ensemble des infrastructures de données comme une infrastructure nationale, il apparaît impossible de constituer un catalogue unique des données publiques ouvertes ou non, notamment du fait de redondances des données dans différents catalogues, d'obsolescence et de métadonnées déficientes.

Il s'agit de répondre aux questions : - existe-t-il une base de données qui répond à mon besoin ? – où se trouve sa version la plus aboutie/légitime ? – qui en est propriétaire ?, quelles sont les informations de qualité, de confidentialité, de licence ? Ces informations devraient être disponibles pour les bases de données publiques ouvertes comme pour les bases de données publiées en diffusion restreintes.

Ainsi, la présence de données dans *data.gouv.fr* ne signifie pas qu'il s'agit de la base de données la plus fraiche, la plus légitime. Par exemple, par le moissonnage des catalogues de données des DDT, de nombreux documents d'urbanisme sont accessibles sur *data.gouv.fr*. Toutefois, il n'est pas garanti qu'il s'agit des documents les plus récents qui font foi, ceux-ci étant disponibles en mairie.

Ainsi, lors du hackathon sur les données d'urbanisme organisé en février 2017 par le ministère en charge du logement, des aménageurs témoignaient de la part trop importante de temps consacré à la recherche des données les plus fiables lors d'une étude d'aménagement qu'ils estimaient à 80 %.

[166] Les schémas de données permettent de décrire des modèles de données : quels sont les différents champs, comment sont représentées les données, quelles sont les valeurs possibles etc.

Les concepts de moissonnage et de syndication

Les nombreuses infrastructures décrites dans ce rapport ne sont pas toutes indépendantes, il existe de nombreux liens permettant sous une forme normalisée d'échanger des informations de métadonnées et d'échanger des données.

La syndication permet à une infrastructure d'informer des « abonnés » des changements concernant ses bases de données. Ainsi, il n'y a pas besoin de consulter régulièrement un portail ou une plateforme pour savoir ce qui a changé ou l'ajout d'une nouvelle donnée, cette information est récapitulée dans une information envoyée par l'infrastructure. En ce sens, le Géoportail de l'Urbanisme mis en place par la DHUP informe ses abonnés des nouvelles publications de documents d'urbanisme sur le territoire national à l'aide de flux conformes au protocole ATOM[167].

Le moissonnage permet à une infrastructure de charger automatiquement des ressources (métadonnées, données) à partir du catalogue de données d'une autre infrastructure. Là aussi, il existe des protocoles bien rodés. Ainsi *data.gouv.fr* propose aux administrations de moissonner leurs sites[168]. Via les métadonnées, on peut construire un catalogue « général » de données par moissonnage. C'est le cas du Géocatalogue mis en œuvre par le BRGM pour le compte de l'infrastructure voulue par la directive INSPIRE. Il n'y a pas besoin de charger les données : la connaissance des métadonnées permet à une infrastructure d'accéder aux données d'une autre infrastructure selon ses besoins par l'intermédiaire d'API.

Le graphe ci-après, construit par le projet de recherche GEOBS[169] illustre l'activité de moissonnage entre IDG (infrastructures de données géographiques) en France en 2018. On perçoit le caractère fractal des infrastructures de données mais aussi l'activité intense d'échange entre ces infrastructures.

[167] https://www.geoportail-urbanisme.gouv.fr/image/UtilisationATOM_GPU_1-0.pdf

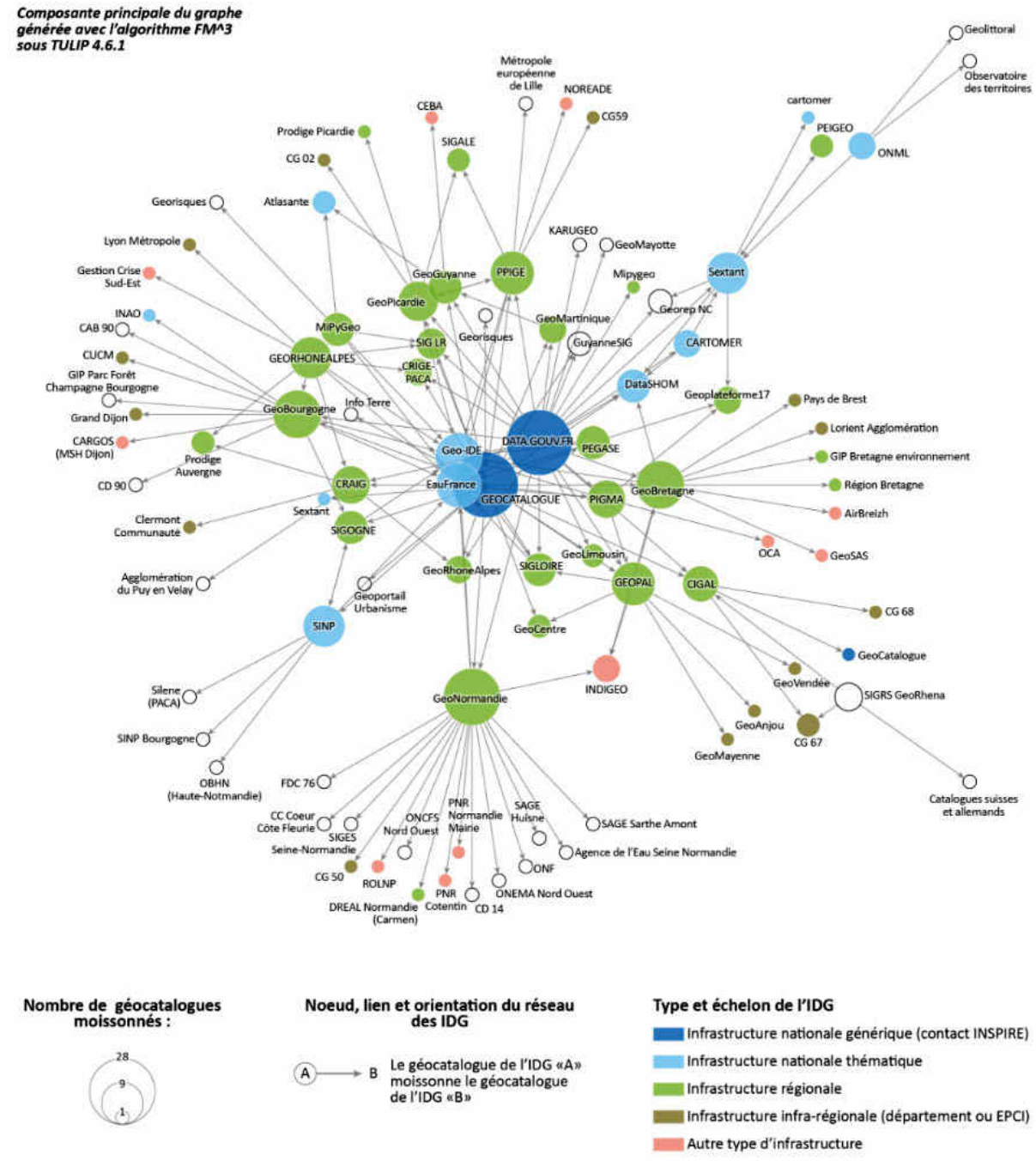
[168] https://doc.data.gouv.fr/jeux-de-donnees/demander-a-datagouvfr-de-moisonner-votre-site/

[169] https://www-iuem.univ-brest.fr/pops/attachments/1746 , voir aussi Adeline Maulpoix, Matthieu Noucher, Olivier Pissoat, Grégoire Le Campion, Françoise Gourmelon, et al.. Enquête 2017 auprès des coordinateurs des Infrastructures de Données Géographiques en France.

Rapport intermédiaire du projet de recherche GEOBS. [Rapport de recherche] Passages UMR 5319;

LETG - Brest Géomer; PRODIG. 2017. halshs-01635844

Schéma du moissonnage des infrastructures de données géographiques (DIG)



Source : Projet Geobs (CNRS 2017).

Dans ce contexte, la mission souhaite mettre l'accent sur la nécessité de combiner deux approches : d'une part la définition au niveau de la DINUM d'une politique interministérielle d'interopérabilité et de qualité de la donnée et d'autre part, sa mise en œuvre au niveau de chaque communauté de travail, pour avoir des règles communes de travail acceptées par tous.

> **Recommandation :** Définir et mettre en œuvre une politique interministérielle d'interopérabilité et de qualité de la donnée (démarches de standardisation, notamment des codes sources, label FAIR, doctrine sur les métadonnées, catalogage)

> **Recommandation :** Encourager les écosystèmes à définir des principes de gouvernance de la qualité, en désignant un référent qualité et en créant des communautés de réutilisation avec participation active des producteurs de la donnée

Les API : un outil utile, mais qui ne doit pas conduire à limiter les réutilisations

Le choix du mode de diffusion de la donnée détermine beaucoup la liberté de réutilisation et la variété des usages qui peuvent en être fait. En particulier, la diffusion peut se faire sous forme de formats « plats » (données dans un tableau par exemple) ou par le biais d'une API (*Application Programming Interface*), interface qui rend disponible les données et des fonctionnalités d'une application existante, à destination d'une application cliente. Physiquement, les bases de données sont stockées à distance (chez leur propriétaire par exemple) et l'accès aux données souhaitées se fait par le biais de l'API, au moment où l'application cliente en a besoin, par requête, selon des protocoles.

Le choix de recourir à une API est le plus souvent justifié par des contraintes techniques, comme la volumétrie de la donnée mise à disposition et des réutilisations, le degré de sécurité, la fréquence de mise à jour ou la granularité des données (par exemple dans le cas de la base SIRENE, exploitée par de nombreux réutilisateurs, comme exposé dans le cas d'usage spécifique).

L'API présente ainsi à la fois des avantages et des inconvénients pour l'utilisateur, et son utilisation doit donc être justifiée par ce bilan d'opportunité, entre la facilitation du service, et la restriction des usages. Si la donnée est un bien non rival, c'est-à-dire que son utilisation par un nouvel utilisateur ne limite pas l'utilisation qui est faite par les autres utilisateurs, ce n'est pas vrai de l'API, sorte de « pont du XXI^{ème} siècle », infrastructure d'accès à la donnée dont l'utilisation engendre des coûts. Par ailleurs, il convient de souligner que l'API suppose d'avoir les moyens de recourir à une application qui interroge l'API et récupère les données, ce qui n'est pas à la portée de l'usager non expert, mais ce qui facilite l'appropriation de la donnée pour les réutilisateurs qui souhaitent exploiter la donnée par le biais d'une application.

Le choix d'une API détermine notamment la disponibilité des champs de données et les séries temporelles. En mettant à disposition les données les plus « fraîches », elle permet une mise à jour en continue des données pour l'utilisateur, mais elle rend aussi impossible l'accès aux anciennes données, sauf dans le cas où les paramètres de l'API le permettent. Par exemple, dans le champ de l'environnement, Hub'eau est un portail d'API qui expose de façon simplifiée les données du SI eau. L'API « Hydrométrie » expose les observations de hauteur d'eau du réseau de mesure français, à partir de la plateforme du service central d'hydrométéorologie et d'appui à la prévision des inondation (SCHAPI), mises à jour toutes les deux minutes, avec 24 heures de profondeur, et maintient un historique d'un mois. Cela permet le développement d'applications visant à informer en temps réel des risques d'inondation.

En théorie, une API pourrait permettre au client d'accéder à l'intégralité des champs, sur l'intégralité de la période temporelle et l'intégralité du champ géographique disponible, mais ce service représenterait un coût en infrastructure disproportionné par rapport au besoin. Ainsi, la structure des données transmises par API est le plus souvent simplifiée et se concentre sur les champs identifiés par les principaux usages. Cela permet aussi d'en simplifier l'usage par les développeurs, qui ne sont pas obligés de maîtriser le schéma originel des données pour les exploiter.

L'API permet également de suivre et de réguler les accès des utilisateurs, et présente à cet égard l'avantage pour le diffuseur de pouvoir gérer des fonctions de sécurité ou de contrat, en exigeant que ces informations juridiques soient envoyées dans la requête (URL, IP, clé de licence, par exemple). L'API permet également de gérer les débits d'un client afin d'éviter la saturation des services de diffusion. Elle permet ainsi de garder une trace des appels et de mieux comprendre l'usage des données, ce que ne permet pas une diffusion de fichiers, qui plus quand quand ils peuvent être à leur tour rediffusés.

À titre d'exemple, Pôle Emploi a présenté à la mission une « trajectoire d'APIsation », qui répond à la fois « *aux orientations stratégiques concernant l'évolution des offres de services de Pôle emploi et à la nécessaire synergie avec [ses] partenaires de proximité qui agissent en complémentarité au bénéfice [des] usagers* », comme les besoins d'échanges liés à l'emploi des jeunes (API avec les missions locales) ou à la mobilité des demandeurs sur les territoires (API pour favoriser l'accès aux aides aux transports).

Pour pallier ces inconvénients, en particulier l'impossibilité de reconstituer la base de données complète (avec tous ses objets et ses champs), il semble raisonnable d'envisager toujours, en complément de la mise à disposition par API, la mise à disposition de la base de données en format de fichier « plat », au moins en partie (comme c'est le cas pour la base SIRENE).

Finalement, recourir à une API n'est pas qu'un enjeu d'infrastructure, mais un enjeu de gouvernance : le diffuseur et le producteur de la donnée doivent être capables de connaître les besoins et les usages de la donnée, et construire le canal de diffusion en fonction de ces besoins, pour calibrer la disponibilité de la machine en fonction des besoins les plus importants**.** La mission considère ainsi que les choix des modes de diffusion doivent être dûment justifiés et proportionnés, pour ne pas limiter la capacité de réutilisation.

2. Des « hubs » indispensables mais qui doivent être interopérables

Valoriser la donnée, notamment dans le domaine de l'IA, suppose dans une majorité des cas de recourir à des données détenues par un tiers. Dans certains cas, l'ouverture des données et des codes sources au public n'est pas possible, ni même souhaitable (pour des raisons notamment de protection des données personnelles, de protection du secret –statistique, fiscal, médical, judiciaire, des affaires, etc.). Il convient alors de prévoir un cadre de partage des données intermédiaire entre l'ouverture au public et la fermeture des données : cela pose la question des moyens techniques de partage (le réutilisateur traite et conserve les données) ou d'accès (le réutilisateur ne conserve pas les données mais uniquement le résultat du traitement). Les conditions de ce partage ne sont pas neutres : il s'agit souvent de données à caractère personnel, nécessitant des mécanismes de protection et de sécurité adaptés.

Le terme *« hub »* ou plateforme de données est polymorphe. Il est souvent utilisé pour désigner une infrastructure de stockage qui regroupe un ensemble de données en provenance de systèmes d'informations multiples. En permettant de regrouper toutes les données de façon centralisée, le *hub* facilite l'accès aux données et leur utilisation de façon sécurisée. Le *hub* propose aussi généralement des outils de traitement et d'analyse de données. Dans ce paragraphe, la mission utilise ce terme dans une acception plus large pour qualifier l'organisation mise en place destinée à fédérer le partage et/ou l'accès à la donnée autour de problématiques communes à un groupe d'acteurs sectoriels ou intersectoriels, au-delà de la composante infrastructures de données évoquées en partie 4 (la mission aborde le cas du CASD, qui est transectoriel, dans la partie 4).

Dans cette acception comprenant le volet de services et d'animation d'une communauté (de recherche, d'acteurs d'une filière économique), la mission a recensé, parmi les grands domaines de politique publique et de manière non exhaustive, huit hubs ou plateformes[170], tous de création récente ou encore à l'état de projet. Elle propose de les répartir en deux groupes selon leur origine et les conditions de leur émergence : ceux issus d'une initiative des pouvoirs publics essentiellement à visée de recherche, et ceux « portés » par les acteurs d'une filière essentiellement destinée à maîtriser l'usage et la valorisation de leurs données.

[170] Le projet **Artémis** (architecture de traitement et d'exploitation massive de l'information multi-sources), piloté par le ministère des Armées est un autre programme phare mais entrant davantage dans la catégorie des infrastructures seules. Il vise à doter les armées d'une infrastructure souveraine de stockage et de traitement massif de données en matière de *big data* et d'IA. Cette infrastructure est constituée d'un socle technique destiné à développer des cas d'usage décentralisés identifiés et opérés par les métiers. Elle vise à la mise en place d'un écosystème permettant aux innovateurs d'apporter leurs créations et de les faire murir jusqu'à des solutions utilisables par les forces armées grâce à une gamme de produit complète (kit de développement et c*loud* dédié pour accompagner les développeurs, version Lab' pour les évaluations par le ministère et les systèmes déployés sur les réseaux classifiés).

Exemples de hubs et projets de hubs de données sectoriels ou intersectoriels

	Health data hub	*Green data hub*	**Energy for Climate (E4C)**	**AgDataHub**	**Alliance Culture Data**	**Apidae tourisme**	**Numalim**
Domaine(s)	Santé	Environnement/santé	Energie/climat	Agriculture	Industries culturelles et créatives	Tourisme	Agroalimentaire
Initiative/ origine	Loi du 24 juillet 2019[171] Arrêté du 29 nov.2019	Publique (action n°17 du PNSE4 actuellement en consultation)	Institut polytechnique de Paris/Ecole des Ponts	Acteurs de la filière (syndicats, chambres consulaires, etc.)	Acteurs de la filière (projet porté par BnF participation)	Acteurs de la filière tourisme en Rhone-Alpes	Acteurs de la filière et états généraux de l'alimentation
Finalités et missions	Réunir, organiser et mettre à disposition des données de santé Diffuser les normes de standardisation pour l'échange et l'exploitation des données de santé Accompagner les porteurs de projets Informer les patients	Réunir, organiser et mettre à disposition des données santé/environnement Faciliter les croisements de données environnementales et sanitaires dans un but de progression des connaissances, et de la recherche, Informer le public	Réunir, organiser et mettre à disposition des données interdisciplinaires pour favoriser les travaux de recherche et l'émergence de solutions opérationnelles pour répondre aux défis posés par le changement climatique	Mettre en place une infrastructure technologique mutualisée associée à une démarche de standardisation collective et structurée dans le but de maîtriser les usages de leurs données pour les producteurs agricoles, permettre l'interconnexion de multiples sources de données, et de partager des modèles et outils innovants	Décloisonner l'accès aux données des industries culturelles et créatives pour mieux les valoriser Fournir des services amont (aux fournisseurs de données) et aval (aux acquéreurs) pour valoriser ces données	Mutualiser les coûts de constitution d'une base de données ; Favoriser le rapprochement entre acteurs de la filière pour valoriser la donnée	Réunir et mettre à disposition des données sur les productions agroalimentaires françaises (composition, conditions de production, etc.) Encourager la création de valeur par les données Informer le public

[171] Relative à l'organisation et à la transformation du système de santé.

Cadre juridique	Groupement d'intérêt public (GIP)	Non arrêté à ce jour. Choix multiples en fonction du dispositif et de la gouvernance du GD4H qui seront considérés.	Non arrêté à ce jour.	SAS (actionnaires : organismes privés ou parapublics)		Société coopérative (Scic SA)	Société coopérative (Scic SAS)
Modalités d'ouverture (ouvert, partage, accès)	Accès distant à des bureaux virtuels	Ouvert/partage ou accès selon les données et les utilisateurs	Ouvert, partage ou accès selon le type de données	Ouvert/partage/accès selon les choix que feront les fournisseurs de données qui seront présents	Ouvert/partage/accès selon les choix que feront les fournisseurs de données qui seront présents (non confirmé à ce jour)	Partage entre ses membres	Partage (opendata indirectement via les applications qui reprendront ces données)
Infrastructures	Plateforme Cloud Centralisée (création d'un entrepôt des données de santé)	Non arrêté à ce jour.	Mutualisées Infrastructure ESPRI de l'institut Pierre Simon Laplace Pour les données sensibles en accès restreint, le CASD	Plateforme Cloud Centralisée	Non arrêté à ce jour.	En propre. Centralisée.	Mixte (centralisée, décentralisée)
Modèle économique	Gratuit ou pouvant être payant pour des acteurs privés	Non arrêté à ce jour.	Tarification en cours d'élaboration	Abonnement (différents niveaux)	Non arrêté à ce jour.	Abonnement (différents types)	Abonnement, commission (place de marché), formation

Source : Mission

2.1. Les *hubs* issus d'une initiative des pouvoirs publics autour des enjeux de recherche

Le Health Data Hub (HDH) ou plateforme des données de santé

Cette plateforme doit permettre à la France de devenir le leader mondial de l'IA en santé et faciliter l'accès aux données de santé pour des projets d'intérêt général innovants de la recherche et développement en matière médicale. Sa création a été actée à la suite du rapport Villani de mars 2018, la santé étant identifié comme un des secteurs prioritaires pour le développement de l'intelligence artificielle en France[172].

Le HDH est un groupement d'intérêt public créé par la loi du 24 juillet 2019 relative à l'organisation et la transformation du système de santé ("Ma Santé 2022"). Il associe 56 parties prenantes, parmi lesquelles la caisse nationale d'assurance maladie (CNAM), le centre national de recherche scientifique (CNRS), la Haute autorité de santé, France Assos Santé, au sein de son assemblée générale, majoritairement publiques, représentant toute la diversité de l'écosystème des données de santé.

Il met en œuvre les grandes orientations stratégiques relatives au Système National des Données de Santé (SNDS) fixées par l'Etat et notamment le ministère des Solidarités et de la Santé. Élargi à un grand nombre de sources de données de santé, notamment cliniques, le SNDS comprend non seulement les données médico-administratives, mais également des données de registres, de cohortes de recherche, d'entrepôts de données hospitalières, etc.

L'article L. 1462-1 du Code de la santé publique détaille ses missions[173].

Concrètement il prend la forme d'un guichet unique permettant un accès aisé et unifié, transparent et sécurisé aux données de santé dans le respect du droit des patients. Ces données sont hébergées sur une plateforme sécurisée qui met à disposition des porteurs de projets des ressources technologiques et humaines mutualisées.

---

[173] *« 1° De réunir, organiser et mettre à disposition les données du système national des données de santé mentionné à l'article L. 1461-1 et de promouvoir l'innovation dans l'utilisation des données de santé ;*

2° D'informer les patients, de promouvoir et de faciliter leurs droits, en particulier concernant les droits d'opposition dans le cadre du 1° du I de l'article L. 1461-3 ;

3° D'assurer le secrétariat unique mentionné à l'article 76 de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés ;

4° D'assurer le secrétariat du comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé [CESREES] ;

5° De contribuer à l'élaboration, par la Commission nationale de l'informatique et des libertés, de référentiels et de méthodologies de référence au sens du b du 2° du I de l'article 8 de la loi n° 78-17 du 6 janvier 1978 précitée. Il facilite la mise à disposition de jeux de données de santé présentant un faible risque d'impact sur la vie privée, dans les conditions prévues au II de l'article 66 de la même loi ;

6° De procéder, pour le compte d'un tiers et à la demande de ce dernier, à des opérations nécessaires à la réalisation d'un traitement de données issues du système national des données de santé pour lequel ce tiers a obtenu une autorisation dans les conditions définies à l'article L. 1461-3 du présent code ;

7° De contribuer à diffuser les normes de standardisation pour l'échange et l'exploitation des données de santé, en tenant compte des standards européens et internationaux ;

8° D'accompagner, notamment financièrement, les porteurs de projets sélectionnés dans le cadre d'appels à projets lancés à son initiative et les producteurs de données associés aux projets retenus ».

Le catalogue proposé par le HDH correspond à une collection de bases de données issues du SNDS que la plateforme est habilitée à mettre à disposition. Parmi ces bases, le SNDS « historique », peut toujours être accédé par le portail de l'Assurance Maladie. Le catalogue rassemblera des copies de bases déjà existantes pour la plupart, jugées les plus pertinentes pour la recherche et l'innovation. C'est un comité stratégique piloté par l'Etat qui fixera les grandes orientations de ce catalogue qui se veut évolutif. Cette collection est construite de manière progressive et itérative, en partenariat avec les responsables de la collecte des données concernées. En créant ce système de bases de données, la loi permet une meilleure visibilité du patrimoine commun pour tout l'écosystème, une homogénéisation des règles d'accès aux données et une facilitation de ces accès avec la mise en place du HDH comme service sécurisé de mise à disposition du catalogue.

Son financement est majoritairement public.

Le projet de Green Data Hub *dans les domaines environnement et santé*

Le *Green Data Hub* (dont la dénomination n'est pas arrêtée) est un projet de *hub* qui vise à disposer d'un espace commun de données environnementales au service de la santé (un intitulé de type *Green Data for Health* – ou GD4H – serait par exemple plus exact). Il s'agit de répondre à la demande d'accès du public et des acteurs du domaine santé-environnement à une information objective et transparente. En l'état actuel du projet, les données sanitaires, d'une part, et environnementales, d'autre part, pourraient être déconnectées et non interopérables entre elles. Le *Green data hub* vise donc à faciliter les croisements de données environnementales et sanitaires dans un but de progression des connaissances, et de la recherche, notamment en matière de causalité entre environnement et santé.

Le GD4H n'est donc pas une infrastructure mais une démarche visant à assurer un partage efficace et effectif de données environnementales au service de la recherche-environnement. Au regard de l'importance de l'open data dans le domaine de l'environnement, le GD4H pourra se traduire différemment sur le plan technologique en fonction des données considérées (portail fédérateur pour les données *open data* et infrastructure ou procédure d'accès sécurisé pour les données d'accès restreint). Il est anticipé que l'identification des bases de données, la vérification de leur utilisabilité dans le cadre d'études santé-environnement (ex : granularité, pertinence de la profondeur temporelle, etc.) ou leur montée en qualité constitueront des enjeux majeurs pour la réussite du GD4H.

Une phase d'investigation et d'identification des besoins de l'écosystème débute en 2021 à l'issue de laquelle des recommandations sur l'architecture du dispositif seront proposées.

Le data hub *du centre* Energy for Climate *(E4C)*

Créé par les établissements supérieurs d'enseignement et de recherche de l'Institut Polytechnique de Paris et l'Ecole des Ponts, le CNRS, le CEA et deux industriels (EDF et Total), le centre Energy for Climate fédère 29 laboratoires de recherche dans des domaines variés (météorologie, climat, sciences de l'ingénieur et des matériaux, mathématiques appliquées, informatique, économie, management, politiques publiques, etc.) dans un but de recherche et d'expérimentation des solutions opérationnelles en réponse aux enjeux d'atténuation du changement climatique.

Le hub de données mis en place à cette fin est une plateforme permettant le partage de données entre les différents partenaires, publics et/ou privés. Certains jeux de données sont accessibles en *open data* (y compris des jeux de données « apportés » par des partenaires industriels), d'autres sont partagés entre les partenaires (sans être ouverts au grand public). Les données les plus sensibles sont ouvertes en accès sécurisé via le CASD (cf. partie 4).

Le financement est majoritairement public mais peut être privé *via* du mécénat ou des contrats de partenariat.

La proposition d'un « Education Data Hub*»*

Ce projet fait l'objet de la proposition n°37 formulée par les États généraux du numérique pour l'éducation réunis en novembre 2020. Cette plateforme de données d'éducation aurait pour objectif de faire avancer la recherche en éducation, d'éclairer les décisions publiques et de construire des services plus performants.

2.2. Des plateformes de données pilotées par les acteurs de filière

La mission a recensé quatre exemples dans les domaines de l'agriculture, du tourisme, de l'agro-alimentaire et de la culture.

Ces quatre plateformes (ou projet) ont pour point commun :

- de permettre aux producteurs de la donnée (appelés « fournisseurs », « apporteurs » selon les hubs) d'avoir la maîtrise de leurs données dans un contexte où la puissance des plateformes numériques ou de certains acteurs économiques (par exemple dans le domaine agricole, les machinistes et semenciers) fait craindre à ces acteurs d'en être dépossédés ;
- d'être avant tout des outils de partage des données à destination d' « utilisateurs » choisis par les « fournisseurs » de donnée. La dimension *open data*, même si elle peut être présente, n'est pas systématique et surtout pas la raison d'être de ces plateformes ;
- d'engager une démarche de standardisation des données au sein de ces secteurs ;
- de proposer des « briques de service » en amont (à destination des fournisseurs de données, par exemple la structuration, mise en qualité et nettoyage de la donnée) ou aval (pour les utilisateurs, par exemple l'analyse de données quantitatives ou qualitatives) ;
- d'être financées au moyen d'abonnements acquittés par les fournisseurs de données et les utilisateurs de données, voire de commissions ou de rétributions pour les prestations de service réalisées. ;
- d'être un puissant levier de développement d'un écosystème aval de valorisation de la donnée (services, applications, etc.).

Le AgDataHub

Pionnier en la matière de plateforme de filière en France, ce hub regroupe, outre le réseau des instituts techniques agricoles (ACTA), les Chambres d'Agriculture (APCA) et le GEVES, une quinzaine de structures privées qui ont investi au capital de la société API-AGRO SAS, maison mère de la SAS *AgDataHub*.

La mission d'AgDataHub est de constituer la base des données agricoles françaises rassemblant potentiellement les données des 280 000 exploitations agricoles professionnelles françaises et 85 000 entreprises, et de permettre à ces acteurs de la filière agricole, apporteurs de données, d'en avoir la maîtrise, en choisissant explicitement avec quels partenaires ils souhaitent les échanger et pour quels usages. La plateforme permet de façon plus large d'accompagner les acteurs des différentes filières afin qu'ils partagent leurs données, pour améliorer les pratiques actuelles ou créer de nouveaux services innovants en matière d'agroécologie, d'agriculture de précision ou de performance économique. Enfin, cette gestion des données agricoles, ambitionne de fournir une meilleure information aux consommateurs finaux. Agdatahub vise ainsi à créer un cercle vertueux de circulation de l'information et de valorisation des données agricoles.

Dans le cadre de l'appel à projets « Accompagnement et transformation des filières » (ATF) pourant sur la mutualisation de moyens au service des filières et plateformes numériques de filières, le AgDataHub a bénéficié d'un contrat d'aide de 3,2 M€ de la part de Bpifrance pour financer le développement technologique de la plateforme de donnée et le dispositif de consentement à l'échange de données. Les travaux financés dans ce cadre ont vocation à répondre aussi à terme aux besoins exprimés par d'autres hubs sectoriels qui auront recours aux solutions développées par Dawex et Orange Business Service (consentement).

La plateforme Apidae Tourisme

La plateforme Apidae Tourisme a pour origine la création, en 2004, dans la région Rhône Alpes, d'un réseau d'acteurs du secteur du tourisme dont l'objectif était de mutualiser des moyens pour développer une base de données des informations touristiques. Progressivement, sa mission s'est élargie avec l'ambition de favoriser la mise en relation des acteurs de l'écosystème local qui utilisent la donnée touristique au service de l'économie locale.

En 2020, la plateforme Apidae héberge plus de 363 712 fiches objets remises à jour quotidiennement par plus de 1 336 clients. Le réseau compte 23 départements, 1 collectivité d'outre-mer et plus de 24 793 utilisateurs de la plateforme

Devant l'ampleur prise par la démarche et face à la multiplication des usages et des services apportés par l'ensemble des acteurs de l'écosystème, le réseau s'est doté en février 2020 d'une personnalité juridique propre, sous la forme d'une société coopérative dont le capital est apporté par 184 sociétaires (exemples) dont Bpifrance à hauteur 100 000 €.

La plateforme Numalim

Issue des travaux des États généraux de l'alimentation en 2017 et de la signature du contrat stratégique de la filière alimentaire en 2018, la plateforme numérique de l'alimentation Numalim, créée en février 2020, a pour objectif de répondre aux enjeux de transparence et de valorisation des démarches de qualité dans l'ensemble de la filière agro-alimentaire.

Elle repose sur une base de données renseignée directement par les fabricants qui a vocation à rassembler non seulement la composition des produits, mais aussi l'origine, les conditions de production (pratiques agricoles), des informations nutritionnelles, des données sur l'impact carbone et le recyclage, etc.

À ce jour, Numalim compte parmi ses sociétaires, 13 organisations représentant notamment les industries agro-alimentaires et les consommateurs, et 8 entreprises. A horizon 2025, elle ambitionne de fédérer les données de la moitié des 18 000 entreprises françaises de l'agroalimentaire, avec un accès pour les 3 500 PME et 12 000 TPE que compte le secteur.

Effet collatéral bénéfique, NumAlim vise d'offrir à la filière alimentaire française un avantage concurrentiel à l'export où l'information sur les données alimentaires est beaucoup moins développée.

Bpifrance a signé un contrat d'aide de 3 M€ sur quatre ans.

Le projet d'Alliance Culture Data

Porté par BnF participation, filiale de la Bibliothèque Nationale de France, ce projet s'adresse aux acteurs des industrie culturelles et créatives (ICC) qui comptent notamment l'édition, le cinéma, la musique, la radio et la télévision.

Son ambition est de permettre la traçabilité des œuvres à des fins de gestion de droits d'auteur et de respect de la propriété intellectuelle, de faire face à l'intermédiation progressive opérée par les plateformes numériques (de type Netflix), d'optimiser le référencement des produits culturels pour en accentuer la « découvrabilité », d'améliorer l'offre de produits culturels, en adéquation avec les goûts des usagers, d'affiner la connaissance des usages pour optimiser l'expérience utilisateur et développer de nouveaux revenus, de permettre l'émergence de start-up notamment en favorisant l'open data.

2.3. Des connexions à créer entre « hubs » sectoriels

Ces hubs de données sont par nature des vecteurs de décloisonnement, puisque leur raison d'être est de permettre d'accéder à des données qui sont aujourd'hui disséminées dans différentes structures, parfois jalousement « couvées » par leur « producteur ». Leur succès dépendra de leur capacité à mobiliser les acteurs de leur secteur d'activité respectif. Ils constituent un atout essentiel pour tirer parti des potentialités offertes par les appariements de données.

Ces différents exemples montrent la variété des initiatives en cours dont la mission considère qu'elle révèle aussi la variété des contextes ayant présidé à la création de ces dispositifs, des acteurs qu'ils fédèrent, et *in fine* des besoins. Cette pluralité illustre le pragmatisme qui doit guider les choix technologiques et organisationnels en la matière.

S'agissant par exemple des infrastructures de stockage, le choix entre infrastructures dédiées ou infrastructures mutualisées, entre stockage centralisé ou décentralisé dépendra notamment de la situation de l'existant (nombre des bases, localisation, par exemple), des exigences en matière de sécurité, et des moyens mobilisables (en interne ou sous forme de prestation).

La centralisation du stockage des données en un seul lieu peut être jugée préférable pour des données sensibles dont un acteur se porte garant de la protection, mais elle peut également accroître le risque pour la sécurité du système d'information (*single point of failure*), là où une solution décentralisée permet de répartir ce risque. Pour des besoins relatifs à des données moins sensibles, comme le comptage de patients dans un établissement de santé, l'utilisation d'un outil mutualisé peut être préféré pour éviter de déplacer les données.

De même, les solutions d'IA en cours de développement comme le *federated learning* sont conçues pour un environnement décentralisé mais nécessitent d'avoir des systèmes d'information compatibles et une maintenance coordonnée dans les différents endroits où l'algorithme est opéré.

Pour ces raisons, il n'apparaît pas possible et pertinent à la mission de définir un cahier des charges type du *hub* de données sectoriels qui serait transposable en l'état dans les différents secteurs de l'action publique. Il convient en revanche de s'assurer que cette pluralité des plateformes de données ne nuise pas à l'interopérabilité des dispositifs de partage et ne constituent pas des freins à l'accès aux données. Derrière l'interopérabilité des plateformes, l'enjeu se situe au niveau de l'interopérabilité des données, elle-même conditionnée par l'adoption d'un minimum de normalisation (cf. paragraphe 1).

> **Recommandation** : Encourager la création de « hubs » sectoriels ou intersectoriels, selon des modalités adaptées à chaque secteur, et dans des conditions assurant leur interopérabilité.

Ces plateformes de données constituent un enjeu d'autonomie stratégique au niveau français mais aussi européen : s'il n'y a pas de moyens (techniques et juridiques) de partage, si l'Europe n'est pas en mesure d'utiliser ses propres jeux de données, elle sera complètement dépendante des données et donc des services et algorithmes développés par d'autres puissances.

3. Faciliter l'accès aux données pour les chercheurs

Pour obtenir un accès sécurisé aux données, les chercheurs adressent aujourd'hui des demandes aux administrations, sans qu'il n'y ait de procédure claire et encadrée. Les chercheurs peuvent ainsi s'adresser directement aux services producteurs, au sein desquels il n'existe pas nécessairement d'interlocuteur ou de référent identifié pour ces demandes (dans certains cas, il s'agit de la personne responsable de l'accès aux documents administratifs, PRADA), mais peuvent aussi solliciter les services statistiques ministériels (SSM), qui jouent souvent un rôle d'interface avec la recherche, en raison de leur compétence de mise à disposition des données statistiques et de leur proximité intellectuelle avec les chercheurs traitant des données. Les SSM connaissent ainsi bien les procédures et le cadre du secret statistique, notamment l'établissement des conventions de mise à disposition des données et le recueil de l'avis du comité du secret statistique, que les chercheurs doivent solliciter lorsque leurs travaux portent sur des données individuelles.

Aujourd'hui, le cadre pour la prise en charge des demandes d'accès sécurisé à des données émanant de chercheurs n'est pas suffisamment clair et structuré, à la fois pour les chercheurs mais aussi pour les administrations. Le cadre prévu par le régime des documents administratifs et notamment la saisine de la CADA ne peut s'appliquer, étant donné que les données sollicitées ne tombent pas dans ce périmètre. Ainsi, surtout lorsqu'il s'agit d'accéder à une source de données n'ayant jamais été ouverte, il peut se produire qu'un chercheur soit :

- n'obtienne pas de réponse de l'administration ;
- soit mal orienté ;
- n'ait aucun recours en cas de refus de communication des données.

En général, ces obstacles peuvent se cumuler. C'est le cas d'un chercheur dont la mission a retracé le parcours pour obtenir la communication de données relatives au travail détaché, dans le cadre d'un projet d'évaluation des effets économiques du recours au travail détaché en France (cf. encadré).

Évaluer le travail détaché : récit d'un parcours du combattant-chercheur

Une chercheuse française, économiste et doctorante dans une prestigieuse institution de recherche parisienne, a cherché à évaluer le recours au travail détaché en France. Plus de vingt mois après sa première requête, sa demande d'accès aux données n'a toujours pas abouti.

En mai 2018, elle adresse une première demande à la direction générale du travail (DGT), qui lui indique pouvoir mettre à disposition les données brutes sur le détachement. Restée plusieurs mois sans réponse à la suite de ce premier rendez-vous, la chercheuse recontacte la DGT en **novembre 2018**, et est informée que sa demande est en cours de traitement. Entre **janvier 2019 et mai 2019**, la chercheuse contacte à plusieurs reprises la DGT, sans succès. **Le 5 juin 2019**, la chercheuse renouvelle sa demande auprès de la DGT en s'adressant cette fois-ci au directeur général.

Le 19 juillet 2019, la DGT la renvoie vers le service statistique ministériel (SSM), la direction de l'animation de la recherche, des études et des statistiques (DARES), qui dispose des données individuelles sur le détachement à partir de 2017. Le lendemain, le mail de la chercheuse est transmis à la DARES, et le **23 juillet 2019**, la DGT met à disposition de la chercheuse un premier jeu simplifié et incomplet de données relatives au détachement, plus d'un an après sa première requête.

Le 19 août 2019, la DARES assure à la chercheuse que les données brutes exhaustives seront mises à sa disposition après un retraitement statistique, et lui indique qu'elle doit effectuer une demande formelle auprès du comité du secret statistique. D'**août 2019 à mars 2020**, la chercheuse échange régulièrement avec le service statistique de la DARES, qui coopère pour l'aider à établir sa convention de mise à disposition. **Le 24 avril 2020**, la chercheuse demande un accès à des données supplémentaires pour compléter son évaluation (données individuelles conservées par le réseau de l'inspection du travail). De **juin à juillet 2020**, la chercheuse est en attente de réponse du service juridique de la DARES, qui est en charge de valider la convention concernant l'accès aux données, et qu'elle relance à plusieurs reprises.

Le **20 août 2020**, après deux demandes auprès du service juridique, la chercheuse reçoit l'accord de la DARES pour l'instruction de sa demande auprès du comité du secret. Le **17 septembre 2020**, la chercheuse reçoit l'accord du comité du secret pour obtenir les données individuelles sur le détachement. Sans nouvelles pendant deux mois en dépit de l'achèvement de ces formalités légales, la chercheuse contacte à nouveau le service statistique de la DARES le **2 novembre 2020**. Elle est finalement informée le **13 novembre 2020** que l'accès aux données doit être reporté, et que la mise à disposition ne pourra se faire avant quatre mois a minima, **en mars 2021**. La DARES invoque la nécessité d'un temps de retraitement supplémentaire, ainsi qu'un revirement sur la mise à disposition des données qui s'opérerait désormais via le CASD. À date, la chercheuse ne dispose donc toujours pas des données sur le détachement, plus de deux ans et demi après les premiers contacts, un retard de mise à disposition des données qui empêche non seulement une évaluation quantitative rigoureuse de la politique des travailleurs détachés, mais entraîne également des surcoûts importants pour la chercheuse et son laboratoire, en matière de financements de thèse comme d'accès payant au CASD depuis septembre 2020.

Cet exemple montre le besoin de **renforcer l'action des SSM et leur interaction avec les directions métiers** pour la mise à disposition des données auprès des chercheurs, mais aussi celui de **mieux intégrer le travail des chercheurs dans l'évaluation des politiques publiques, la sensibilité d'un sujet, par ailleurs enjeu public majeur, ne pouvant justifier une absence de communication sur plusieurs années.**

Un délai aussi conséquent entre la demande initiale et la décision du comité du secret pourrait être évitée avec un cadre simple de prise en charge des demandes des chercheurs au sein de l'administration. À cet égard, la Commission européenne offre un exemple de bonnes pratiques :

- des adresses mail fonctionnelles identifiables et disponibles sur les sites de la Commission existent pour les chercheurs ;
- quand un chercheur sollicite un service de la Commission pour obtenir des données, il reçoit automatiquement un message d'accusé réception ;
- la Commission a une obligation de répondre à la demande sous quinze jours ;
- l'accusé réception automatique envoyé par la boîte fonctionnelle précise l'existence de ce délai et la date à laquelle il arrive à échéance[174].

Il s'agit donc tout à la fois d'un régime de réponse aux demandes d'accès aux documents plus large et contraignant que le droit français (prévu par un règlement de 2001[175]) et de bonnes pratiques « de bon sens » (adresses fonctionnelles, communication sur Internet sur la procédure, message automatique précisant les modalités pour le chercheur). Il semble tout à fait accessible aux pratiques françaises de se mettre à ce niveau, et au droit français d'envisager une obligation de réponse à ces demandes spécifiques, afin d'en garantir la traçabilité et la prise en charge.

La mission considère qu'il n'est pas nécessairement souhaitable d'avoir une porte d'entrée unique pour les demandes des chercheurs, compte tenu de la variété des situations. Toutefois, elle identifie nettement le besoin de créer un cadre minimal de prise en charge de leurs demandes. Ce rôle peut être confié notamment aux AMDAC et aux SSM, qui, par leurs fonctions et leur vision souvent centralisée des données existantes, sont les plus à même de pouvoir y répondre. La configuration exacte de la prise en charge doit être décidée à l'échelle la plus pertinente, notamment des ministères et des opérateurs dans le cas de l'État. Elle doit cependant garantir les outils indispensables à la traçabilité et à la réponse minimale qui doit être apportée aux demandes, comme les boîtes mail fonctionnelles, les accusés réception et le rappel du cadre juridique applicable. Un suivi de ces demandes doit être effectué, et un recours doit être possible en cas de réponse trop tardive.

[174] Exemple de message adressé automatiquement par la Commission : « Nous vous remercions pour votre message électronique du 07/10/2019. Par la présente, nous accusons réception de votre demande d'accès à des documents, qui a été enregistrée le 08/10/2019 sous le numéro de référence GESTDEM 2019/XXXX. Conformément au règlement (CE) n° 1049/2001 relatif à l'accès du public aux documents du Parlement européen, du Conseil et de la Commission, votre demande sera traitée dans un délai de quinze jours ouvrables. Ce délai arrivera à expiration le 29/10/2019. S'il s'avère nécessaire de le prolonger, nous vous en aviserons en temps utile. »

[175] Article 7 du règlement n° 1049/2001 du Parlement européen et du Conseil du 30 mai 2001 relatif à l'accès du public aux documents du Parlement européen, du Conseil et de la Commission : « *Les demandes d'accès aux documents sont traitées avec promptitude. Un accusé de réception est envoyé au demandeur. Dans un délai de quinze jours ouvrables à partir de l'enregistrement de la demande, l'institution octroie l'accès au document demandé et le fournit dans le même délai conformément à l'article 10, soit communique au demandeur, dans une réponse écrite, les motifs de son refus total ou partiel et l'informe de son droit de présenter une demande confirmative conformément au paragraphe 2 du présent article.* »

En complément, afin d'accroître la confiance des services producteurs de la donnée vis-à-vis des chercheurs, la mission recommande de sensibiliser les administrations à la faculté qui leur est ouverte depuis l'adoption de l'article 36 de la loi pour une République numérique[176], de recourir de manière volontaire, au comité du secret statistique. « *Ce passage par le comité du secret statistique [permet] de sécuriser les administrations afin de les inciter à autoriser l'accès aux bases de données qu'elles détiennent dans des conditions présentant toutes les garanties requises. Cet accès sécurisé [peut] être assuré par le producteur ou être fourni via des services tels que le CASD (centre d'accès sécurisé aux données) ou l'ODR (open data room) géré par la Banque de France* »[177].

Enfin, l'outil que constitue le centre d'accès sécurisé aux données (CASD) gagnerait à être mieux connu au sein de l'administration. Créé par arrêté interministériel du 29 décembre 2018, le CASD est un groupement d'intérêt public (GIP) rassemblant l'État représenté par l'INSEE, le GENES, le CNRS, l'École Polytechnique et HEC Paris. Il a pour *« objet principal d'organiser et de mettre en œuvre des services d'accès sécurisé pour les données confidentielles à des fins non lucratives de recherche, d'étude, d'évaluation ou d'innovation, activités qualifiées de « services à la recherche», principalement publiques »*. Contrairement aux hubs sectoriels décrits plus haut, il n'a pas vocation à animer un écosystème d'acteurs dans un secteur donné, et se réduit ainsi davantage à une fonction d'infrastructure. Néanmoins, le CASD est susceptible d'apporter une réponse aux demandes d'accès à des données de manière sécurisée – en particulier pour les chercheurs. Il dispose de près de 350 jeux de données de différents secteurs. **L'intervention du CASD et des hubs sectoriels pour les projets de recherche doit donc être promue.**

> **Recommandation** : Améliorer la prise en charge des demandes des chercheurs, en associant les AMDAC et les SSM (délai de réponse obligatoire, création d'un recours, recours à la consultation du comité du secret statistique à titre facultatif)

[176] Cet article modifie l'article L. 311-8 du code des relations entre le public et l'administration (CRPA) et prévoit que lorsqu'une demande d'accès anticipé « porte sur une base de données et vise à effectuer des traitements à des fins de recherche ou d'étude présentant un caractère d'intérêt public, l'administration détenant la base de données ou l'administration des archives peut demander l'avis du comité du secret statistique institué par l'article 6 bis de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques ». Le comité pourra recommander le recours à « une procédure d'accès sécurisé aux données présentant les garanties appropriées », son avis devant tenir compte i) « des enjeux attachés aux secrets protégés par la loi, notamment la protection de la vie privée et la protection du secret industriel et commercial » ainsi que ii) « de la nature et de la finalité des travaux pour l'exécution desquels la demande d'accès est formulée ».

[177] *«* L'accès des chercheurs aux données administratives. État des lieux et propositions d'actions », Rapport au secrétaire d'État charge de l'industrie, du numérique et de l'innovation, Conseil national de l'information statistique, A. Bozio, P.-Y. Geoffard, mars 2017.

4. Adapter le cadre juridique national et européen en conciliant innovation et protection des droits fondamentaux

4.1. En matière de données personnelles : adopter un cadre protecteur et adapté aux besoins d'expérimentation en matière d'IA

Les technologies d'IA comportent deux phases bien distinctes qui appellent des régimes juridiques différenciés en matière de protection des données personnelles

La phase de développement des outils d'intelligence artificielle dans un nombre croissant de domaines relevant des autorités publiques (santé, sécurité, environnement, etc.) suppose de disposer de jeux de données, souvent personnelles, destinées à *« entraîner les machines »*. Afin de produire des algorithmes performants, ces données doivent être réelles, représentatives des situations opérationnelles d'emplois car des données reconstituées peuvent entraîner des biais dans l'algorithme affectant la qualité du développement et *in fine* la confiance mise dans l'outil. Elles doivent en outre être en nombre conséquent, d'où la nécessité de recourir à des données déjà collectées, alors même que la finalité de développement d'un outil d'IA n'a pas systématiquement été prévue lors de la collecte de ces données. Le caractère massif mais aussi rétrospectif de la collecte des dites données rend de fait impossible le recueil du consentement des individus à cette réutilisation. On peut citer par exemple les millions de radiographies pulmonaires nécessaires pour mettre au point un algorithme d'aide à la détection des tumeurs pulmonaires, ou aux milliers d'heures de vidéo nécessaires pour élaborer un algorithme détectant la présence d'un colis abandonné. En outre, les données d'apprentissage étant couteuses à produire, annoter et stocker, et les algorithmes d'IA devant être régulièrement réévalués et testés, il est capital de pouvoir conserver ces données pendant une durée supérieure aux durées de conservation de droit commun.

Lorsque l'apprentissage porte sur des données personnelles, la possibilité de réutiliser des données déjà collectées à une autre fin devrait être offerte tout en assurant un cadre protecteur des droits fondamentaux, d'autant plus qu'il est facile de garantir que les données personnelles utilisées dans la phase d'entrainement le sont « à blanc », sans aucune incidence sur l'individu propriétaire de ces données.

Passée la phase d'entrainement, **l'algorithme utilisé en phase opérationnelle** est ensuite « alimenté » par des données nouvelles, indépendantes des données d'apprentissage et relevant du régime de protection de droit commun s'appliquant au traitement de données alors créé (principes de finalité et de proportionnalité, durées strictes de conservation, contrôle indépendant, etc.).

Dans le cadre juridique actuel, il n'existe pas de différence entre les formalités applicables au traitement créé pour la phase d'expérimentation et celles requises pour la phase opérationnelle[178]. Ainsi, la phase d'élaboration d'un algorithme dit de vidéo intelligente (comme dans le cas du projet VOIE) suppose l'adoption d'un décret en Conseil d'État après avis motivé et publié de la CNIL, ce qui prend en moyenne 18 mois, alors même qu'il s'agit d'une expérimentation menée sur un temps (quelques semaines) et un espace (une demi-douzaine d'emprises ferroviaires ou RATP) limités, comme l'exige la mise en place pérenne d'un même dispositif entré en phase opérationnelle.

[178] La loi de 1978 modifiée a supprimé les formalités préalables sauf pour les traitements « *mis en œuvre pour le compte de l'Etat et : 1°) qui intéressent la sûreté de l'Etat, la défense ou la sécurité publique ; 2°) ou qui ont pour objet la prévention, la recherche, la constatation ou la poursuite des infractions pénales ou l'exécution des condamnations*

L'absence actuelle d'un cadre juridique sur la protection des données personnelles adapté à l'apprentissage en IA conduit aujourd'hui à ce que, selon les secteurs et les projets, les acteurs renoncent à tirer profit des avancées en matière d'IA ou en limitent les développements, recourent à des jeux d'apprentissage reproduisant artificiellement les données nécessaires, ou acquièrent des jeux de données pré-entrainées par d'autres, au risque, non seulement de compromettre la survie des acteurs français et européens de l'IA[179], mais aussi de rendre moins performants les outils développés.

C'est particulièrement le cas lorsque l'expérimentation se situe dans une des matières couvertes par les articles 31[180] et 32[181] de la loi informatique et libertés. Dans ces cas, toute expérimentation requiert à la fois la modification du texte réglementaire régissant le traitement ayant recueilli les données dont la réutilisation est souhaitée (arrêté ou décret en CE après avis motivé et publié de la CNIL) et l'adoption d'un texte réglementaire créant le nouveau traitement dont l'expérimentation est la finalité. Si l'expérimentation débouche sur une mise en œuvre opérationnelle de l'algorithme d'IA, un troisième texte est alors nécessaire selon le même formalisme.

Une telle lourdeur n'est pas compatible avec le temps de l'innovation technologique et impose de trouver une manière agile d'encadrer juridiquement ce type d'expérimentation.

Les marges de manœuvre offertes par le RGPD

Le règlement général sur la protection des données offre aux Etats une assez large faculté d'adapter les principes qu'il développe, et notamment en matière de recherche.

D'une part, l'article 5.1.b) du RGPD[182] prévoit, par dérogation au principe selon lequel les données personnelles sont collectées pour « *des finalités déterminées, explicites et légitimes et ne peuvent être traitées ultérieurement d'une manière incompatibles avec ces finalités* », trois motifs de réutilisation des données personnelles (les archives, la statistique et la recherche) détaillés dans l'article 89 du même règlement.

---

pénales ou des mesures de sûreté » (article 31), « *qui portent sur des données génétiques ou sur des données biométriques nécessaires à l'authentification ou au contrôle de l'identité des personnes* » (article 32), ainsi que pour certains traitements dans le domaine de la santé.

[179] Cet argument « écosystémique » a d'ailleurs été souligné avec force par le rapport « Donner un sens à l'intelligence artificielle » de la mission présidée par le député Cédric Villani (avril 2018). *« L'accès à la donnée reste néanmoins une condition essentielle de l'émergence d'une industrie française et européenne de l'IA. Dans un monde de plus en plus automatisé, c'est de cet accès que dépendent la vitalité et la performance de notre recherche et de l'action publique, mais aussi notre capacité collective à déterminer la trajectoire de l'intelligence artificielle, à dessiner les contours de notre société automatisée.*

Or, sur l'IA, la situation actuelle est caractérisée par une asymétrie critique entre les acteurs de premier plan – les GAFAM, auxquels il faut ajouter IBM pour l'IA, d'un côté, les BATX, de l'autre – qui ont fait de la collecte et de la valorisation des données la raison de leur prééminence ; et les autres entreprises et administrations, dont la survie à terme est menacée. Cette première asymétrie en emporte une seconde, critique, entre l'Europe et les Etats-Unis. »

[180] « Traitements de données à caractère personnel mis en œuvre pour le compte de l'Etat et : 1° Qui intéressent la sûreté de l'Etat, la défense ou la sécurité publique ; 2° Ou qui ont pour objet la prévention, la recherche, la constatation ou la poursuite des infractions pénales ou l'exécution des condamnations pénales ou des mesures de sûreté ».

[181] « Les traitements de données à caractère personnel mis en œuvre pour le compte de l'Etat, agissant dans l'exercice de ses prérogatives de puissance publique, qui portent sur des données génétiques ou sur des données biométriques nécessaires à l'authentification ou au contrôle de l'identité des personnes ».

[182] « Les données à caractère personnel doivent être : (…) b) collectées pour des finalités déterminées, explicites et légitimes, et ne pas être traitées ultérieurement d'une manière incompatible avec ces finalités; le traitement ultérieur à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques n'est pas considéré, conformément à l'article 89, paragraphe 1, comme incompatible avec les finalités initiales (limitation des finalités) »

D'autre part, l'alinéa 4 de l'article 6[183] du RGPD indique, « en creux » que « le droit de l'Union ou le droit d'un Etat membre qui constitue une mesure nécessaire et proportionnée dans une société démocratique pour garantir les objectifs visés à l'article 23 paragraphe 1 » peuvent autoriser la réutilisation de données personnelles à d'autres fins que celles initialement déclarées.

Enfin, le considérant 159 du RGPD[184] adopte une vision large de la notion de recherche. Il précise ainsi que « le traitement de données à caractère personnel à des fins de recherche scientifique devrait être interprété au sens large et couvrir, par exemple, le développement et la démonstration de technologies, la recherche fondamentale, la recherche appliquée et la recherche financée par le secteur privé. Il devrait, en outre, tenir compte de l'objectif de l'Union mentionné à l'article 179, paragraphe 1, du traité sur le fonctionnement de l'Union européenne, consistant à réaliser un espace européen de la recherche. Par «fins de recherche scientifique», il convient également d'entendre les études menées dans l'intérêt public dans le domaine de la santé publique ».

La lecture combinée de ces trois dispositions permet d'estimer que le droit européen ne s'oppose pas à l'adoption de dispositions nationales qui autoriseraient la réutilisation des données personnelles à des fins de recherche appliquée en matière d'innovation technologique, pour autant que le traitement est mis en œuvre dans le respect des dispositions du RGPD et de la LIL applicables à de tels traitements et qu'il n'est pas utilisé pour prendre des décisions à l'égard des personnes concernées. S'agissant du droit français, cette adoption suppose de modifier la loi 78-17 du 6 janvier 1978.

---

[183] « *Lorsque le traitement à une fin autre que celle pour laquelle les données ont été collectées n'est pas fondé sur le consentement de la personne concernée ou sur le droit de l'Union ou le droit d'un État membre qui constitue une mesure nécessaire et proportionnée dans une société démocratique pour garantir les objectifs visés à l'article 23, paragraphe 1, le responsable du traitement, afin de déterminer si le traitement à une autre fin est compatible avec la finalité pour laquelle les données à caractère personnel ont été initialement collectées, tient compte, entre autres :*

a) de l'existence éventuelle d'un lien entre les finalités pour lesquelles les données à caractère personnel ont été collectées et les finalités du traitement ultérieur envisagé ;

b) du contexte dans lequel les données à caractère personnel ont été collectées, en particulier en ce qui concerne la relation entre les personnes concernées et le responsable du traitement ;

c) de la nature des données à caractère personnel, en particulier si le traitement porte sur des catégories particulières de données à caractère personnel, en vertu de l'article 9, ou si des données à caractère personnel relatives à des condamnations pénales et à des infractions sont traitées, en vertu de l'article 10 ;

d) des conséquences possibles du traitement ultérieur envisagé pour les personnes concernées ;

e) de l'existence de garanties appropriées, qui peuvent comprendre le chiffrement ou la pseudonymisation. »

[184] « *Lorsque des données à caractère personnel sont traitées à des fins de recherche scientifique, le présent règlement devrait également s'appliquer à ce traitement. Aux fins du présent règlement, le traitement de données à caractère personnel à des fins de recherche scientifique devrait être interprété au sens large et couvrir, par exemple, le développement et la démonstration de technologies, la recherche fondamentale, la recherche appliquée et la recherche financée par le secteur privé. Il devrait, en outre, tenir compte de l'objectif de l'Union mentionné à l'article 179, paragraphe 1, du traité sur le fonctionnement de l'Union européenne, consistant à réaliser un espace européen de la recherche. Par «fins de recherche scientifique», il convient également d'entendre les études menées dans l'intérêt public dans le domaine de la santé publique. Pour répondre aux spécificités du traitement de données à caractère personnel à des fins de recherche scientifique, des conditions particulières devraient s'appliquer, en particulier, en ce qui concerne la publication ou la divulgation d'une autre manière de données à caractère personnel dans le cadre de finalités de la recherche scientifique. Si le résultat de la recherche scientifique, en particulier dans le domaine de la santé, justifie de nouvelles mesures dans l'intérêt de la personne concernée, les règles générales du présent règlement s'appliquent à l'égard de ces mesures.* »

Le principe de bacs à sable ou « sandbox *» expérimenté par l'homologue de la CNIL au Royaume-Uni*

Depuis un peu plus d'un an, l'*Information Commissioner's Office* (ICO),homologue de la CNIL au Royaume-Uni, développe un service dénommé « *Regulatory Sandbox* »[185] destiné à soutenir les organisations (de la start-up aux grandes organisations, privées, publiques ou associatives) qui créent des produits et services utilisant les données personnelles de manière innovante. Elle accompagne chaque projet sélectionné grâce à une équipe dédiée.

Selon le commissaire à l'information, Elizabeth Denham, « *l'ICO soutient l'innovation technologique et les nouvelles utilisations passionnantes des données, tout en veillant à ce que la vie privée et les droits des personnes soient protégés. (…) La confidentialité et l'innovation ne s'excluent pas mutuellement et il n'y a pas besoin de choisir entre les deux. Le bac à sable aidera les entreprises et les organismes publics à proposer de nouveaux produits et services au profit du public, avec l'assurance qu'ils ont abordé la protection des données en l'intégrant dès le départ. S'engager avec les entreprises et les innovateurs dans le bac à sable est également un exercice précieux pour l'ICO qui peut ainsi identifier les avancées en matière de technologie et d'innovation et les opportunités et les défis qu'ils peuvent offrir.*»

Lancé en septembre 2019, la phase Beta du s*andbox* a permis à un échantillon d'une dizaine organisations d'essayer le service parmi lesquelles :

- ONFIDO pour étudier comment identifier et atténuer les biais algorithmiques dans les modèles d'apprentissage automatique utilisés pour la vérification d'identité biométrique à distance ;
- le JISC[186] pour accompagner le développement de son code de bonnes pratiques développé avec les établissements d'enseignement qui souhaitent utiliser les données sur les activités des étudiants pour soutenir le bien-être mental, émotionnel et physique des étudiants ;
- Novartis pour expérimenter la technologie vocale dans les soins°,
- l'aéroport d'Heathrow pour l'utilisation de la biométrie en vue de fluidifier le parcours des usagers lors des différents contrôles réalisés ;
- l'autorité du Grand Londres, sur le croisement de fichiers de données sanitaires, sociales et criminelles pour prévenir et réduire les violences.

Un deuxième appel à manifestation d'intérêt a été ouvert en août 2020 dans les domaines suivants : le partage de données et la protection de la vie privée des mineurs. Les projets *« doivent être à la pointe de l'innovation et peuvent opérer dans des domaines particulièrement difficile de la protection des données, où il existe une réelle incertitude sur ce à quoi ressemble la conformité »*[187].

[185] Bac à sable.

[186] Association à but non lucratif des secteurs de l'enseignement secondaire et supérieur et de la formation au Royaume-Uni pour les services et les solutions numériques.

[187] Site internet de l'ICO.

Une première introduction du concept de bac à sable réglementaire en droit français dans les domaines des télécommunications et de l'énergie

La loi pour une République numérique (dans son article 92) a introduit en droit français **la faculté de déroger au cadre d'attribution des fréquences et numéros** à des fins d'expérimentation[188]. Elle autorise l'Autorité de régulation des communications électroniques et des Postes (ARCEP) à accorder des dérogations temporaires aux obligations d'un acteur afin de l'accompagner dans le développement d'une technologie et d'un service innovants, au plan technique ou commercial (jusqu'à deux ans). Ce dispositif doit permettre d'accompagner l'innovation dans l'internet des objets, les applications mobiles utilisant des numéros de téléphone, les réseaux associatifs, etc. L'adoption de ce dispositif fait suite aux constats selon lesquels les entrepreneurs ou les entreprises innovantes de petite taille ne sont pas nécessairement rompus aux démarches administratives et obligations liées à leur statut, et le déploiement à destination du grand public d'innovations potentiellement disruptives peut se trouver limité par un cadre réglementaire peu flexible.

Toute dérogation aux obligations doit faire l'objet d'une décision au cas par cas par l'ARCEP, et ne saurait remettre en cause les objectifs de régulation de l'Autorité. En particulier, l'ARCEP est vigilante à ce que la levée des obligations ne remette pas en cause la protection du secret des correspondances, de la santé ou encore de l'environnement. En outre, les obligations de service public prévues aux articles L. 35 à L. 35-7 du CPCE, comme l'acheminement gratuit des appels d'urgence par exemple, ne peuvent pas être concernées par ces dérogations.

Ces exonérations peuvent être valables y compris dans le cas où le demandeur souhaiterait tester un service commercial auprès du grand public. Néanmoins, ce cadre dérogatoire n'est envisageable que si l'expérimentation concerne un nombre limité d'utilisateurs et génère un chiffre d'affaires raisonnable : le chiffre d'affaires hors taxes réalisés chaque semestre doit être inférieur à 500 000 € et le nombre d'utilisateurs de la technologie ou du service doit être inférieur à 5 000.

Quatre ans après son entrée en vigueur, le dispositif de bac à sable ouvert auprès de l'ARCEP compte 9 cas de mise en œuvre pour des demandes déposées entre juin 2017 et septembre 2019, par cinq sociétés, tous dans le domaine de la numérotation[189]. Les ressources ouvertes à titre dérogatoire dans un but expérimental vont prochainement être ouvertes de manière pérenne, à la suite de la transposition d'une directive visant à introduire en droit français le code européen des télécommunications. Le bac à sable a ainsi permis d'anticiper la modification pérenne de la réglementation dans ce secteur.

Par ailleurs, l'article 61 de la loi Energie-Climat[190] a introduit **un dispositif d'expérimentation réglementaire dans le secteur de l'énergie**, créant ainsi un cadre juridique favorable aux expérimentations visant à déployer des technologies ou des services innovants en faveur de la transition énergétique et des réseaux et infrastructures intelligents, tout en garantissant la sécurité, la sûreté et la qualité de fonctionnement des réseaux et des installations. Le législateur a limité les champs de la réglementation ouverts à l'expérimentation.

[188] Introduction du IV de l'article L. 42-1 et IV de l'article 44 du code des postes et des communications électroniques (CPCE).

[189] Dans le domaine des fréquences, la faculté qui préexistait dans le code d'autoriser des expérimentations techniques de courte durée, sans fin commerciale, peut expliquer l'absence de demandes dans le cadre du bac à sable.

[190] Loi n°2019-1147 du 8 novembre 2019 relative à l'énergie et au climat.

La commission de régulation de l'énergie (CRE)[191] et l'autorité administrative « peuvent, chacune dans leur domaine de compétence, par décision motivée, accorder des dérogations aux conditions d'accès et à l'utilisation des réseaux et installations pour déployer à titre expérimental des technologies ou des services innovants en faveur de la transition énergétique et des réseaux et infrastructures intelligents ». Ces dérogations temporaires d'une durée maximale de quatre ans (renouvelables une fois au plus pour la même durée et dans les mêmes conditions que la dérogation initialement accordée) « sont assorties d'obligations relatives à l'information des utilisateurs finaux concernant le caractère expérimental de l'activité ou du service concerné ainsi qu'aux modalités de mise en conformité, à l'issue de l'expérimentation, avec les obligations auxquelles il a été dérogé ». Ces expérimentations doivent contribuer à l'atteinte des objectifs de la politique énergétique définis à l'article L. 100-1 du code de l'énergie. Le ministre chargé de l'énergie et, le cas échéant, le ministre chargé de la consommation disposent d'un délai de deux mois à compter de la notification de la demande de dérogation, pour s'opposer à l'octroi de tout ou partie de ces dérogations. La Commission de régulation de l'énergie publie chaque année un rapport sur l'avancement des expérimentations pour lesquelles une dérogation a été accordée et en publie une évaluation lorsqu'elles sont achevées.

Un premier guichet de candidature a été ouvert par la CRE du 15 juin au 15 septembre 2020 et lui a permis de recueillir une quarantaine de projets dont la moitié entre dans les domaines ouverts à l'expérimentation. La CRE prévoit de rendre ses décisions sur ces projets d'ici février 2021. S'il est trop tôt pour tirer un bilan, la commission estime, sans préjuger des décisions qu'elle prendra, qu'il permet de faire émerger des solutions très innovantes et dignes d'intérêt. Elle envisage de reconduire l'appel à projet à une fréquence annuelle ou biannuelle.

La nécessité de l'élargir ce dispositif dérogatoire de bac à sable au droit des données personnelles pour la constitution de jeux d'apprentissage en matière d'intelligence artificielle

La création d'un dispositif expérimental permettant à titre dérogatoire, la réutilisation de données déjà collectées pour une finalité autre que la constitution de jeux d'apprentissage d'IA, et leur conservation sur une durée plus longue, conditionne largement l'agilité et la capacité d'innovation dans notre pays.

Étant donné la grande diversité des applications, des secteurs et des usages de l'IA, ce cadre dérogatoire devrait privilégier la définition de règles génériques (critères d'éligibilité de l'expérimentation, principe d'une durée maximale d'expérimentation, d'une évaluation de l'expérimentation à l'issue du projet, publication d'un rapport annuel sur les expérimentations en cours, etc.) et renvoyer à un protocole, négocié au cas par cas, avec l'autorité chargée du contrôle de sa mise en œuvre (la CNIL).

Ce dispositif expérimental devrait prévoir une acception suffisamment large (tout en étant précise) des données nécessaires à l'apprentissage dans la mesure où l'intérêt recherché par beaucoup des outils d'IA est de pouvoir discriminer entre deux situations (par exemple pathologie/non pathologie en matière médicale, situation représentative d'une menace/situation non porteuse d'une telle menace en matière de sécurité intérieure). Ainsi, pour apprendre à un algorithme à différencier ces deux situations, il convient de disposer d'exemples des deux catégories, et donc de données n'ayant rien à voir avec la pathologie ou la menace, ce que ne contiennent pas forcément les bases de données pouvant servir de « matière première » à l'apprentissage.

Ce cadre dérogatoire devrait également s'appliquer aux phases de test, de contrôle-qualité et d'audit des algorithmes d'IA notamment lorsque la puissance publique est appelée à choisir entre différentes solutions techniques.

[191] Autorité administrative indépendante chargée de veiller au bon fonctionnement du marché de l'énergie et d'arbitrer les différents entre utilisateurs et exploitants de l'électricité et du gaz naturel.

En contrepartie, la faculté de recourir à ce cadre expérimental devrait être encadrée, s'agissant notamment de données particulièrement sensibles comme les images et vidéo, par des conditions portant sur :

- la démonstration que ces jeux de données d'apprentissage apporteraient une plus-value significative par rapport aux jeux d'apprentissage disponibles en source ouverte ou constitués par simulation d'acteurs volontaires ;
- l'application de techniques de pseudonymisation compatibles avec les usages du jeux d'apprentissage (par exemple en matière d'images, effacement des métadonnées de lieu ou date de prise de vue, floutage automatique du visage ou de la silhouette lorsque l'algorithme n'a pas besoin de cette information) ;
- le recours à un tiers de confiance garantissant une totale étanchéité des jeux de données d'apprentissage par rapport à l'exploitation opérationnelle de l'outil développé ou testé, et l'impossibilité d'y accéder à toute personne non directement impliquées dans la mise au point en R&D du dit dispositif.

Au-delà, une expérimentation en matière d'IA, qui voudrait pratiquer la « fouille de données » sans *a priori*, peut se heurter au principe de proportionnalité. Le dispositif de bac à sable proposé devrait également prendre en compte cette difficulté.

S'agissant d'une dérogation au droit régissant les données personnelles, la mission estime que ce pouvoir de dérogation devrait revenir à la CNIL en tant que garante de la protection desdites données, et lui permettre de déroger aux textes régissant les traitements source du jeux d'apprentissage, que ces textes aient été pris sur le fondement de la loi informatique et liberté ou sur le fondement d'autres textes sectoriels. La création de ce bac à sable au bénéfice de la CNIL implique une modification de la loi informatique et liberté. Une modalité de mise en œuvre serait de confier à la CNIL l'autorisation de la réutilisation des données à des fins d'apprentissage et, le cas échéant, leur conservation pour une durée supérieure à la durée prévue dans le traitement initial, dès lors que ce dernier est régi par un acte réglementaire pris sur le fondement des articles 31 ou 32, ou a fait l'objet d'un avis au titre de l'article 8.I.4°) a) de la loi de 1978. Depuis la révision de la loi, consécutive à l'entrée en vigueur du RGDP, la CNIL ne dispose d'un pouvoir d'autorisation que dans le domaine de la santé pour les traitements qui ne sont pas conformes à un référentiel (article 66.III). En contrepartie, la commission devrait statuer sur la demande de dérogation dans un délai raisonnable (de l'ordre de deux mois, renouvelable une fois) et serait dans l'obligation de motiver son refus.

> **Recommandation** : Créer un dispositif de bac à sable expérimental permettant à la CNIL de déroger aux textes existants pour autoriser la réutilisation de données personnelles dans des jeux d'apprentissage d'algorithmes d'intelligence artificielle, et leur conservation pour une durée plus longue que celle autorisée lors de leur collecte initiale.

Au-delà du droit des données personnelles, la mission estime nécessaire d'examiner si d'autres réglementations portant sur l'usage des donnés (par exemple en matière statistique, fiscale, etc.) pourraient utilement faire l'objet de dérogations de type « bac à sable », au cas par cas, encadrées et évaluées, afin de faciliter l'expérimentation en matière d'intelligence artificielle.

Promouvoir au niveau européen une vision plus proactive du cadre réglementaire au regard des évolutions technologiques induites par l'IA

L'adaptation réactive des règles juridiques par rapport à des technologies en progrès rapide doit imprégner la conduite de nos politiques publiques au plus haut niveau décisionnel, particulièrement dans le cadre européen, à défaut, pour les pays de l'Union européenne, d'être durablement voire irrémédiablement distanciés par les écosystèmes asiatiques et anglo-saxons. La sortie de l'UE (donc du cadre RGPD) du Royaume-Uni, premier écosystème d'innovation numérique en Europe, renforce l'acuité des débats sur la question des règles équitables de concurrence en matière de libre-échange (*level-playing field*).

Le Conseil européen[192] ouvre la voie à une évolution en matière de sas réglementaires, définis comme des cadres concrets qui, en offrant un contexte d'expérimentation structuré, permettent le cas échéant de tester en situation réelle des technologies, des produits, des services ou des approches innovants – tout particulièrement à l'heure actuelle, dans le cadre de la transition numérique – pendant une durée limitée et dans une petite partie d'un secteur ou d'un domaine, sous contrôle réglementaire, en veillant à ce que des garanties appropriées soient en place. Les clauses d'expérimentation, qui constituent souvent la base juridique des sas réglementaires, sont définies comme étant des dispositions juridiques permettant aux autorités chargées de mettre en œuvre et de faire appliquer la législation de faire preuve au cas par cas d'une certaine souplesse pour ce qui est de tester des technologies, des produits, des services ou des approches innovantes.

Ces conclusions attestent d'une prise de conscience des enjeux en la matière dont la traduction/le prolongement concrèt(e) sont attendus en matière de droit des données personnelles.

Dans sa revue du RGPD de juin 2020, la Commission européenne a appelé de ses vœux la rédaction de lignes directrices sur l'IA. Ce sujet figure dans la feuille de route du Comité européen de la protection des données (CEPD), même si aucun document spécifique n'a été publié à ce stade.

L'éthique de l'intelligence artificielle

Permettre de développer une intelligence artificielle en France, et renforcer ainsi notre autonomie stratégique, pour éviter une dépendance vis-à-vis d'algorithmes qui ne correspondent pas à nos valeurs et nos choix de société, nécessite d'engager des actions pour une éthique de l'intelligence artificielle. Sans elles, l'acculturation des populations à l'usage de l'IA comporte le risque d'un rejet massif. 87 % des Français se disent mal informés sur l'utilisation faite de leurs données personnelles et 60 % des Français déclarent se méfier de l'IA[193].

Conscient des enjeux nationaux et européens autour de l'intelligence artificielle, le gouvernement souhaite voir émerger des initiatives françaises innovantes. Lors de la conférence de remise du rapport Villani, un des points centraux soulevés a été le besoin d'une certification de l'IA pour prévenir les risques de discrimination. De nombreux exemples venus des États-Unis montrent que, si on n'y prend pas garde, l'IA peut rapidement produire des actions à caractère sexiste et raciste. On peut ainsi redouter l'utilisation d'outils, particulièrement de reconnaissance faciale, qui comporteraient des biais dans leur constitution, par exemple en n'ayant été entraînés qu'à partir de visages d'hommes seulement, ou de femmes seulement, et dont on voudrait se servir sur toute la population. Ce risque doit d'autant moins être sous-estimé que le recours à l'IA peut être perçu, à tort, comme l'utilisation d'une solution infaillible parce que fondée sur des paramètres technologiques avancés.

Il est donc essentiel que l'intelligence artificielle puisse être formée à partir de données aussi vastes et riches que possible, et qu'une éthique de son utilisation soit élaborée.

À l'étranger, plusieurs pays se sont engagés dans le développement d'outils permettant d'évaluer les impacts éthiques de leurs projets. Le Canada a ainsi publié en 2020 un outil d'évaluation de l'incidence algorithmique, dont l'usage est obligatoire pour tous les nouveaux projets gouvernementaux[194].

[192] « Conclusions sur le rôle des sas réglementaires et des clauses d'expérimentation dans un cadre réglementaire de l'UE propice à l'innovation, à l'épreuve du temps, durable et résilient » adoptées le 16 novembre 2020

[193] Sondage Harris-interactive pour Quantmetry, 2016.

[194] https://www.canada.ca/fr/gouvernement/systeme/gouvernement-numerique/innovations-gouvernementales-numeriques/utilisation-responsable-ai.html

La France a participé à la constitution d'un groupe d'experts de haut niveau à l'échelle de l'Union européenne, qui a produit des lignes directrices sur l'éthique de l'IA en 2019. Ces lignes directrices doivent être désormais déclinées de manière opérationnelle, ce qui fait l'objet d'un travail européen auquel la France participe, afin que les principes conformes à nos choix de société puissent être mis en application.

En outre, un comité pilote d'éthique du numérique a été créé au sein du comité consultatif national d'éthique, compétent en matière d'IA[195]. Le comité est constitué d'environ 30 personnes et dirigé par Claude Kirchner, directeur de recherche émérite de l'INRIA, et réunit des spécialistes du numérique, académiques ou issus des entreprises, des philosophes, des médecins, des juristes et des membres de la société civile. Le comité a commencé à rendre des avis, les premiers portant sur les agents conversationnels, le véhicule autonome et le diagnostic médical à l'ère de l'IA.

Par ailleurs, sur l'une des dimensions de l'éthique, en l'occurrence la transparence et l'explicabilité, la France s'est dotée, avec **la loi pour une République numérique**, d'une législation à l'avant-garde des pratiques de ses partenaires étrangers, **imposant aux administrations de fournir des explications détaillées sur les décisions administratives individuelles prises sur le fondement d'un traitement algorithmique.**

Des initiatives privées existent également, comme l'émergence du label *Fair Data Use* (FDU) de la société Maathics, qui se présente comme le premier « smart label » pour une utilisation équitable de l'IA, et qui s'appuie sur la confiance des utilisateurs, la conformité au RGPD, et une démarche RSE. L'objectif du label FDU est de s'assurer qu'un service d'IA soit équitable et sans discrimination.

4.2. Mieux valoriser les données personnelles en mettant en œuvre les dispositions permettant le croisement de fichiers à partir du NIR « haché » ou statistique

Les administrations françaises disposent de nombreux fichiers comportant des données personnelles protégées par les dispositions de la loi informatique et libertés du 6 janvier 1978. Ces fichiers sont une source d'information très riche pour la statistique publique et pour la recherche, surtout lorsqu'on peut mettre en relation les données que comportent deux traitements (« appariement » des fichiers). Le moyen le plus sûr pour garantir un appariement de qualité entre deux fichiers est de recourir au numéro d'identification au répertoire national d'identification des personnes physiques (NIR) dit aussi « numéro de sécurité sociale ». En raison de son caractère signifiant, il a été historiquement réservé au secteur de la sécurité sociale. La CNIL en a progressivement élargi les usages possibles.

Il existe deux régimes juridiques encadrant les modalités de recours au NIR pour des besoins d'appariement de fichiers :

- en premier lieu, le décret « cadre NIR »[196] indique, au sein d'un nombre de secteurs identifiés (protection sociale, santé, logement, travail, justice, financier, fiscal et douanier, statistique publique et recensement, éducation), quels sont les usages possibles de ce numéro. Dès lors qu'une administration souhaite recourir au NIR pour un autre usage, ce décret en conseil d'Etat doit être modifié, ce qui implique une procédure lourde et longue (un délai de 18 mois est généralement évoqué) ;

[195] Comité consultatif national d'éthique pour les sciences de la vie et de la santé, communiqué de presse, 2 décembre 2019.

[196] Décret n° 2019-341 du 19 avril 2019 relatif à la mise en œuvre de traitements comportant l'usage du numéro d'inscription au répertoire national d'identification des personnes physiques ou nécessitant la consultation de ce répertoire

- pour faciliter l'usage du NIR dans les travaux de la statistique ou de la recherche publique, l'article 34 de la loi pour une République numérique a donc créé une procédure spécifique permettant de créer un code statistique non signifiant (CSNS), dérivé du NIR (« NIR haché » ou « NIR statistique »), traduite dans l'article 30 de la loi informatique et libertés modifiée. Le décret d'application[197] en a défini les modalités d'application en distinguant deux cas d'usage : un code pérenne pour les usages dans le domaine statistique d'une part, un code à durée limitée pour la recherche d'autre part, ce numéro étant spécifique à l'étude, afin d'éviter tout croisement ultérieur de données.

Cet article qui a constitué une importante avancée dans la valorisation des données n'est pas opérationnel à ce jour. L'encadré ci-dessous présente les travaux conduits par l'INSEE

Mise en œuvre du projet « code statistique non signifiant » par l'INSEE

Le projet code statistique non signifiant (CSNS) développé au sein de l'INSEE a pour ambition de proposer une aide aux appariements de données. Il s'insère au sein d'un environnement plus global de traitements de fichiers, dont le but est de rapprocher entre elles des données qui concernent des individus. En facilitant les appariements, ce projet devrait permettre notamment de réduire la charge statistique d'enquêtes auprès des ménages.

Le projet a aussi pour objectif de proposer une organisation facilitant la diffusion raisonnée du CSNS au sein du système statistique public (SSP), de façon sécurisée et abordable pour tous par souci d'équité. Pour encourager son utilisation et faire des gains d'efficience, son insertion progressive dans des fichiers « pivots[198] » pourra servir collectivement des traitements identifiés ou futurs. Les fichiers « pivots » ne sont pas encore formellement identifiés, mais on peut penser aux sources fiscales ou sociales, voire au recensement de la population.

Le produit final est un service offert à l'ensemble du SSP, qui consiste à délivrer un CSNS à un responsable de traitement, ce dernier ayant en amont fourni soit un NIR, soit des éléments d'état civil permettant de le déterminer. Le service a vocation à être pérenne, par le renouvellement du CSNS dès que nécessaire (a minima tous les 10 ans) et doit fonctionner selon une procédure fiable garantissant la sécurité des échanges de données.

Pratiquement, le CSNS est obtenu par une opération cryptographique consistant, à l'aide d'une clé secrète, à transformer le NIR en un identifiant non porteur d'information, soit non signifiant. Ce processus est éventuellement précédé d'une étape d'identification au Répertoire national d'identification des personnes physiques (RNIPP) si le NIR n'est pas connu du responsable de traitement mais que seuls des éléments d'état civil (traits d'identité) sont fournis.

Conformément à l'article 6 du décret n°2016-1930, le CSNS est détruit par le service Insee dès sa validation par le destinataire ; ce dernier doit alors le conserver dans un lieu sécurisé, hermétique à celui des données statistiques et des données identifiantes.

Source : Mission ; INSEE.

L'INSEE indique qu'une première version permettant d'obtenir un CSNS à partir d'un NIR sera disponible fin 2020 et une seconde version, intégrant un moteur d'identification statistique pour les cas où le NIR n'est pas disponible, est prévue pour fin 2021. Enfin, une version finale, offrant un moteur d'identification à son terme et une interface d'utilisation complète devrait intervenir au 4e trimestre 2022.

[197] Décret n° 2016-1930 du 28 décembre 2016 portant simplification des formalités préalables relatives à des traitements à finalité statistique ou de recherche.

[198] Un fichier est qualifié de « pivot » s'il contient des variables potentiellement utiles dans de multiples appariements de fichiers (par exemple, des variables fiscales ou socio-démographiques).

L'INSEE souhaite par ailleurs mettre en place une gouvernance pour les usages du CSNS et plus largement les traitements mobilisant des appariements, dans une politique de transparence, garantissant l'information des personnes concernées et la justification de leurs finalités. Elle souhaite, pour ce faire, s'appuyer sur le Conseil national de l'information statistique (Cnis). Il sera amené à émettre un avis sur tous les traitements d'appariements formulés par le SSP, à l'image de ce qui est fait pour les enquêtes, l'appariement étant un mode de collecte comme un autre. Ces traitements seront ainsi spécifiés et portés à la connaissance du public. Tout traitement impliquant le CSNS à des fins de croisement de données devra par ailleurs, de la part du responsable de traitement concerné, répondre de sa conformité aux obligations du RGPD et de la loi informatique et libertés, notamment son inscription au registre des traitements et l'information des personnes concernées, sur les finalités, les durées de conservation des données et sur leurs droits. Il devra également au préalable faire l'objet d'une présentation lors d'une commission du Cnis.

> **Recommandation** : Mettre en œuvre les dispositifs techniques permettant d'utiliser la procédure d'appariement de fichiers sur la base du code statistique non signifiant à des fins de statistique publique et de recherche scientifique et historique.

CAS D'USAGE – Nam.R

La start-up Nam.R a été fondée en 2017 dans le but de permettre l'exploitation de données géo-localisées dans un objectif de développement durable. Précisément, Nam.R a développé une expertise dans les données concernant tous les secteurs de la transition écologique : développement des énergies renouvelables, opérations d'efficacité énergétique, *smart grids*, circuits courts. Nam.R a également développé ses propres plateformes pour l'exploitation de ces données afin de proposer à ses clients ou partenaires une offre de service complète. Sa force est son expertise en *machine learning*, traitement de l'image et du langage naturel. L'équipe se compose de 40 personnes.

Du point de vue économique, l'entreprise n'utilise que des données ouvertes, ce qui lui permet d'être totalement indépendante vis-à-vis d'éventuels fournisseurs de données puisque l'information est récupérée sans condition ni contrepartie. Par ailleurs, ces données géolocalisées sont toujours anonymisées et non liée aux consommateurs/utilisateurs. Enfin, l'entreprise est entièrement financée par fonds propres. Il n'y a donc pas de dépendance financière envers un éventuel prêteur.

Pour valoriser au mieux les données, Nam.R a développé une « Data Library ». L'objectif était de construire une base de connaissance structurée la plus vaste possible. L'entreprise a recensé toutes les sources d'*open data* et effectue une veille permanente sur ce sujet. Cela lui permet de mettre à jour quotidiennement, via une technique de *scrapers*, ses jeux de données et d'en extraire des métadonnées. Une technique de *miners* permet d'harmoniser les données *via* un traitement du langage naturel. Enfin, un processus de traitement en Machine Learning vient compléter l'ensemble. Cette base de connaissance fait l'objet de partenariats avec des acteurs comme OpenData France, Etalab ou la Cour des comptes, entre autres.

Un exemple d'application, sur lequel Nam.R a travaillé en 2019, est la solarisation de la France, à savoir optimiser le rendement des panneaux solaires via le choix de leur emplacement. Nam.R a commencé par récolter les données (routes, ponts, maisons, bâtiments, etc.) mises à disposition par les acteurs publics ou par le biais de partenaires comme l'Institut national de l'information géographique et forestière (IGN). En ajoutant de multiples autres paramètres tels que le nombre d'étages d'un bâtiment, la présence d'un groupe de climatisation sur le toit, le matériau, la forme ou la pente du toit, et en s'appuyant sur sa Data Library, Nam.R a été alors en capacité d'optimiser le choix de l'emplacement des infrastructures solaires, que ce soit à l'échelle d'un bâtiment ou d'un quartier.

CAS D'USAGE – Le projet Vidéoprotection ouverte et intégrée (VOIE)

Contenu et cadre du projet

Le projet Vidéoprotection Ouverte et IntégréE (VOIE), financé par Bpifrance jusqu'en novembre 2018 visait à tester du point de vue technique et opérationnel, des algorithmes permettant la détection d'évènements violents (rixes, agressions), la détection automatique d'anomalies (intrusions, franchissement de barrières et de voies) et le suivi de personnes (par exemple à la suite de la découverte d'un bagage ou colis abandonné) sur des images issues de dispositifs de vidéoprotection. Le projet avait également pour but d'explorer les dimensions légales et d'acceptabilité sociétale, permettant, in fine, de mesurer la proportionnalité entre les « contraintes » de ces nouveaux outils sur les données personnelles et leur efficacité réelle sur le terrain.

S'agissant du volet RATP, le traitement des incidents concernait le pôle multimodal urbain de Châtelet. L'objectif était d'évaluer l'efficacité de la vidéo intelligente associée à d'autres systèmes ; une centaine de scénarios de détection d'alertes sans IA (machine learning) ont été étudiés, comportant notamment des rixes et agressions.

Un travail similaire a été conduit sur plusieurs sites SNCF dont la gare du Nord et ses approches, afin de varier les conditions et d'être plus représentatif des situations réelles (extérieur, intérieur, tunnel, grande mezzanine, zone publique et zone à accès limité). Les scénarios étudiés ont porté sur l'intrusion, la présence de vendeurs à la sauvette et les sauts de portiques.

Le projet VOIE a rencontré plusieurs obstacles dans l'expérimentation envisagée en raison du cadre juridique régissant la protection des données personnelles. Les restrictions imposées par la CNIL ont porté notamment sur :

- l'interdiction de recours à la biométrie dans les algorithmes, à défaut de décret en Conseil d'Etat pris après avis motivé et publié de la CNIL autorisant cette expérimentation ; le traitement testé ne doit pas permettre d'identifier un individu de manière unique ;
- une limitation stricte des conditions de captation des images concernant à la fois les images utilisées (s'agissant de l'expérimentation RATP, enregistrement « à la volée » de deux séquences de deux heures par semaine pendant 10 semaines –soit une quarantaine heures), et pour le volet « algorithme de suivi », le recours à des personnes volontaires ayant formalisé leur consentement par écrit ;
- la suppression des images utilisées à l'issue de l'expérimentation, comme cela fut le cas pour le projet Secur-id mené en 2011/2014 sur les mêmes thématiques, et pour le projet Victoria conduit de 2017à 2020, dans le prolongement de VOIE.

Finalement, les opérateurs ont dû reconstituer des scénarios joués par des plastrons ce qui limite grandement l'étendue des tests. En outre, la destruction systématique des jeux de données ne permet pas de mesurer les progrès réalisés. Enfin, les restrictions posées à l'expérimentation sur le suivi de personnes (interdiction des dispositifs permettant de retrouver ou reconnaître la personne physique) ne permettent pas de mettre au point un outil fiable fondé sur l'apparence générale si l'individu change volontairement d'apparence, ou lorsque les codes vestimentaires ont tendance à s'uniformiser (par exemple en hiver).

Les enseignements à tirer : un cadre juridique inadapté à l'expérimentation des innovations technologiques en matière de vidéo-protection dans l'espace public

Les opérateurs de vidéo-protection ne peuvent pas, même sous un contrôle étroit et selon des protocoles stricts qui protégeraient les images d'entraînement de toute utilisation illégitime,

coopérer avec des développeurs pour leur propre compte dans de bonnes conditions de sécurité juridique.

Le cadre législatif et réglementaire en vigueur, ainsi que la doctrine de la CNIL qui en découle, ne fait pas la distinction entre une démarche expérimentale (donc encadrée dans l'espace et le temps) qui requiert l'utilisation de jeux de données d'apprentissage pertinentes, et l'exploitation de tels dispositifs en phase opérationnelle relevant d'un cadre s'appliquant sur le territoire et en toute situation.

Dans plusieurs décisions d'autorisation délivrées ces dernières années à des entreprises privées de traitement vidéo ainsi qu'à des opérateurs publics de vidéo-protection, la CNIL a conditionné son accord 1/ au recours exclusif à des volontaires après recueil exprès de leur consentement, ce qui ne permet pas de modéliser toutes les situations de manière « naturelle », et pose des problèmes pratiques difficilement surmontables pour les situations dont les occurrences sont rares ; 2/ à la suppression des données collectées à l'issue de l'expérimentation. Ainsi, dans le projet Victoria précité, près d'un quart des moyens consacrés au projet l'ont été à constituer des jeux de données, y compris en rédigeant des scénarios et en faisant tourner les scènes par des comédiens.

Cette situation présente de nombreux inconvénients :

- l'entraînement se fait majoritairement sur des données collectées à l'étranger (par exemple celles du *National Institute of Standards and Technology* – NIST – qui dépend du ministère de l'industrie américain, ou celles d'universités étrangères) qui ne coïncident pas exactement avec la réalité du territoire national : les formats de véhicules, d'équipements publics, de bâtiments, les caractéristiques physiques et comportementales des individus etc., présentent des différences qui limitent la performance des algorithmes ;
- l'entraînement des algorithmes peut être insuffisamment contextualisé par rapport aux cas d'usages concrets dans lesquels ils seront employés, notamment lorsque la base d'entraînement ne s'effectue que sur des données artificiellement reconstruites avec des plastrons qui ne « jouent » pas en conditions réelles ;
- les chercheurs ou industriels français peuvent moins facilement contrôler les biais éthiques des algorithmes pré-entraînés ou des jeux de données ainsi acquis, ce qui limite leur auditabilité ; il en va de même pour la puissance publique vis-à-vis des protocoles d'entraînement en cas de litige pénal ou mettant en jeu la responsabilité civile ;
- cette situation présente enfin un désavantage compétitif pour les sociétés de traitement vidéo françaises, qui a des conséquences directes pour notre tissu industriel et pour notre autonomie stratégique.

Partie 4
Se donner les moyens de nos ambitions

1. Renforcer les compétences

1.1. Une acculturation nécessaire des acteurs publics au numérique

La très grande majorité des acteurs interrogés et la consultation publique conduite par la mission soulignent un fort besoin d'acculturation des agents publics aux enjeux numériques, à commencer par les cadres dirigeants. Une administration a par exemple cité le fait de mettre en ligne un des textes de loi dont elle est responsable comme une pratique d'ouverture de code source. Un autre témoignage de ce manque de compréhension se lit dans une contribution de la consultation publique, faisant état de la « *confusion dans l'esprit de nombre d'agents entre l'ouverture et la publicité des données individuelles d'une part* », et de « *l'absence de perception de la valeur économique de la donnée*. »

L'utilité et le potentiel des données et des codes sont mal compris, et restent identifiés dans l'imaginaire collectif comme des sujets purement « techniques », qui ne sont pas au service de l'élaboration et du pilotage des politiques publiques. Rappelons que certaines directions métiers sollicitées par la mission ont répondu ne pas se sentir concernées, renvoyant à leur direction du numérique le soin d'apporter des réponses. Pour répondre à cette problématique, des actions sont menées, par exemple par la direction du numérique des ministères sociaux en complémentarité avec les services statistiques ministériels (la direction de la recherche, des études, de l'évaluation et des statistiques (DREES) et la direction de l'animation, de la recherche, des études et des statistiques (DARES)), afin de sensibiliser sur l'importance des données : des groupes de travail et des ateliers de démonstration sur les données sont organisés avec les directions et services métiers en vue de développer une offre sur l'ouverture des données, leur visualisation, l'aide au ciblage et le prédictif.

Par ailleurs, **les cadres de la fonction publique sont insuffisamment formés au potentiel des données publiques et aux moyens à y consacrer**, ce qui a pour conséquence un manque de portage managérial de cette politique au sein des directions métiers. Au-delà du portage politique nécessaire (cf. partie 2, paragraphe 3.1), la politique publique de la donnée et des codes sources a besoin d'une impulsion qui soit donnée par les cadres et les décideurs publics à tous les niveaux.

Conscientes de cet enjeu, la direction interministérielle du numérique (DINUM) et la direction générale de l'administration et de la fonction publique (DGAFP) mènent une action concernant la « mise en œuvre d'un dispositif d'accompagnement des cadres dirigeants leur permettant d'intégrer l'impact du numérique sur la sphère publique »[199]. Cette mesure s'intègre dans les actions prioritaires prévues dans le schéma directeur de la formation professionnelle tout au long de la vie des agents de l'État[200]. Cette action comporte deux axes majeurs : d'une part, un mentorat de haut niveau pour les cadres dirigeants en poste, et d'autre part, des immersions dans les start-up d'État pour les cadres dirigeants entre deux postes.

En sus de ce plan de formation, l'organisation annuelle d'un ou de séminaires à destination des cadres dirigeants de la fonction publique (en débutant par les directeurs des administrations centrales, puis les sous-directeurs) sur la politique d'ouverture des données et des codes sources pourrait compléter les actions d'ores et déjà menées par la DINUM et la DGAFP.

[199] Action 15 du plan d'actions dans le plan d'actions pour la filière numérique et des systèmes d'information et de communication, circulaire DINUM / DGAFP du 2 mai 2019.

[200] Axe III – Action prioritaire n° 7 : « Diffuser une culture managériale commune en rendant obligatoire une formation au management pour tout primo-encadrant et tout agent nommé à la direction d'un opérateur ministériel et en proposant des modules adaptés aux besoins de formation des managers tout au long de leurs parcours professionnels » notamment l'objectif partagé : «accompagner la transition numérique ».

Par ailleurs, la mission considère qu'il convient de prévoir un plan de formation aux enjeux du numérique non seulement pour les cadres dirigeants, mais aussi pour l'ensemble des niveaux hiérarchiques, avec un volet portant sur la politique d'ouverture des données et des codes sources. Du reste, un budget de 22 M€, décidé dans le cadre du plan de relance, doit venir financer des actions d'accompagnement de l'ensemble des agents dans l'appropriation des enjeux du numérique.

En outre, la mission recommande d'une part que l'offre de formation interministérielle fasse l'objet d'une plus grande communication au niveau des ministères et d'autre part que le plan interministériel de formation proposé par la DGAFP et la DINUM soit complété au niveau ministériel afin de prendre en compte plus spécifiquement les besoins de chaque ministère en lien avec l'AMDAC.

Au-delà de la formation continue des agents publics, il conviendrait d'inclure, dans les différents cursus de formation des agents publics, des modules sur les enjeux du numérique, en incluant un volet sur la politique d'ouverture des données et des codes sources.

> **Recommandation :** Développer une politique de formation de la fonction publique plus ambitieuse sur les enjeux du numérique (obligation de formation des cadres dirigeants aux enjeux du numérique, séminaires de cadres dirigeants, offre de formation pour tous les niveaux hiérarchiques, plans de formation ministériels complémentaires à l'offre interministérielle, modules dans l'ensemble des cursus de formation de la fonction publique)

Par ailleurs, la problématique d'acculturation aux enjeux du numérique se pose avec au moins autant d'acuité pour ce qui concerne les élus. Du reste, les élus et les agents travaillant au sein de collectivités territoriales auditionnés par la mission ont tous souligné l'importance du soutien des élus dans la démarche d'ouverture des données et des codes sources : sans portage politique fort, la politique d'ouverture des données et des codes sources ne peut s'envisager faute d'une priorisation des moyens humains, techniques, financiers et juridiques. Aussi, la mission recommande de former les élus aux enjeux de la donnée et des codes sources dans les politiques publiques.

Il convient ici de rappeler que la formation des élus locaux est actuellement structurée autour de deux cadres distincts :

- d'une part, les collectivités sont dans l'obligation d'inscrire, chaque année, un montant minimal de crédits dédiés à la formation de leurs élus dans leur budget prévisionnel. Les formations éligibles à ces financements sont uniquement les formations liées à l'exercice du mandat, qui ne peuvent être dispensées que par un organisme agréé à cet effet par arrêté du ministre chargé des collectivités territoriales, pris après avis du Conseil national de la formation des élus locaux (CNFEL) ;
- d'autre part, le droit individuel à la formation (DIF), créé par la loi du 31 mars 2015 visant à faciliter l'exercice, par les élus locaux, de leur mandat, permet à l'ensemble des élus d'acquérir des droits à formation à raison de de 20 heures par année complète de mandat. Les formations éligibles à ce DIF recoupent un champ plus large, puisqu'elles peuvent concerner l'exécution du mandat comme la réinsertion professionnelle ; l'élu est libre d'en disposer. Le DIF est financé par des cotisations prélevées sur les indemnités de fonction des élus, et les collectivités territoriales ne participent donc pas à son abondement.

Le contenu des ordonnances visant à réformer ces dispositifs, conformément à l'habilitation législative prévue à l'article 105 de la loi n° 2019-1461 du 27 décembre 2019 relative à l'engagement dans la vie locale et à la proximité de l'action publique, est en cours d'arbitrage. Ses principaux objectifs sont de faciliter le recours à la formation par les élus, en améliorant les dispositifs existants et leur articulation, tout en clarifiant le champ des formations qui relèvent spécifiquement de l'exercice du mandat.

Sans aller jusqu'à imposer une liste de formations obligatoires en début de mandat à l'élu – ce type d'obligation n'existant pas à l'heure actuelle - la mission recommande de développer une offre de formation adaptée aux élus et de l'inclure dans le périmètre des formations liées à l'exercice du mandat, tout en laissant le choix à l'élu.

Recommandation : Proposer une offre de formation dédiée aux élus sur les enjeux de la donnée et des codes sources dans les politiques publiques

1.2. Recruter et de fidéliser les compétences dans les métiers du numérique et des systèmes d'information et communication (NSIC)

Au-delà des enjeux d'acculturation des agents publics à la politique de la donnée et des codes sources, se pose la question des compétences disponibles pour mettre en œuvre cette politique. La DGAFP et la DINUM ont ainsi identifié 15 métiers en tension dans le numérique au sein de l'État, qui figurent ci-dessous :

- **études et développement** : concepteur-développeur ; intégrateur d'applications ;
- **exploitation des infrastructures** : administrateur outils/systèmes/réseaux & télécoms ; intégrateur d'exploitation ;
- **urbanisation** : architecte technique ; urbaniste des SI ;
- **gestion de projet des systèmes d'exploitation** : chef de projet maîtrise d'œuvre ; chef de projet maîtrise d'ouvrage ; urbaniste des SI ;
- **métiers transverses** : acheteur IT ;
- **gestion et exploitation des données** : *data scientist* ;
- **sécurité** : analyse en détection d'intrusions ; auditeur en sécurité ; responsable sécurité ;
- **agilité** : *scrum master*.[201]

La DGAFP et la DINUM ont donc élaboré un plan d'actions pour la filière numérique et des systèmes d'information et de communication permettant d'attirer, de recruter et de fidéliser les profils adéquats. Ce plan fait l'objet d'une circulaire du 2 mai 2019 qui vise trois objectifs : i) le recrutement et le vivier de compétences ; ii) la mobilité et les parcours ; iii) le recours au contrat pour renforcer l'attractivité des postes.

Le premier objectif est d'attirer et de recruter les bons profils dans le vivier spécifique des agents de la filière NSIC.

Pour parvenir à cet objectif, la circulaire prévoit la rénovation des modalités de recrutement, la constitution de viviers, le développement de la « marque État employeur du Numérique et des SIC» et des partenariats avec les écoles, ainsi que par une démarche de gestion prévisionnelle des emplois, des effectifs et des compétences (GPEEC) interministérielle spécifique à la filière NSIC.

Le renforcement en compétences NSIC s'appuie notamment sur la rénovation du corps à vocation interministérielle des ingénieurs SIC (ISIC) géré par le ministère de l'Intérieur. Ainsi, la circulaire prévoit la création d'un concours interne spécial, suivi d'une formation, ouvert à l'ensemble des agents publics, quelles que soient leur filière ou catégorie ainsi qu'une modernisation des modalités de recrutement des agents titulaires dans le corps des ISIC, afin d'assurer une meilleure adéquation des profils recrutés aux postes offerts. À fin 2018**, le corps des ISIC comptait ainsi 673 ingénieurs, chiffre en croissance de 43% depuis 2012 et 35% depuis 2016, croissance expliquée avant tout par l'augmentation des effectifs de moins des ingénieurs de moins de 30 ans** (qui ont triplé entre 2012 et 2018, et plus que doublé entre 2016 et 2018).

Par ailleurs, l'élaboration d'une démarche opérationnelle de gestion prévisionnelle des emplois, des effectifs et des compétences (GPEEC) vise en premier lieu à constituer une base de données exhaustive, fiable et actualisée, qui fasse apparaître le rattachement de chaque poste et agent au Répertoire interministériel des métiers et de compétences NSIC. La circulaire prévoit la réalisation d'une cartographie des effectifs par métier et compétences ainsi que des besoins futurs en effectifs par métier et compétence, en s'appuyant sur les travaux réalisés en 2017. Cette démarche doit également permettre d'identifier les besoins en formation.

[201] Source : DGAFP, DINSIC, 2017.

La mission ne peut que soutenir cette démarche de GPEEC prévue dans la circulaire. Cependant, elle invite à prioriser cette action qui n'a pas réellement démarré, faute de priorisation suffisante. En outre, **elle considère nécessaire de prendre en compte l'ensemble des corps susceptibles d'apporter des compétences NSIC**, parmi lesquels figurent les administrateurs de l'INSEE, les attachés de l'INSEE, les ingénieurs des Mines, etc. À cet égard, elle s'interroge sur l'évolution des effectifs dans les corps des administrateurs et des attachés de l'INSEE compte tenu des besoins identifiés en *data scientists*. Ainsi, le corps des administrateurs de l'INSEE comptait 494 administrateurs à fin 2018, en croissance de 2 % depuis 2016 mais avec une baisse de 9 % des effectifs chez les moins de 30 ans. Quant au corps des attachés de l'INSEE, il comptait 1662 attachés à fin 2018, en croissance de 3 % depuis 2016, avec une augmentation de 1% des effectifs chez les moins de 30 ans qui ne saurait masquer une baisse tendancielle des effectifs sur cette tranche d'âge de 10 % sur la période 2012-2018.

En complément de l'identification des compétences NSIC au sein des différents corps, le recrutement des contractuels disposant de compétences non disponibles dans les corps d'État est nécessaire. L'attractivité de ce statut doit toutefois être renforcée pour ces profils rares, au travers de contrats en CDI et de parcours professionnels intéressants (cf. infra).

> **Recommandation :** Poursuivre les travaux relatifs à la gestion des emplois, des effectifs et des compétences du numérique et structurer une filière technique de la fonction publique pour les métiers experts du numérique, en créant des parcours pour les corps techniques et en pérennisant en CDI les agents contractuels apportant des compétences non disponibles dans les corps existants

Le deuxième objectif du plan d'actions DGAFP-DINUM est de développer la mobilité et les parcours professionnels des agents de la filière dans une logique de « management des compétences » et comme levier de motivation des agents.

La circulaire de mai 2019 souligne que *« dans une filière où l'expertise technique et l'évolution rapide des métiers conditionnent la performance des services, le management des compétences est fondamental »*. Afin de répondre à cet enjeu, le plan d'actions vise à *« proposer des parcours de carrière riches et diversifiés, à la fois dans le domaine du management et dans la valorisation de l'expertise (« parcours d'expert »), développer la mobilité interministérielle et la formation professionnelle »*.

Dans son rapport de septembre 2020 sur les agents contractuels dans la fonction publique[202], la Cour des comptes souligne qu'*« une partie de ces compétences émergentes, notamment dans le domaine des métiers du numérique, peut ne pas présenter un caractère pérenne et / ou être à obsolescence rapide. La création de nouveaux corps ou cadres d'emplois ou le recrutement d'agents titulaires dans le cadre de corps ou cadres d'emplois existants par les employeurs publics ne serait pas, dans ce cas, une solution au problème posé. »* Aussi, la mission estime nécessaire d'accompagner la réflexion sur l'évolution des corps susceptibles d'apporter des compétences NSIC, d'une réflexion sur la mise à jour des compétences tout au long de la carrière. Cette problématique concerne également les agents contractuels recrutés en CDI.

Au-delà des actions de formation des agents, se pose la question des parcours, sujet évoqué dans la circulaire de mai 2019. Au-delà des actions prévues par cette circulaire, la mission estime nécessaire d'engager une réflexion spécifique avec les corps de l'INSEE afin de diversifier les parcours des administrateurs et des attachés dans l'ensemble des administrations, au-delà des services statistiques ministériels, et valoriser le travail et la carrière des agents choisissant ces parcours.

> **Recommandation :** Diversifier les parcours des administrateurs et des attachés de l'INSEE dans l'ensemble des administrations, au-delà des services statistiques ministériels, et valoriser le travail et la carrière des agents choisissant ces parcours

[202] Les agents contractuels dans la fonction publique (exercices 2010-2019), Cour des Comptes, septembre 2020.

La construction de parcours professionnels est enfin un facteur d'attractivité essentiel pour ces profils rares. Aussi, une expérimentation visant à faciliter le partage ponctuel d'expertise NSIC entre administrations afin de valoriser les compétences développées par les agents de la filière a été mise en place en 2019. En outre, la mobilité entre le secteur public et le secteur privé public des agents NSIC est encouragée afin de bénéficier d'un apport croisé d'expertise et de compétences. Au-delà des actions prévues par la circulaire de mai 2019, un rapport du CGE de décembre 2019 sur l'échange de compétences pour la réalisation de projets numériques de l'État formule un certain nombre de propositions autour d'un *« troc de compétences »* sous forme d'expérimentation.

Le troisième objectif du plan d'actions est ainsi de faciliter le recours au contrat pour les métiers NSIC.

Au-delà des compétences disponibles au sein des corps d'État, le recours aux agents contractuels permet d'apporter des compétences complémentaires nécessaires. Ainsi, la circulaire de mai 2019 souligne que *« compte-tenu de la forte technicité requise dans certains métiers NSIC et de l'évolution rapide des compétences associées, la gestion « duale » d'agents titulaires et contractuels au sein de la filière a vocation à perdurer »*.

Elle prévoit en premier lieu de faciliter le recrutement direct en contrat à durée indéterminée (CDI). La loi de transformation de la fonction publique du 6 août 2019 facilite ainsi le recours au CDI pour recruter des contractuels disposant de compétences que ne présenterait pas les candidatures de fonctionnaires. Néanmoins, la DGAFP n'a pas été en mesure de fournir des données sur le nombre de contractuels en CDI dans ces métiers de la filière NSIC, ce qui interroge sur le suivi et le pilotage de cette politique.

Le programme des entrepreneurs d'intérêt général (EIG) constitue une initiative prometteuse d'Etalab, initié en 2016 et intégré aujourd'hui au sein de la DINUM. Le principe du programme consiste à intégrer pendant 10 mois des profils numériques d'exception dans les administrations pour relever des défis d'amélioration du service public à l'aide du numérique et des données. Au-delà de l'intérêt de bénéficier de profils rares au sein des différents services de l'État à un coût moindre qu'une prestations équivalente[203], ce programme participe également à leur acculturation.

Ainsi, **133 EIG ont été accueillis dans 40 administrations depuis janvier 2017 au sein des différentes directions des ministères** (qui participent au minimum à hauteur de 40% du budget d'un EIG). 30 % d'EIG ont été pérennisés comme contractuels dans la fonction publique, témoignant du succès du programme. Néanmoins, **les effectifs des différentes promotions restent modestes au regard des besoins. Si la mission n'a pas pu obtenir de données sur le nombre d'EIG en CDI, il ressort des auditions que les EIG ont été très majoritairement pérennisés en CDD, atténuant le bilan positif de ce programme en matière d'intégration durable de nouveaux profils au sein des administrations**. Le succès relatif de ce programme montre non seulement que de nombreuses administrations sont prêtes à innover, mais aussi que de nombreux profils techniques sont volontaires pour contribuer au secteur public, surtout si leur contribution aide à l'ouverture de données, de codes sources et une façon différente de concevoir les métiers techniques dans le secteur public.

Ce programme EIG attire entre outre des personnes souvent très jeunes, qui trouvent là soit une première expérience dans le secteur public à leur sortie de formation, soit des problématiques différentes après une première expérience dans le privé.

> **Recommandation :** Passer à l'échelle et inscrire dans la durée le programme d'entrepreneurs d'intérêt général

Par ailleurs, la rémunération constitue un levier important pour renforcer l'attractivité du secteur public pour les métiers en tension. **Un premier référentiel sur les métiers en tension assorti d'un mode opératoire pour faciliter l'embauche de contractuels a été mis en place en mai 2019**. Ce référentiel, basé notamment sur les rémunérations servies dans le secteur privé, vise à faciliter les demandes de visa aux contrôleurs budgétaires et comptables ministériels (CBCM) sur les recrutements sous contrat, ainsi que de réduire la concurrence entre ministères sur les niveaux de rémunération offerts.

[203] Un EIG embauché sur dix mois représente un coût de 70 000 €, soit en moyenne un coût trois fois moins élevé que celui d'une prestation équivalente.

Un premier bilan réalisé par la DGAFP et la DINUM montre que le référentiel de rémunération a été utilisé pour 23 % des recrutements, soit 179 agents sur les 772 recrutés au total sur la période. D'après cette enquête, les difficultés rencontrées dans l'application du référentiel portent sur l'absence d'appropriation par les services (22 %), une difficulté d'articulation entre les services RH et les services budgétaires (22 %), un dialogue social perturbé (11 %), une interprétation du référentiel par le SRH (11 %), une difficulté de positionnement de certains postes qui ne recouvrent pas totalement les métiers décrits (11 %) et enfin un fort décalage entre la publication du référentiel de rémunération et sa mise en œuvre effective (11 %). Enfin, 29 % des répondants soulignent des effets reconventionnels (demande de réévaluation de salaire et d'IFSE) et des risques de départs d'agents déjà en poste du fait de l'application de ce référentiel.

Compte tenu des effets positifs de ce référentiel, ayant permis de recruter des profils adaptés aux postes, il est pérennisé et une réflexion est en cours pour élargir le référentiel à tous les métiers de la filière et pour proposer un nouveau cadre d'application du référentiel permettant de favoriser notamment les recrutements de candidats qui ne possèdent pas le niveau de diplôme requis mais qui présentent des compétences recherchées.

La mission ne peut que souligner les effets positifs de tels travaux et regrette néanmoins son application très partielle : elle engage donc à rendre ce référentiel obligatoire.

> **Recommandation :** Accroître l'attractivité de l'État pour les métiers du numérique en tension (rendre le référentiel de rémunération obligatoire, développer la communication auprès des formations spécialisées)

Enfin la circulaire de mai 2019 met en place un dispositif spécifique de gestion des contractuels de « haut niveau » concernant à la fois une politique de rémunération tenant compte des rémunérations servies dans le secteur privé et un parcours professionnel allant jusqu'à la nomination sur emploi-fonctionnel ou équivalent.

2. Développer l'utilisation des logiciels libres

Les efforts en matière de mutualisation des codes sources produits au sein de l'administration sont aujourd'hui assez faibles. Or, les gains associés à cette démarche aboutissant à la production de logiciels permettraient une meilleure utilisation des fonds publics et une plus grande transparence. La mission recommande donc d'élargir le périmètre de l'AGDAC et de lui confier une mission visant à promouvoir la publication et la réutilisation des codes sources au sein de l'administration au travers d'un Open Source Program Office. Pour accomplir cette mission, il convient d'une part de renforcer et rationaliser l'utilisation des logiciels libres dans la sphère publique, et d'autre part, d'animer une communauté au sein de l'État de sorte à accroître l'attractivité des profils.

2.1. Renforcer et rationaliser l'utilisation des logiciels libres

La mission recommande de mettre en place une stratégie sur l'utilisation des logiciels libres au sein de l'État ainsi qu'une animation interministérielle rassemblant les développeurs de l'État au travers de l'OpenSource Program Office (OSPO). Le besoin de soutien du logiciel libre dans le secteur public est la principale demande de la consultation publique conduite par la mission.

Au niveau interministériel, l'AGDAC aurait pour mission d'accompagner les ministères dans la définition et la mise en œuvre de leur politique d'utilisation des logiciels libres, et ce, en s'appuyant sur les AMDAC.

En matière d'organisation au sein de l'État, le décret relatif au système d'information et de communication de l'État et à la DINUM d'octobre 2019 prévoit qu'*« elle élabore et met à disposition des ressources numériques partagées ainsi que des méthodes et outils d'usage commun [...et] développe et soutient des produits, services et programmes innovants en vue de leur intégration dans les ministères »*. Dans ce cadre, le département Etalab propose des services relatifs aux logiciels libres, à travers notamment le maintien du socle interministériel des logiciels libres (SILL). Celui-ci est le catalogue de référence des logiciels libres dont l'usage est significatif dans l'administration. Pour chaque logiciel libre référencé, il existe un « référent SILL » au sein de l'administrations utilisatrice dont le rôle est d'indiquer la version minimale recommandée et de répondre aux éventuelles questions des autres administrations sur le logiciel porté. Par ailleurs, le SILL recense en grande majorité des logiciels libres développés par le secteur privé, et ne permet que partiellement de faire connaître des logiciels libres développés dans le secteur public comme le logiciel VITAM, logiciel libre d'archivage numérique développé conjointement par plusieurs ministères et opérateurs, pionnier en la matière.

Au niveau ministériel, il pourrait ainsi être demandé de décliner la politique d'utilisation des logiciels libres définie au niveau interministériel par l'AGDAC, au sein de chaque ministère, puis de la publier. Par exemple, le ministère des armées a défini une politique qui expose la stratégie de recours aux logiciels libres ainsi : « *L'emploi de logiciels libres favorise cette architecture modulaire et autorise une plus grande efficience dans son évolutivité là où cela est nécessaire. Cela constitue un apport très fort à l'innovation. Les logiciels libres permettent le développement de modules sur mesure, limités au strict besoin initial. L'usage du logiciel libre permet en effet de « piocher », en cohérence avec le Cadre de Cohérence Technique (CCT), dans les souches disponibles proposées par les communautés du logiciel libre.* »

Pour ce faire, des référents logiciels libres pourraient être nommés en appui des AMDAC, sur le modèle des « FLOSS evangelist »[204] de Renater (le réseau national de télécommunications pour la technologie l'enseignement et la recherche). En effet, le réseau est impliqué dans des communautés internationales du logiciel libre disposant parfois de plusieurs centaines de milliers d'utilisateurs. En 2017, l'organisme s'est doté d'un « *FLOSS evangelist* » bénéficiant d'un rôle consultatif pour les aspects de veille, stratégie et interactions avec les communautés libristes : il a en particulier pour mission de faire la pédagogie du logiciel libre en interne et auprès de ses partenaires, de produire une expertise dans ce domaine pour les activités de Renater et d'identifier, activer et développer des synergies avec des communautés libres existantes. Un « focus *open source* » a permis une obligation morale d'exemplarité, dans l'esprit de la loi pour une République numérique, ainsi qu'un impératif de faire le meilleur usage des fonds publics et ce en pensant au-delà de Renater en matière d'exploitabilité et réutilisabilité.

Ce volet de la stratégie de l'OSPO pourrait être évalué en identifiant les moyens investis au travers de la part de budget des ministères consacrée à l'action de référents logiciels libres des ministères ou à la mise en œuvre de leur politique open source.

2.2. Renforcer l'attractivité en valorisant les logiciels libres

Développer les formations techniques valorisant des outils libres issus de la recherche française

En 2020, la DINUM a signé un partenariat avec l'INRIA pour de l'échange d'expertise sur des sujets d'exploitation de la donnée publique et d'intelligence artificielle.

Récemment, l'INRIA a lancé INRIA Académie, un programme de formation autour de logiciels libres issues de ses projets de recherche, tous reconnus comme des projets à forte valeur ajoutée à l'international.

La DINUM et l'INRIA pourraient enrichir leur partenariat d'un volet « formation » à destination des agents publics pour qu'ils se forment aux outils libres indispensable au traitement de la donnée publique.

Au-delà de l'INRIA, la DINUM pourrait envisager d'autres partenariats avec des communautés dont l'activité de formation peut intéresser les agents publics, comme la communauté R[205], très active en France et dont les liens avec des agents publics de l'INSEE, entre autres, sont déjà anciens.

Imaginer des parcours de carrière pour des « champions » de l'open source

Debian est l'un des projets historiques du logiciel libre. C'est la distribution dont sont dérivées de nombreuses autres, dont Ubuntu. Être « Debian Leader » pour ce projet est une charge et un motif de fierté dans un univers du logiciel libre se réclamant par ailleurs fortement de la méritocratie.

Lorsqu'il devient critique pour elles de prendre part à des projets libres, les grandes entreprises tentent d'embaucher les développeurs les plus talentueux et influents de ces projets. L'Open Source Program Office pourrait avoir vocation à définir les projets libres critiques sur lesquelles il vaut la peine de recruter des talents, tout en travaillant avec la DGAFP pour la définition de carrières valorisantes pour ces personnes clefs.

[204] *Free Libre Open-Source Software*.

[205] Voir http://forums.cirad.fr/logiciel-R/ et https://github.com/frrrenchies/frrrenchies pour des exemples de ressources.

Attirer les jeunes talents soucieux de l'intérêt général et adhérents aux valeurs du logiciel libre

Le succès du programme Entrepreneur d'intérêt général (EIG) est un exemple de dispositif permettant d'attirer les jeunes talents de la donnée au service de l'intérêt général. La mission s'est interrogée sur la possibilité de créer un dispositif spécifique dans le domaine du code et du logiciel libre.

À titre de comparaison, le principe du « *Google Summer of Code* » est simple : Google paie des étudiants pour qu'ils contribuent à des projets open source pendant quelques mois d'été, ces étudiants étant suivis et accompagnés par des participants actifs de ces projets.

Il pourrait être envisagé de créer un « **BlueHats**[206] **Summer of Code** » où l'AGDAC s'engagerait à mobiliser des étudiants pour des stages bien rémunérés, étudiants qui contribueraient à résoudre un problème dans un logiciel libre utilisé par une administration. Pour ces étudiants, ce serait une première expérience d'implication dans un projet libre ; pour les administrations porteuses, une façon d'internaliser des compétences autour de ces logiciels qu'elles utilisent ; pour les projets libres, un apport supplémentaire de contributions.

Les logiciels libres comme composants stratégiques de nos systèmes d'information

Les infrastructures nécessaires à la donnée sont de plus en plus exposées à des formes de dépendances logicielles, ce qui soulève un enjeu d'autonomie stratégique. Les technologies web ont favorisé l'émergence du logiciel en tant que service, ou *Software as a Service* (SaaS) : au lieu d'exécuter un programme sur sa machine, l'utilisateur utilise son navigateur (ou un autre logiciel client) pour se connecter à une machine d'Internet sur laquelle s'exécute le logiciel rendant le service. Les organismes proposant des services peuvent assurer l'hébergement de ce service eux-mêmes (sur site, *on premises* en anglais), soit en déléguer tout ou partie à une entreprise.

Un organisme peut déléguer tous les aspects du service, auquel cas elle se tournera vers une offre SaaS ; il peut déléguer toutes les couches nécessaires au fonctionnement du service, sauf la couche applicative et les données, auquel cas elle utilisera une offre de *Platform as a Service* ; il peut enfin ne déléguer que les couches « profondes » et prendre à sa charge les applicatifs, les données, ainsi que le système d'exploitation et tous les services intermédiaires permettant que l'ensemble fonctionne, auquel cas on parlera d'*Infrastructure as a Service*[207]. De même que le serveur Web libre Apache Httpd était celui qui faisait tourner la majorité des sites web à la fin des années 2000, c'est avec des logiciels libres que tournent aujourd'hui la majorité des couches profondes de ces infrastructures, à commencer par le noyau Linux, omniprésent. **Il y a un fort enjeu d'autonomie stratégique dans la capacité à participer à la gouvernance de ces projets libres critiques, car ces services se développent majoritairement à l'étranger.**

[206] Les BlueHats désignent ceux qui se reconnaissent à la fois dans les valeurs du service public et du logiciel libre. La DINUM a initié le mouvement BlueHats lors du Paris Open Source Summit 2018 et elle continue de contribuer à ce mouvement en publiant une infolettre et en organisant des ateliers de partage d'expérience.

[207] https://www.bmc.com/blogs/saas-vs-paas-vs-iaas-whats-the-difference-and-how-to-choose/

De façon plus générale encore, les logiciels se construisent de nos jours à partir de codes sources existants, qu'ils s'agissent de cadriciels (*frameworks*) ou de bibliothèques de code – chaque *framework* ou bibliothèque pouvant à son tour dépendre d'autres cadriciels ou bibliothèques. La plateforme *code.etalab.gouv.fr* indique que les codes sources publiés par le secteur public dépendent de 1 738 bibliothèques[208]. La vaste majorité de ces bibliothèques sont libres, et forment un bien commun numérique que les acteurs privés et publics ont tous intérêt à maintenir. En 2014, le monde entier a découvert qu'une brique libre essentielle au bon fonctionnement du web, celle assurant la fiabilité des échanges sécurisés sur Internet, était maintenue par un individu sur son temps libre[209]. La Linux Fondation s'est emparée du sujet et a lancé la *Core Infrastructure Initiative*, une initiative pour inciter le plus d'acteurs possibles à prendre soin de cette infrastructure essentielle[210] que sont devenus les logiciels libres.

À titre illustratif de la présence des **logiciels libres dans les infrastructures du secteur public**, il est à noter l'implication déjà ancienne de l'Éducation nationale, avec des serveurs recourant à des logiciels libres dont *Red Hat*[211] assure le support[212]. Du côté des entreprises privées, citons l'exemple d'Airbus Defence & Space qui fait le choix d'une infrastructure reposant entièrement sur des logiciels libres[213].

[208] Il y en a en réalité beaucoup plus : ne sont comptées là, pour des raisons d'efficacité de la navigation, que les dépendances incluses dans au moins deux dépôts.

[209] https://fr.wikipedia.org/wiki/Heartbleed

[210] *Report on the 2020 FOSS Contributor Survey*, publié par The Linux Foundation & The Laboratory for Innovation Science at Harvard

[211] L'entreprise Red Hat a été rachetée par IBM en 2019 pour 34 milliards de dollars.

[212] https://www.generation-nt.com/linux-red-hat-education-migration-actualite-44840.html

[213] https://www.lemagit.fr/etude/Les-prochains-satellites-dAirbus-entierement-exploites-par-lOpen-source

3. Investir dans les infrastructures

Que l'on se situe dans le cadre de l'ouverture des données et des codes sources ou dans le domaine du partage et de l'accès dans un cadre plus restreint, la question des investissements est capitale quand on sait que l'accélération de la transformation numérique conduit le volume de données à augmenter de manière exponentielle.

3.1. Se doter d'infrastructures capables d'accueillir les données

Une notion recouvrant des acceptions plus ou moins larges

Selon une approche très large, elle peut être définie comme tout élément, matériel ou immatériel, qui en conditionne un autre et demande un traitement et une maintenance particulière. D'un point de vue plus technique, la notion d'infrastructure renvoie aux couches inférieures du système d'exploitation, allant du *hardware* à la mise en réseau des machines.

La définition d'infrastructures de données géographiques donnée par Rajabifard et al. en 2002[214] - à savoir des *« dispositifs qui rassemblent, dans un cadre dynamique,* ***les informations, les systèmes informatiques, les normes et standards, les accords organisationnels, les ressources humaines et les communautés nécessaires pour faciliter et coordonner l'accès et le partage des informations*** *géographiques »* - peut être généralisée.

Des besoins de stockage de plus en plus importants

Avec l'accélération de la transformation numérique, les masses de données augmentent et le stockage représente un poste important d'investissement. Par exemple, Météo-France prévoit qu'après la mise en œuvre de son « supercalculateur », les résultats de modélisation probabilistes représentent un volume de 6To de données par jour contre 1,5 actuellement avec un stockage des prévisions numériques du temps représentant un volume de 15 000 To à l'horizon 2025. L'augmentation du budget dédié représenterait 1,5 M€ par an.

Par ailleurs, **la constitution de « lacs de données »** est essentielle pour pouvoir stocker des mégadonnées et faciliter de nouveaux usages de la donnée, tel que le développement de l'intelligence artificielle. Cette notion a été développée par James Dixon, CTO de Penthao, en 2010, comme une solution pour le stockage de données sans prétraitement et sans connaître précisément l'usage future qu'il en sera fait. Ceci nécessite néanmoins des investissements importants.

Ces investissements importants doivent être donc d'autant plus coordonnés et pensés pour un usage le plus large possible. À titre d'illustration, la DGFiP a mis en place un lac de données grâce à un financement du FTAP fin 2019, pour un montant de 14,9 M€, et nécessitant un effectif important[215]. La mise en place de l'infrastructure devrait être opérationnelle fin 2020. Celle-ci doit accueillir toutes les données de la DGFiP afin de permettre des croisements dans leur format natif. Dans le même temps, la DGDDI a lancé son projet de lac de données, également financé par le FTAP (pour un montant de 18,8 M€ mais en incluant d'autres projets). Si la DGDDI indique qu'elle « étudiera à moyen terme la question d'une possible mutualisation avec la DGFiP », la mission considère que cette mutualisation devrait être davantage envisagée dès la conception du projet, voire constituer un critère de sélection des projets du FTAP.

[214] Rajabifard A., M.-E. Feeney, I. P. Williamson, 2002, *Future directions for SDI development.* International Journal of Applied Earth Observation and Geoinformation, 4, 1, pp. 11-22.

[215] Le projet nécessite 7 personnes en interne, ainsi que 4 prestataires pour la partie infrastructure et développement, et 1 personne en interne et 2,5 prestataires pour la partie dictionnaire.

Ces projets révèlent l'importance d'une cartographie et d'une documentation de la donnée, préalables à l'infrastructure : en général, les acteurs publics ont d'abord besoin de recenser leurs propres données et de les faire figurer dans des dictionnaires, indispensables à leur connaissance et à leur réutilisation.

Le développement du cloud*, un enjeu d'indépendance numérique*

Le *cloud* présente des avantages en matière de mutualisation des ressources. Un cloud permet de mettre à disposition d'un utilisateur plusieurs ressources informatiques (espace de stockage, puissance de calcul en particulier), sans que l'utilisateur ne doive les administrer directement (cf. encadré ci-dessous). Ces ressources informatiques sont réparties sur plusieurs zones géographiques et sont liées par un réseau. Le principe du *cloud* est donc d'être un service à la demande : le client doit pouvoir faire une demande à n'importe quel instant et doit pouvoir mesurer l'usage des ressources. Aussi, la mutualisation des ressources réside dans le fait que les hébergeurs du *cloud* partagent les ressources à l'ensemble des clients en fonction de la demande.

Les modèles de *cloud*

Il existe trois modèles de cloud : le cloud public (service partagé et mutualisé entre clients), le cloud privé (service administré par le client) et le cloud hybride (combinaison des cloud publics et privés).

Les trois catégories principales de cloud sont : l'IaaS, le PaaS et le SaaS.

- l'IaaS (*Infrastructure as a Service*) est le niveau de service le plus bas. L'IaaS consiste à donner accès à des services informatiques tels que des serveurs, réseaux et du stockage. Cet accès se fait au travers de machines virtuelles où l'utilisateur peut installer des applications. L'installation du système d'exploitation, des logiciels, des données sont sous la responsabilité du lient ;
- le PaaS (*Platform as a Service*) est un niveau au-dessus de l'IaaS. Le PaaS fournit en plus de l'infrastructure technique des composants logiciels comme le système d'exploitation. L'utilisateur doit donc juste gérer l'ajout d'applicatifs. Le fournisseur est responsable de l'infrastructure et de la sécurité ;
- le SaaS (*Software as a Service*) fournit des applications prêtes à l'emploi s'exécutant sur l'infrastructure du fournisseur cloud et accessible via le navigateur de l'utilisateur ou une API. Cette catégorie est par exemple utilisée pour l'accès à une messagerie en ligne, à Google doc ou à Office 365.

Source : Mission.

En 2012 face à l'utilisation de plus en plus importante de *clouds* en France, la CNIL a publié cinq recommandations[216] pour les entreprises utilisant ce type de service suite à une consultation publique[217] sur le *Cloud Computing,* à savoir :

- identifier clairement les données et les traitements passant dans le *cloud* ;
- définir les exigences de sécurité technique et juridique ;
- conduire une analyse de risques afin d'identifier les mesures de sécurité essentielles pour l'entreprise ;
- identifier le type de *cloud* pertinent pour le traitement envisagé (SaaS, PaaS, IaaS, public, privé, hybride) ;

[216] CNIL, Recommandations pour les entreprises qui envisagent de souscrire à des services de *Cloud computing*

[217] CNIL, Synthèse des réponses à la consultation publique sur le *Cloud computing* lancée par la CNIL d'octobre à décembre 2011 et analyse de la CNIL

– choisir un prestataire présentant des garanties suffisantes.

3.2. Des infrastructures adaptées aux différents usages de la donnée

Un prérequis à l'ouverture des données : créer des interconnexions pour faciliter la circulation des données

Chaque organisation, dès lors qu'elle produit, collecte ou utilise de la donnée, met en place une infrastructure plus ou moins complète, sans que cette mise en place soit toujours consciente d'ailleurs. Il en résulte généralement au sein d'une même organisation un morcellement des infrastructures selon les domaines (infrastructure pour les données de ressources humaines, une pour les données de gestion, une pour les données de production, etc) même si le degré de maturité des organisations vis-à-vis de la donnée est très variable.

La mission ne prétend pas avoir une vision consolidée des besoins mais a identifié quelques constats clés, à travers les auditions menées. Quelques exemples permettent d'illustrer la problématique liée au morcellement des infrastructures : celle-ci rend en effet difficile la circulation de la donnée et nécessite de concevoir l'architecture globale des systèmes d'information afin qu'ils puissent communiquer entre eux.

Dans le secteur de la santé, il y a peu de partage de données du point de vue des professionnels de santé. La feuille de route de la délégation du numérique en santé vise ainsi à créer le « code de l'urbanisme » autour de référentiels socles (éthique, sécurité, interopérabilité) et les services socles. Le rôle des pouvoirs publics est de faire en sorte que les services numériques proposés aux citoyens dans leur espace numérique en santé (parmi lesquels le dossier médical partagé, e-Prescription etc.) soient cohérents les uns avec les autres.

Dans le secteur médico-social, l'éclatement des systèmes d'information inhérent à la décentralisation de ces politiques, rend difficile la remontée des données : la construction de systèmes d'information constitue le prérequis d'une meilleure circulation des données. La loi d'adaptation de la société au vieillissement du 28 décembre 2015 confie à la CNSA la mise en place d'un système d'information commun des maisons départementales des personnes handicapées (MDPH), afin de contribuer à une plus grande performance et qualité du service rendu, à une harmonisation des pratiques et à une facilitation de pilotage national et local. Ce SI vise aussi à contribuer à améliorer la connaissance des personnes en situation de handicap et des réponses apportées. La consolidation de l'entrepôt de données devrait permettre d'obtenir des données relatives aux profils des données des personnes handicapées qui font des demandes aux MDPH, mais également d'obtenir les droits alloués par les CAF. Concernant le système d'information de l'autonomie liée à l'âge, le système d'information est encore plus balbutiant. L'agence du numérique en santé analyse les besoins des conseils départementaux pour instruire les droits des personnes sur les décisions d'attribution (les SI sont décentralisés). La CNSA est confrontée à l'impossibilité de connecter les données à un entrepôt de données relativement disparate.

En matière d'environnement, les infrastructures agrègent les données sur le territoire français et offrent une réponse à toutes les fonctionnalités attendues d'une infrastructure décrites plus haut. Elles représentent des points de vue thématiques sur l'écosystème national des données publiques. Elles peuvent interagir avec des infrastructures de territoires ou de thématiques plus réduits par la gouvernance, le financement, voire le réglementaire. Citons les 3 SI fédérateurs[218] de la direction de l'eau et de la biodiversité du ministère de la transition écologique, définis au sein des schémas nationaux des données sur l'eau, la biodiversité et le milieu marin où sont répertoriés les « SI métiers » avec lesquels elles interagissent.

[218] L'objectif principal d'un SI fédérateur est d'organiser, dans un domaine thématique (pour un milieu), la standardisation, la collecte, la préservation et la diffusion des données portant sur son domaine de manière à faciliter leur croisement et leur utilisation, y compris au-delà des missions de service publics qui président à leur création. Ils

Le projet de Géoplateforme de l'IGN, lauréat du fonds de transformation de l'action publique (FTAP) pour un montant de 3,6 M€, présenté par Valéria Faure-Muntian dans son rapport au Gouvernement sur les données géographiques souveraines[219], s'appuie sur une infrastructure fonctionnant selon le modèle d'une plateforme mutualisée (par opposition à une fédération de plateformes). Elle offre la possibilité aux communautés de créer leur infrastructure propre au sein de la plateforme en autonomie d'administration, profitant de briques logiciels communes, de facilités d'interopérabilité et d'une garantie de service et de sécurité.

> **Recommandation :** Orienter les investissements du plan de relance vers les infrastructures favorables à la circulation de la donnée (appels à projets de la DINUM et appels à projets sectoriels)

Data.gouv.fr*, une infrastructure à consolider pour l'ouverture des données*

La valorisation d'une donnée nécessite préalablement qu'elle soit découverte puis accessible et suffisamment documentée (notamment ses critères qualité) pour susciter sa réutilisation. Pour identifier et s'approprier une donnée, la comparaison des différentes infrastructures peut prendre du temps et nécessiter une expertise pour les réutilisateurs.

Conçu au départ comme un outil pour donner de la visibilité sur les données existantes et rassembler ces données sur un point d'entrée unique, *data.gouv.fr* est rapidement considéré comme une infrastructure, comme mentionné par Henri Verdier dans son rapport d'administrateur général des données de 2017, « *La donnée comme infrastructure essentielle* ».

La plateforme *data.gouv.fr* n'est cependant pas qu'une infrastructure et rassemble un certain nombre de services, dont l'API Geo et l'API Entreprises, les catalogues *api.gouv.fr*, *schema.data.gouv.fr*, entre autres.

Une assise réglementaire et des obligations : dans les dispositions du décret n° 2019-1088 du 25 octobre 2019 relatif au système d'information et de communication de l'État et à la direction interministérielle du numérique, data.gouv.fr est indiqué comme « le portail unique interministériel destiné à rassembler et à mettre à disposition librement l'ensemble des informations publiques de l'État, de ses établissements publics et, si elles le souhaitent, des collectivités territoriales et des personnes de droit public ou de droit privé chargées d'une mission de service public ».

Dans les dispositions de décret n° 2017-331 relatif au service public de mise à disposition des données de référence[220], *data.gouv.fr* apparaît comme le portail de mise à disposition des bases de données référencées par le service public des données de référence (SPD) lorsque l'administration productrice d'une base de données de référence ne peut ou ne souhaite pas diffuser celle-ci selon les obligations de qualité de service précisées dans un arrêté du Premier ministre[221].

Ces obligations réglementaires visent à garantir une pérennité et une qualité de service du portail *data.gouv.fr*.

---

permettent de valoriser ces données à la fois pour piloter les politiques publiques par la donnée mais aussi pour répondre aux enjeux de société dans leur domaine. Ils permettent de standardiser les données produites en s'appuyant sur des référentiels communs, de les qualifier et de les concentrer, ce qui facilite la mise en place de solutions performantes (par ex. API) pour les diffuser en opendata dans la logique « FAIR » (Facile à trouver ; Accessible ; Interopérable ; Réutilisable) et par exemple qu'une donnée puisse être utilisée dans le cadre de plusieurs politiques publiques ;

[219] Rapport au Gouvernement de Valéria Faure-Muntian, *Les données géographiques souveraines*, juillet 2018.

[220] Ces dispositions sont codifiées dans le CRPA aux articles R.321-7 et R.321-8

[221] Arrêté du 14 juin 2017 relatif aux règles techniques et d'organisation de mise à disposition des données de référence prévues à l'article L. 321-4 du code des relations entre le public et l'administration. Les règles de publication des données de référence comprennent un liste de métadonnées, une fréquence maximale de mise à jour, des règles de disponibilité et de niveau de performance du service, des modalités de mise à disposition des données, une procédure de signalement au producteur de données et une contrainte de délai d'information des usagers pour toute modification sur la donnée ou sa mise à disposition.

L'utilité du service public de la donnée de référence, qui impose à ces données des critères qualité sur leur diffusion et de mise à jour, est unanimement reconnue par les acteurs de la donnée, qu'ils soient publics ou privés. Toutefois seuls 9 jeux de données de référence y figurent : depuis sa création, aucun autre jeu de donnée l'a rejoint.

De par sa conception, l'« infrastructure *data.gouv.fr* » a offert dès sa création des services pour rassembler l'ensemble des acteurs afin de devenir ce point central de découverte des données publiques. Pour les infrastructures existantes possédant leurs propres catalogues de données, data.gouv.fr offre des services de moissonnage dans une **vision fédératrice** (74% des jeux de données proviennent de moissonnage, 40% des collectivités possédant une infrastructure sont régulièrement moissonnées). En parallèle il propose aux administrations et notamment aux collectivités qui ne peuvent se permettre d'investir dans de la compétence et des technologies, des services simples de dépôt de données, de documentation et de mise en relation avec les réutilisateurs, dans une **vision d'infrastructure communautaire**.

Autour de cette vision, la DINUM a développé des services pour faciliter l'exposition et l'appropriation des données. *Geo-datagouv* permet de moissonner les catalogues compatibles avec les principes de la directive INSPIRE, essentiellement des infrastructures de données géographiques. *Api.gouv.fr* propose un catalogue des API des administrations. *Schéma.gouv.fr* offre la possibilité aux administrations de référencer les schémas des bases de données que *data.gouv.fr* relaie dans une logique de qualité vertueuse d'adéquation des données au schéma qui les structure pour en faciliter l'appropriation et en permettre l'agrégation.

Enfin, l'expérience de transport.data.gouv.fr démontre la faisabilité de bénéficier des briques de service de l'infrastructure (« le back-end ») dans une approche décentralisée. C'est une approche mise en avant par Etalab.

Si on reprend le besoin des utilisateurs de données : « comment puis-je trouver et m'approprier très rapidement la donnée dont j'ai besoin ou découvrir celle qui pourra me servir, au plus proche de son producteur pour être certain de sa légitimité ? », la vision fédératrice et communautaire de data.gouv.fr paraît adaptée à cet enjeu puisqu'elle vise à entretenir un point « unique » (et non « central ») à la fois dans le respect de la subsidiarité pour les infrastructures existantes et dans l'accompagnement des producteurs qui ne disposent pas de telles infrastructures.

Toutefois l'offre de service de data.gouv.fr ne suffit pas aux besoins de découverte. Ce point unique ne doit pas se contenter de ne cataloguer que les données ouvertes, mais il doit aussi rendre compte des données partageables sous conditions (ce qui figurait dans le plan annoncé par l'Administrateur général des données dans son rapport au Premier ministre), les données accessibles ou consultables voire des données fermées des administrations si elles le souhaitent. Savoir qu'une donnée existe et pour quelle raison elle n'est pas publiée fait gagner le temps d'une recherche vaine (on peut penser par exemple aux bases de données des services statistiques ministériels couvertes par le secret statistique).

Le service *api.gouv.fr* permet quant à lui de référencer des API dont l'accès est restreint aux administrations habilitées. La multiplication des jeux de données (plus de 35000 à ce jour) impose de faire monter le moteur de recherche en échelle et de répondre au besoin de simplicité et de pertinence attendu des usagers. Le travail éditorial est d'autant amplifié.

De même la multiplication des accès (450 millions de requêtes par semaine en novembre 2020 sur *adresse.data.gouv* pour 3,7 millions de visiteurs uniques) impose de dimensionner correctement l'infrastructure socle au regard de ses obligations (SPD et autres bases de données socles hébergées).

Par ailleurs, si on prend l'exemple du service *Geo.data.gouv.fr,* celui-ci n'est plus maintenu et des plateformes régionales d'informations géographiques compatibles avec les principes INSPIRE ne sont actuellement plus moissonnées. Il conviendrait d'organiser la remontée des informations en lien avec le ministère de la transition écologique en charge de la mise en œuvre de la directive INSPIRE, et avec le bureau de recherches géologiques et minières (BRGM) qui entretient le géocatalogue, point focal des données géographiques entrant dans le champ de la directive en charge d'alimenter le rapportage européen annuel de la France.

Il existe des chantiers ouverts à Etalab mais cela demande des ressources : s'intéresser aux données « sources » pour permettre de rejouer des algorithmes, travailler sur la montée en qualité des données exposées (à l'image par exemple des investigations et rapports publiés par le SANDRE sur les données du SI eau[222]), améliorer la visibilité des contenus, notamment en investissant dans l'éditorialisation : le SPD est le haut de la pyramide par l'intérêt qu'il suscite. Les espaces de données sectorielles intéressent les communautés qui les animent. Pour les autres données, l'intérêt reste à susciter.

Favoriser l'émergence de plateformes de partage et/ou d'accès mutualisées

Plusieurs plateformes existent dont la mission considère qu'elles méritent d'être promues en temps qu'outils mutualisés facilitant le partage ou l'accès (cf. partie 3). Elles se distinguent par les exigences en matière de sécurité d'accès à la donnée (données en partage ou en accès restreint) mais aussi sur leur finalité : offrir un accès pérenne à un catalogue de données ou permettre un accès ponctuel (le temps d'un projet de recherche par exemple) à un ou plusieurs jeux de données.

Loin de prétendre à l'exhaustivité, la mission a identifié, à l'occasion de ses auditions différentes infrastructures répondant à ce besoin.

En premier lieu, le Réseau Quetelet PROGEDO diffusion est le portail français d'accès aux données dans le domaine de la recherche en sciences humaines et sociales, qui permet aux chercheurs français et étrangers d'obtenir des bases de données anonymisées nécessaires à leurs traitements, notamment des fichiers issus des grandes enquêtes, recensements et autres bases de données provenant de la statistique publique française. Passent par ce canal de mise à disposition gratuite, les fichiers confidentiels (dits fichiers de production et de recherche) dont le niveau de détail a été réduit, afin de diminuer de manière importante le risque de ré-identification. Une procédure allégée a été mise en place en 2018, afin de permettre un accès plus rapide des chercheurs aux données.

Le Centre de Données et Services ESPRI est, quant à lui, un service transverse de la fédération des laboratoires Institut Pierre Simon Laplace (IPSL) et l'un des 4 centres de données de l'infrastructure nationale de données et de services pour l'atmosphère, AERIS, mise en place en 2014. Il est l'infrastructure de partage des données dans le cadre du datahub Energy for climate (E4C) en matière de recherche sur l'énergie et le climat (cf. partie 3).

L''infrastructure de données géographiques Sextant[223] développée par l'Ifremer a permis à l'établissement de prendre le leadership de projets européens en hébergeant ces projets dans une infrastructure multi-thématiques, multi-partenaires et multi-projets, conforme aux normes ISO et OGC.

Enfin, le centre d'accès sécurisé distant aux données (CASD) se distingue des solutions qui précèdent par trois facteurs :

- **il propose un équipement conçu pour permettre aux chercheurs ou statisticiens de travailler, dans des conditions de sécurité élevées** (accès via un boitier sécurisé par une empreinte digitale et uniquement sur place au CASD), sur des données confidentielles ; historiquement issues de la statistique publique, elles sont soumises à des règles de confidentialité renforcées en raison de leur niveau de détail plus fin que les fichiers présents sur le réseau Quetelet ;
- **il répond à un besoin ponctuel d'accès à la donnée** et ne dispose pas d'un catalogue de données mobilisables en un instant t mais plutôt d'une liste des jeux de données ayant été rendus accessibles par son intermédiaire ;
- **il est intersectoriel : bien que son ADN soit statistique, le CASD permet l'accès à des données de nature administrative qui représentent aujourd'hui 20 % des jeux de données mis à disposition** (70 % de nature statistique ou administrative ayant fait l'objet d'un retraitement et 10 % de nature privée).

[222] http://mdm.sandre.eaufrance.fr/geo/rapportsv3

[223] https://sextant.ifremer.fr/Presentation

Il convient de noter le développement d'une offre « CASD Light », permettant un accès sans boitier d'accès, qui s'apparente aux services proposés actuellement par les plateformes sectorielles : cette offre a cependant vocation à être utilisée pour des projets ponctuels.

Face aux besoins de partage et d'accès à la donnée tant dans le secteur public que dans le secteur privé, les besoins sont amenés à croître de manière très forte ; la réponse à ces besoins combinera nécessairement une offre publique du même type que le CASD (y compris l'offre « CASD light ») et une offre privée, ce qui laisse entrevoir l'utilité d'un encadrement public des tiers de confiance privés agissant comme *« datatrust »*[224]. Ce modèle mérite d'être exploré et expertisé, en vue de poser, le cas échéant, un cadre juridique pour cette pratique.

3.3. Jouer sur différents leviers pour investir dans les infrastructures

Compte tenu des besoins importants en matière d'infrastructure, il apparaît essentiel de pouvoir mobiliser les différents leviers de financement. Sans viser l'exhaustivité, la mission a identifié différentes pistes de financement pouvant contribuer à cet effort d'investissement dans les infrastructures qui représentent un coût important.

Le fond de transformation de l'action publique (FTAP) constitue un premier levier de financement pour le renforcement des infrastructures. Par exemple, le lac de données de la DGFiP a pu être mis en place grâce à un financement du FTAP fin 2019. Le projet de Géoplateforme de l'IGN a également pu bénéficier de ce mode de financement pour un montant de 3,6 M€.

Par ailleurs, le programme d'investissements d'avenir (PIA), piloté par le secrétariat général pour l'investissement (SGPI), a été mis en place par l'État pour financer des investissements innovants et prometteurs sur le territoire, afin de permettre à la France d'augmenter son potentiel de croissance et d'emplois. Il offre un levier d'investissement supplémentaire aux besoins en infrastructures.

Ainsi, l'Agence nationale de la recherche (ANR) a lancé un appel à manifestations d'intérêt pour le financement d'équipements d'excellence (EquipEx+) pour la recherche scientifique dans le cadre de l'action « Equipements structurants pour la recherche » du troisième programme d'investissements d'avenir (PIA3) : celui-ci vise à doter la recherche française d'équipements de haut niveau afin de lui permettre d'accueillir les expérimentations conduites par les chercheurs nationaux comme internationaux dans les meilleures conditions de travail et aux plus hauts standards. Cette démarche intervient après le PIA 1 qui a mené un important effort pour les équipements intermédiaires dans la recherche scientifique et le PIA 2 qui s'est concentré sur un nombre plus limité d'équipements structurants.

À titre d'illustration, le CASD a été retenu comme équipement d'excellence dans le cadre PIA 1 en février 2010, après une phase pilote mise en œuvre conjointement par le GENES et l'INSEE en 2009. À ce titre, le projet Equipex CASD, porté par le GENES en tant que coordinateur, est bénéficiaire d'un co-financement géré par l'ANR jusqu'à la fin de l'année 2019.

L'appel à projets « accompagnement et transformation des filières » de Bpifrance s'inscrit également dans le cadre du programme d'investissement d'avenir. L'action a vocation à renforcer la compétitivité des filières stratégiques françaises par l'innovation, en permettant notamment le recours à des outils de partage de données destinées au développement de solutions d'intelligence artificielle. Les plateformes numériques de filière ont vocation à mutualiser au sein des filières des outils numériques contribuant à la compétitivité des acteurs de la filière et à la performance des échanges entre ces acteurs. Les initiatives de mutualisation et de partage des données destinées au développement de solutions d'intelligence artificielle pourront viser la constitution, la mise à disposition et l'exploitation de nouvelles bases de données mutualisées.

Par ailleurs, il est à noter un effort d'investissement considérable dans le secteur de la santé à la suite des difficultés constatées dans la remontée des données dans le cadre de la crise de la Covid19 (cf. cas d'usages des données et modèles épidémiologiques dans le cadre de la crise de la Covid19) : le Ségur de la santé a ainsi consacré 2 Mds€ d'investissements dans le numérique en santé.

[224] Personnes morales publiques ou privées qui mettent en commun leurs données dans un cadre de gouvernance préétabli à l'avance.

Plus globalement, le plan de relance du gouvernement prévoit 500 M€ d'investissements sur le volet numérique, dont 208 M€ pour développer les outils numériques des agents, 204 M€ pour la transformation numérique de l'État, et 88 M€ à destination des collectivités territoriales. Ces financements offrent une opportunité inédite pour investir dans les infrastructures nécessaires à la collecte et à la circulation de la donnée et des codes sources.

Dans le cadre du plan de relance, la DINUM porte plus particulièrement une mesure concernant *le « soutien à l'innovation et à la transformation numérique de l'Etat et des territoires »* qui s'articule autour de sept thèmes, parmi lesquels certains visent à *« développer la collaboration des administrations avec leurs écosystèmes grâce au numérique »*, *« développer l'usage de la donnée au service de l'action publique (IA, décisionnel, échange de données, open data, archivage) »* ou encore à *« accélérer des projets de transformation nécessitant l'activation de leviers multiples et une mobilisation financière importante dont la dimension numérique apparaît structurante ».*

> **Recommandation :** Orienter les investissements du plan de relance vers les infrastructures favorables à la circulation de la donnée (appels à projets de la DINUM et appels à projets sectoriels)

3.4. Le besoin d'une autonomie stratégique européenne plus forte

Au-delà des règles protégeant l'usage des données personnelles, ce sont également les outils et les infrastructures sur lesquels ces données transitent, et les services dont elles sont l'objet, qui constituent un paramètre majeur pour susciter la confiance des citoyens dans le développement des nouvelles technologies et de l'IA en particulier.

À ce titre, plusieurs sujets appellent une réponse harmonisée dans le cadre du projet de règlement « *Data Governance Act* » (DGA) comme :

- la localisation en Europe de certaines catégories de données sensibles (personnelles et non personnelles) détenues par des entités du secteur public, et leur non exposition à des lois extraterritoriales en l'absence d'accord bilatéral autorisant leur transfert vers des pays tiers ;
- la régulation des services de partage de données au sein de l'Union européenne : les tiers de confiance qui mettent en relation des « producteurs » de données et des utilisateurs devraient respecter des règles de neutralité et d'absence de conflit d'intérêt et être ainsi contraints à séparer leurs activités de *data brokerage* de celles de commercialisation de leurs propres services de traitement de données.

La stratégie européenne en matière d'économie de la donnée, annoncée le 19 février 2020, vise à renforcer l'autonomie stratégique de l'Europe sur ses données avec :

- un cadre général d'autorisation pour les fournisseurs de services de partage de données s'adressant aux utilisateurs professionnels ou aux personnes concernées (exemple : les *Data Brokers* ou places de marché de la donnée), afin d'accroître la confiance dans la fourniture de services de partage de données au sein de l'Union et d'éviter la fragmentation du marché intérieur ;
- un cadre général d'autorisation pour la collecte et le traitement de données mises à disposition à des fins altruistes (« altruisme numérique ») par des personnes ou des entreprises (exemple : ses données de santé après la mort).

Le DGA constitue la première pierre réglementaire de la stratégie européenne en matière d'économie de la donnée, annoncée le 19 février 2020, qui vise à renforcer l'autonomie stratégique de l'Europe sur ses données. Il devrait être suivi, dans la deuxième partie de l'année 2021, par un *Data Act* (qui devrait aborder les questions de partage/accès à des données dans certains secteurs comme le transport, de propriété et de portabilité des données).

Le projet de cloud européen GAIA-X constitue une initiative franco-allemande à dimension européenne intéressante, qui doit permettre de développer une infrastructure de données pour l'Union européenne, officiellement lancé le 4 juin 2020.

Au niveau national, la stratégie de l'État en matière d'informatique en nuages repose sur les trois niveaux de *cloud* définis pour permettre de concilier sécurité des données et accélération des usages. Cette stratégie est définie par la circulaire du Premier ministre en date du 8 novembre 2018 sur la doctrine d'utilisation de l'informatique en nuage par l'État.

Plusieurs travaux parlementaires ont alimenté la réflexion sur cette problématique, notamment la commission d'enquête sur la souveraineté numérique du Sénat en 2019, et plus anciennement les travaux de la sénatrice Catherine Morin-Desailly (par son rapport intitulé « L'Union européenne, colonie du monde numérique ? »). Une mission d'information de l'Assemblée nationale est également en cours, à date (« Bâtir et promouvoir une souveraineté numérique nationale et européenne », présidée par Jean-Luc Warsmann).

CAS D'USAGE – Trois succès de logiciels libres d'information géographique

Le développement des logiciels libres par des acteurs publics offre des opportunités que la mission a choisi d'illustrer par trois exemples emblématiques dans le domaine de l'information géographique : QGIS, Prodige, Geotrek.

Geotrek

En 2011, les parcs nationaux des Ecrins et du Mercantour se sont associés pour développer une application web de gestion et de valorisation de leurs sentiers avec un budget de 70 000 €. Ils ont retenu des technologies libres et une approche générique pour réaliser un outil qui pourrait servir à d'autres structures, et choisi de publier l'application sous licence libre pour en faciliter la réutilisation. Avec le prestataire, ils ont aussi pris soin de la documenter et de communiquer sur celle-ci sous le nom de Geotrek pour la faire connaître et la partager.

Huit ans après sa première version, l'application est utilisée par plus d'une centaine de structures en France et commence à être déployée à l'étranger. Au-delà du partage de l'outil Geotrek, ce projet a permis de fédérer des projets de différents territoires et de grouper des financements, pour mieux gérer et valoriser leurs territoire, leurs randonnées et leur patrimoine, et de mieux partager et diffuser ces données à différents niveaux (portails touristiques locaux, systèmes d'information touristique régionaux, OSM, IGN, *data.gouv.fr*).

QGIS

QGIS est un logiciel qui permet de mettre en place un SIG (système d'information géographique), c'est-à-dire un système d'information conçu pour collecter, conserver, traiter, analyser, gérer et présenter tous types de données spatiales et géographiques. En 2010, une enquête est faite au sein du ministère de l'écologie : l'usage des SIG libres (QGIS, GVSIG, etc.) est confidentiel, avec 80 utilisateurs contre 5200 licences de SIG propriétaires. Quelques sites pilotes participent dès 2010 à une utilisation expérimentale de QGIS. Cette période est utilisée par le ministère pour bien comprendre le fonctionnement des projets libres, notamment leur « résilience » et l'importance de contribuer en retour.

En 2012, la circulaire du Premier ministre sur l'usage des logiciels libres dans l'administration (dite « circulaire Ayrault »)[225] agit comme catalyseur en insistant sur les économies possibles grâce aux logiciels libres et le fait qu'ils facilitent l'expérimentation et l'adaptation. À partir de 2014, le ministère de l'écologie s'engage sur un déploiement plus massif de QGIS dans les services en visant à terme le remplacement de l'ancien logiciel propriétaire, avec un dispositif de formation. Désormais, la quasi-totalité des utilisateurs a basculé sur QGIS, logiciel techniquement évolué et au moins aussi puissant que l'ancien logiciel propriétaire.

Prodige

Préexistant à la directive européenne Inspire de 2007, Prodige a été développé pour gérer les échanges entre les services de l'État et les collectivités. Sa première version date de 2007. L'ADAE (Agence pour le développement de l'administration électronique, service prédécesseur de la DINUM) avait lancé et mutualisé plusieurs initiatives pour développer l'usage de l'information géographique, dont Prodige initialisé en Région Rhône-Alpes. Ces activités cartographiques ont été transmises au ministère de l'écologie, dont la reprise de Prodige a permis une généralisation aux

[225] Circulaire du Premier ministre du 19 septembre 2019 relative aux orientations pour l'usage des logiciels libres dans l'administration.

régions et départements. L'ADAE a également été à l'origine des applications Géocatalogue et Géoportail.

Aujourd'hui, les principaux utilisateurs de Prodige sont les services de l'État, les collectivités territoriales, et différents opérateurs publics. Il est en particulier utilisé par le Cerema pour sa plateforme *open data* « CeremaData », et par les ARS pour réaliser atlaSanté, projet mutualisé des agences régionales de santé et du ministère de la santé depuis 2013. Plus de la moitié des régions utilise à l'heure actuelle une plateforme Prodige. En outre, certaines plateformes Prodige sont déployées au niveau d'un département ou sur un territoire particulier (et non pas de la région).

Un vecteur puissant de mutualisation

Ces trois projets ont mis en évidence les gains pour les acteurs publics, en matière de mutualisation des ressources, permettant de réduire les coûts de développement, voire à certains acteurs d'accéder à des services qu'ils ne pourraient pas développer seuls. L'exemple de Géotrek illustre bien cet enjeu.

Des structures comme des parcs nationaux et régionaux, des départements, des communautés de communes, des comités du tourisme et de la randonnée ont ainsi déployé Geotrek, facilité par le besoin commun, la généricité et l'absence de licences payantes. Ils ont bénéficié du travail de conception et des développements initiaux sans les repayer et ont ainsi pu concentrer leurs ressources sur le développement de nouvelles fonctionnalités et modules, bénéficiant à leur tour au collectif.

Aujourd'hui, plus de 100 structures utilisent cet outil dont le coût total est estimé à environ de 1,5 M€[226]. Au-delà du coût mutualisé de l'outil, il y a aussi de nombreux gains et économies indirects (partage de ressources techniques, de méthodologie, projets communs, partage d'expérience, offres de service au territoires que l'on aurait pu réaliser chacun de notre côté, partage et ouverture des données, etc.).

Les évolutions de Géotrek reposent sur des financements de multiples parcs nationaux, avec une coordination technique organisée en groupes de travail, ainsi qu'un comité de pilotage pour la dynamique générale du projet. Un groupement de commande a aussi été mis en place pour mutualiser le financement de certains développements communs.

Une participation nécessaire à la communauté de développeurs

Recourir à des outils libres nécessite de trouver une gouvernance adaptée et d'identifier la communauté propre à maintenir la dynamique de l'outil dans le temps. Pour réussir, il est essentiel qu'un écosystème complet se mette en place autour de l'outil (sociétés de services, communauté d'utilisateur). C'est pourquoi la mission propose de mettre en place un programme permettant d'accompagner l'émergence de ces écosystèmes au sein de l'État.

S'agissant de Geotrek, le Parc national des Ecrins anime cette communauté des utilisateurs qui se réunit chaque année pour faire avancer le projet de manière cohérente et collective. Néanmoins, l'animation du projet concentrée sur le Parc national des Ecrins comporte des risques. S'il existe un fort intérêt des utilisateurs pour cette démarche de partage et d'ouverture de l'outil, il y a un faible investissement humain des différents utilisateurs au-delà des développements qu'ils financent. La pérennisation des ressources pour assurer l'animation du projet et sa maintenance technique constitue un enjeu essentiel.

S'agissant de QGIS, en juin 2013, le ministère de l'écologie comptait une communauté interne active, avec plus de 350 abonnés à la liste de discussion, attirés par la gratuité de l'outil, la possibilité d'utilisations avancées et le développement d'applications métiers. Lors du déploiement au sein des

[226] A partir des estimations proposées sur https://www.openhub.net/p?query=geotrek

services en 2014, le ministère de l'écologie s'implique dans la feuille de route du projet et participe aux échanges *via* la forge QGIS et la communauté des développeurs.

Enfin, dans le cas de Prodige, le ministère de l'écologie prend en charge l'assistance à maîtrise d'ouvrage « coordinatrice » du projet, par le biais de l'Ecolab (service du Commissariat général au développement durable) en lien avec le ministère de la santé. Les adaptations liées à la conformité à la directive Inspire ou à l'urbanisation avec d'autres plateformes, comme la future GéoPlateforme de l'IGN, sont financées par les crédits du ministère, tandis que les besoins spécifiques sont financés par les acteurs qui les demandent.

Prodige permet de valoriser les données des collectivités, notamment des régions, et de garantir les partenariats avec une méthodologie et mise en œuvre commune, en particulier pour le bon respect des obligations de la directive Inspire. Chaque région utilisant Prodige déploie une instance, et les partenaires locaux, y compris des départements, peuvent utiliser la plateforme de leur territoire. Prodige a été enrichi de modules développés par des acteurs territoriaux, comme par exemple le module "Base communale". Cette coordination des différents développements a permis une meilleure écoute des besoins des utilisateurs.

Un enjeu d'indépendance vis-à-vis de prestataires informatiques

Le fait de ne pas dépendre d'un seul prestataire constitue un enjeu important pour ce type de projet. Plusieurs facteurs peuvent réduire cette dépendance : la mise à disposition d'une documentation du projet, le partage des ressources techniques, le recours à une pluralité de prestataires dans le cadre des marchés publics, quand cela est possible, le groupement des financements pour permettre à d'autres prestataires d'investir du temps dans la découverte et prise en main de l'outil, et enfin la diversification des services attendus autour de l'outil (hébergement, installation, customisation, développements, etc.).

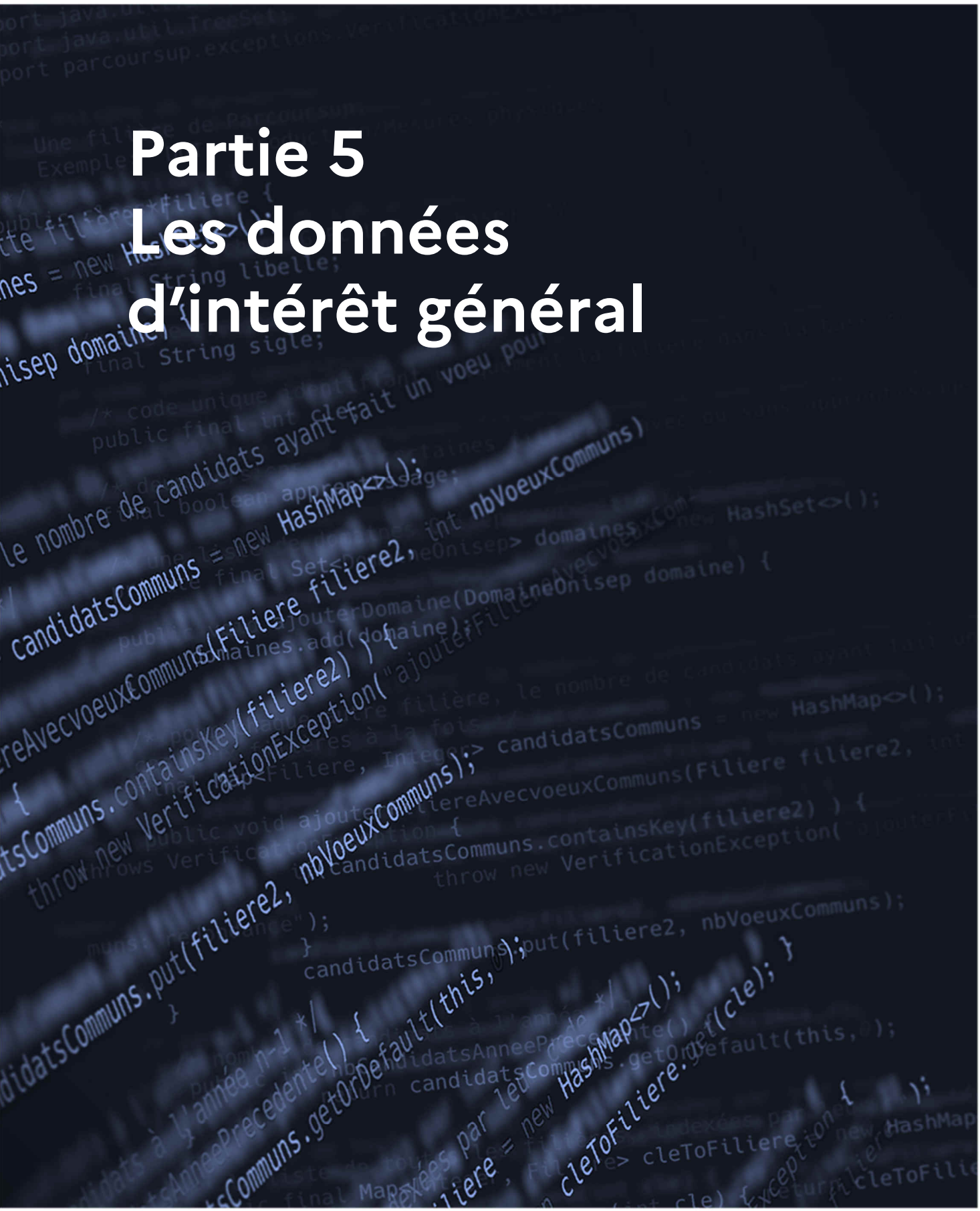
Pour Prodige, plusieurs SSII ont été prestataires au fil des ans pour des développements. La licence CeCILL a permis des contributions extérieures, l'accord-cadre de développement comportant un lot permettant des développements communs avec le logiciel Carmen, porté par le BRGM. Le visualiseur cartographique actuel est commun à Prodige et Carmen, résultat d'un effort partagé entre le ministère de l'écologie, le ministère de la santé et le BRGM. Prodige intègre également des composants libres majeurs comme le logiciel de catalogage GeoNetwork, utilisé aussi par les autres plateformes françaises principales et à l'international.

Une adaptation et une acculturation nécessaire des acteurs publics

Le développement de logiciels libres par des acteurs publics doit encore faire l'objet d'une acculturation des agents.

D'une part, l'utilisation des outils de la commande publique est parfois complexe pour tenir compte des besoins des logiciels libres. Par exemple, dans le cas de Prodige, des améliorations non structurantes du logiciel, ou bien qui ne sont pas encore clairement connues dans leur spécification technique, peuvent difficilement faire l'objet d'un marché public, outil plutôt conçu pour commander un logiciel entier ou bien une fonctionnalité bien identifiée *a priori*. Ce besoin d'adaptation est aujourd'hui pris en compte dans les travaux conduits par la direction des affaires juridiques des ministères économiques et financiers et l'agence du patrimoine immobilier de l'État (APIE) sur les cahiers des clauses administratives générales applicables aux marchés publics de techniques de l'information et de la communication (CCAG-TIC).

D'autre part, l'administration privilégie, pour des motifs budgétaires, le recours à des sociétés sélectionnées par appels d'offres, au recrutement de développeurs, lorsqu'elle souhaite des évolutions en matière de logiciels libres. Pour QGIS, il y a toutefois des contributeurs importants de QGIS dans la société titulaire du marché public. Par ailleurs, le dispositif d'accompagnement a été dès le départ identifié comme un point critique pour l'adoption du logiciel QGIS par les agents. Un dispositif très complet de formation, en particulier en distanciel (FOAD) a été mis en place.



Partie 5
Les données d'intérêt général

1. Une notion imprécise et dont la traduction juridique manque de cohérence

1.1. Les données d'intérêt général : un concept récent encore mal défini d'un point de vue juridique

Plusieurs tentatives de définition mais pas de traduction juridique

La notion de données d'intérêt général sonne comme une évidence : des données permettent de décrire, d'éclairer voire d'agir pour l'intérêt général, même si elles ont été produites ou collectées par la sphère privée. D'ailleurs, la loi du 17 juillet 1978 sur l'accès aux documents administratifs prévoit dès l'origine que « les documents administratifs sont de plein droit communicable aux personnes qui en font la demande, qu'ils émanent des administrations de l'État, des collectivités territoriales, des établissements publics ou des organismes, fussent-ils de droit privé, chargés de la gestion d'un service public »[227]. Le caractère privé du détenteur ne suffit donc déjà pas en soi à faire obstacle à l'ouverture des donnés.

Pour autant, les auditions conduites par la mission n'ont pas permis de mettre en évidence l'existence d'une définition consensuelle et stabilisée des données d'intérêt général. Cette absence de définition est également corroborée par l'étude de la doctrine et des dispositions légales en vigueur sur les données. Bien souvent, les tentatives de définition relèvent de la tautologie : sont d'intérêt général les données privées dont l'ouverture correspond à un motif d'intérêt général, sans que l'intérêt général en question soit précisément caractérisé ni qu'une méthode pour l'identifier soit établie, au-delà du renvoi à des secteurs d'intervention légitime de la puissance publique. Il s'agit là d'une difficulté importante pour toute tentative de traitement unifié du sujet, aussi bien sur le plan juridique que pratique. Toutefois, plusieurs distinctions conceptuelles peuvent être utilement mobilisées pour tenter de clarifier le débat.

Le rapport Jutand de mars 2015 sur l'ouverture des données de transport[228] est l'un des premiers à proposer d'employer la notion d'intérêt général pour guider la réflexion sur l'ouverture des données. Se heurtant à la difficulté de définir le statut juridique de certains services de mobilité (autopartage, libre-service, etc.) dont les données pourraient néanmoins intéresser la puissance publique et le citoyen, il met en évidence les insuffisances d'une approche par le périmètre strict des missions de service public et en propose deux élargissements. D'une part, il suggère de retenir une piste organique, « consistant à tenir compte de la participation d'une personne publique à la mise en place, au fonctionnement ou au financement du service considéré ». D'autre part, il propose une piste téléologique qui appréhende les données selon leur finalité, via l'introduction d'une notion « d'information d'intérêt général » (IIG) qui permet d'englober « toutes les données de l'information transport, en fonction des évolutions éventuelles ».

[227] Article 2 de la loi n° 78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal

[228] Francis JUTAND (dir.), « Ouverture des données de transport », *Rapport remis au secrétaire d'État chargé des Transports de la Mer et de la Pêche*, mars 2015, 146 p..

De son côté, le rapport relatif aux données d'intérêt général de septembre 2015[229] identifie des données d'opérateurs privés dont la puissance publique a déjà imposé l'ouverture pour mettre fin à des abus de position dominante[230] ou pour les besoins de certaines politiques publiques. Prolongeant la classification du rapport Jutand, il distingue les données d'intérêt général produites ou contrôlées par une personne privée qui entretient un lien avec la puissance publique (concessionnaire, délégataire, bénéficiaire d'une subvention), des données qui sont d'intérêt général « par nature ». Pour cette deuxième catégorie, il met en avant quatre types de motifs d'intérêt général reconnus par la jurisprudence constitutionnelle et qui peuvent fonder une obligation de communication des données par la loi : l'optimisation de la conduite des politiques sectorielles, l'information du citoyen, la recherche scientifique et le développement économique.

Toutefois, il relève que l'élaboration d'une loi générale d'ouverture des données privées pour des motifs d'intérêt général échouerait à définir ces motifs de façon suffisamment précise pour être normative et qu'elle présenterait dès lors un risque d'inconstitutionnalité et d'inconventionnalité. En conséquence, il recommande de privilégier une approche sectorielle, jugée mieux à même de circonscrire les atteintes portées à la liberté d'entreprendre, au droit de propriété et au secret des affaires, et permettant de mieux les proportionner aux motifs d'intérêt général visés.

C'est une telle démarche qui a été retenue dans un rapport complémentaire de mars 2016[231], lequel préconisait une ouverture de jeux de données précis et ciblés dans différents secteurs (mobilité, emploi et formation, logement et foncier, finance, énergie), en faisait la cartographie, l'exposé des motifs, l'analyse juridique et l'étude d'impact, et proposait des modalités de mise en œuvre pratique. Si les exemples sont convaincants, il convient cependant de relever que les données identifiées par le rapport, « dont l'ouverture est à la fois juridiquement possible et significative en matière de potentiel économique », sont principalement des données de la sphère publique et parapublique, à rebours de l'orientation initiale qui se voulait centrée sur les données des acteurs privés, ce qui traduit bien les difficultés à manipuler le concept de données d'intérêt général.

Sur le plan de sa traduction législative, la loi pour une République numérique se contente de reprendre l'expression dans le titre de sa section 2[232], sans en donner une définition juridique ni lui conférer une réelle portée normative. Elle suit la recommandation d'une approche au cas par cas et prévoit une série de dispositions visant à imposer l'ouverture des données principales générées par un concessionnaire ou un délégataire de service public (dans les transports, l'eau, la gestion des déchets ou des réseaux d'énergie) ou par un opérateur bénéficiant de subventions publiques, ainsi que des données privées utiles à la statistique publique.

[229] T. Aureau, L. Cytermann, C. Duchesne, M. Morel, L. Vachey, Rapport relatif aux données d'intérêt général, Rapport du Conseil Général de l'Économie, de l'Inspection Générale des Finances et du Conseil d'État, septembre 2015, 93 p.

[230] Décision 14-MC-02 du 9 septembre 2014, Autorité de la concurrence, qui a ordonné à GDF Suez de communiquer une partie des données de son fichier clients à ses concurrents afin qu'ils puissent mieux faire connaître leurs offres.

[231] C. Duchesne, M.Meyer, C. Oslina, M. Perrière, L. Ruat, G. Tirot, L. Vachey, Les données d'intérêt général – Phase 2, Rapport du Conseil Général de l'Économie, de l'Inspection Générale des finances et du Conseil d'État, mars 2016, 412 p.

[232] Articles 17 à 24 de la loi n° 2016-1321 du 7 octobre 2016 pour une République numérique

Une notion souvent convoquée, mais dans des situations très variées, signe d'un réel besoin de qualifier et de désigner une pratique existante

En l'absence de définition stabilisée, de nombreuses contributions nourrissent l'analyse conceptuelle autour des données d'intérêt général, qui connaissent un regain d'actualité lié à l'essor des nouveaux usages de la donnée, en matière d'intelligence artificielle notamment, et aux enjeux d'un pilotage éclairé de l'action publique sur des matières complexes, comme l'environnement ou la santé. On assiste en effet à une véritable dispersion des définitions, qui mêlent des catégories de données multiples, hétérogènes et dont les usages différenciés possèdent tous leur part de légitimité : données environnementales, données d'innovation, données d'apprentissage algorithmiques, données territoriales dans le cadre des projets de smart cities, données à haute valeur ou à fort potentiel. Ce foisonnement s'accompagne d'une certaine complexité, qui est renforcée par le fait que la destination de ces données n'est pas toujours évidente (confidentielles, ouvertes, partagées), ni leur producteur clairement identifié.

Parallèlement, on constate que la notion rencontre un grand succès et qu'elle est régulièrement mobilisée dans les discours publics sur la donnée. Certains invoquent la notion de données d'intérêt général comme une extension du domaine de l'ouverture des données publiques c'est à dire un levier supplémentaire pour obtenir des acteurs privés une ouverture de leurs données. Le rapport du député Luc Belot au Premier ministre plaidait en 2017 pour la création d'un statut pour les données d'intérêt territorial, qu'il illustre par la capacité qui serait offerte aux collectivités d'accéder **de manière automatique et obligatoire** aux données des opérateurs de mobilité tels Uber ou Waze. Pour leur part, les membres du Conseil national du numérique proposent que les données environnementales soient considérées comme des données d'intérêt général et « constituent une brique de la transition écologique et solidaire ».

Dans une autre approche, plusieurs contributeurs de la consultation réalisée dans le cadre de cette mission soulignent que les données d'intérêt général pourraient avoir un intérêt pour d'autres acteurs que les pouvoirs publics, par exemple pour permettre à des tiers (journalistes, société civile, associations) de s'assurer du respect, par les acteurs privés de dispositions légales. On voit ici dans les données d'intérêt général une opportunité pour étendre les obligations de transparence et de redevabilité à certains acteurs du secteur privé.

In fine, la plasticité et la polysémie de la notion présentent autant d'avantages que d'inconvénients. S'il est difficile de lui donner une portée juridique générale en tant que telle, et qu'il faut donc multiplier les lois sectorielles, cette imprécision offre des flexibilités qui permettent de décliner concrètement la notion en tenant compte de la diversité des données, des acteurs et des finalités considérées.

1.2 Les données d'intérêt général : un outil ancien qui s'est développé de façon progressive et hétérogène au profit de politiques publiques sectorielles

La collecte de données privées par la puissance publique est une réalité ancienne qui s'amplifie et se diversifie

Historiquement, la puissance publique a régulièrement produit une partie des données dont elle a besoin pour son action propre. Ainsi, l'Institut national de l'information géographique et forestière (IGN) trouve ses racines dans le service géographique de l'armée, lui-même issu du Dépôt de la Guerre, créé en 1688[233]. L'information géographique est en effet cruciale pour organiser la défense du pays, mener une guerre et asseoir la puissance régalienne de l'État, autant sous Louis XIV qu'à l'ère de la géolocalisation.

[233] https://www.ign.fr/institut/notre-histoire

Aujourd'hui, une large partie de l'action et du rôle même de la puissance publique demeure de collecter, de traiter et d'archiver des données fournies par les particuliers et les organisations privées. Les entreprises sont aujourd'hui soumises à de multiples régimes de déclaration obligatoire, par exemple sur l'embauche d'un nouveau salarié ou encore en matière fiscale. De même, les foyers sont tenus de déclarer leurs revenus, mais aussi les principaux évènements de vie (naissances, unions, décès par exemple). De même, la puissance publique oblige déjà des acteurs privés à révéler publiquement des données (création d'entreprise, comptabilité, permis de construire). De ce point de vue, l'accès par la puissance publique à des données du secteur privé n'est pas un concept si nouveau qu'il pourrait paraître.

Avec la modernisation de l'action publique, l'intervention de la puissance publique se diversifie et un nombre croissant de données sont collectées dans des secteurs de plus en plus variés : éducation, environnement, culture, vie des entreprises, activité économique. En outre, les modalités de l'action publique ont muté avec de nouvelles formes institutionnelles (décentralisation, agencification de l'Etat, privatisations d'entreprises publiques, statistique publique, etc.) qui multiplient les producteurs et destinataires de données.

Les individus et les organisations privées sont donc conduits à fournir un nombre croissant de données, ce qui incite d'ailleurs la puissance publique à rationaliser ses exigences, par exemple avec le dispositif « Dites-le-nous une fois »[234]. En principe, la fourniture de données suppose une base légale, qui oblige le particulier ou l'entreprise. Ainsi, en matière de statistique publique par exemple, la loi prévoit que « les personnes sont tenues de répondre, avec exactitude, et dans les délais fixés, aux enquêtes statistiques qui sont rendues obligatoires »[235].

Les types de fourniture de données à la puissance publique par les entreprises sont nombreux. Sans en dresser la liste exhaustive on peut citer les **obligations de déclaration** fiscales et sociales, les **demandes de subvention**, les **demandes d'autorisations** environnementales, le **dépôt de brevet**, la **transmission des données environnementales**[236], la demande de permis de construire, la **réponse à un appel d'offres** public, la **réponse à une enquête** de la statistique publique, les **réquisitions judiciaires**, etc. La fourniture de données peut être obligatoire ou nécessaire à l'accomplissement d'une démarche volontaire. Tous ces exemples correspondent à la catégorie des données d'intérêt général destinées à la puissance publique (B2G).

Au fil du temps, les rapports sociaux et économiques se sont complexifiés et le service public s'est étendu pour répondre aux attentes de la population. Les données nécessaires à l'accomplissement des missions de service public sont donc toujours plus nombreuses, et permettent notamment de :

- délivrer des autorisations et des titres ;
- contrôler le respect des obligations légales et réglementaires ;
- assurer le fonctionnement des institutions ;
- développer et améliorer les services rendus au public et aux entreprises ;
- permettre des projets de recherche ;
- favoriser le développement économique.

[234] https://www.modernisation.gouv.fr/home/dites-le-nous-une-fois-un-programme-pour-simplifier-la-vie-des-entreprises

[235] Article 3 modifié de la Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques.

[236] Base IREP – registre des émissions polluantes – site géorisques, en application du règlement européen 166/2006 du 18 janvier 2006

En outre, le développement des technologies numériques a produit une rupture quant à la nature des données qui peuvent intéresser la puissance publique et le potentiel qu'elle peut tirer de leur utilisation. Les données dont il est désormais souvent question sont, pour la plupart, des données massives (*Big Data*) liées à des transactions ou générées sous la forme de traces (par exemple les données de caisse ou encore les trajets parcourus par un véhicule). Selon l'organisation des Nations Unies, le *Big Data* est une opportunité de générer des informations en temps-réel, alors que les statistiques publiques apportent une information validée *a posteriori*, *via* des enquêtes rigoureuses. Les deux sources de données se complètent davantage qu'elles ne s'opposent. Comme l'illustre le cas d'usage sur la statistique publique, la combinaison de ces deux types de données nécessite cependant un important travail (cf. le cas d'usage sur les données du secteur privé utilisées par la statistique publique).

Les données *Big Data* seules présentent tout de même une valeur d'usage certaine. La capacité à suivre, en quasi temps-réel, l'évolution des transactions de cartes bancaires dans un secteur d'activité (par exemple les bars et restaurants) et sur un territoire donné peut être une information précieuse, par exemple pour le suivi des impacts de la crise sanitaire. Cette dimension de « tableaux de bord » offre de réels bénéfices en période de forte incertitude.

La puissance publique collabore selon des modalités diversifiées avec les acteurs privés

De nombreuses initiatives ont vu le jour autour des données d'intérêt général, dans des secteurs d'activité et selon des modalités différentes. Le groupe d'experts mandaté par la Commission européenne[237] a recensé au moins cinq modalités de collaboration entre acteurs publics et privés (entreprises et associations) :

- le fournisseur de données réalise les analyses en interne, à l'aide de ses outils propres, et met à disposition de la puissance publique, voire de la société dans son ensemble, les connaissances qui émergent de ces analyses ;
- le fournisseur de données les met à disposition de la recherche afin de produire de nouvelles connaissances scientifiques ;
- le fournisseur de données partage des données au sein de sa filière et peut partager tout ou partie des connaissances produites avec les autorités publiques ;
- le fournisseur de données partage des données avec un ou plusieurs tiers de confiance identifiés ;
- le fournisseur de données autorise un accès direct à certains ensembles de données.

Pour l'ensemble de ces modalités, les données (et les connaissances) peuvent être partagées sur une base volontaire ou faire l'objet d'obligations légales (comme pour les régimes de déclaration obligatoire). Parmi les initiatives volontaristes, on peut citer le réseau des Banques Populaires (Groupe BPCE) qui met à disposition des données agrégées sur les dépenses et paiements des touristes étrangers dans chaque région de France. La société Waze propose aux collectivités de participer à son programme Connected Citizens, qui prévoit un échange de données entre les parties prenantes.

Certains acteurs publics peuvent aussi faire l'acquisition de données du secteur privé par le biais de marchés publics. Les coûts de transaction peuvent être particulièrement élevés (identification de la source de données et de son producteur, négociation des conditions contractuelles, accord sur les modalités techniques et juridiques) et les risques perçus particulièrement forts (incertitude sur la qualité de la donnée fournie, manque d'expérience au sein du secteur public sur l'utilisation de ce type de données).

[237] « Towards a European strategy on business-to-government data sharing for the public interest / Final report prepared by the High-Level Expert Group on Business-to-Government Data Sharing », Publications Office of the European Union, 2020

Il convient enfin de distinguer une grande diversité de finalités et de motivations, pour la puissance publique, à accéder à des données des acteurs privés. La régulation d'un secteur (l'énergie, les télécommunications aujourd'hui, demain sans doute les grandes plateformes numériques) ou l'amélioration des politiques publiques, comme la protection de l'environnement, font notamment partie des motifs identifiés.

Les premières bases juridiques sectorielles qui encadrent l'utilisation des données d'intérêt général sont particulièrement hétérogènes et peu cohérentes

Le cadre juridique français en matière de partage des données d'intérêt général est tout autant fragmenté et dispersé que les initiatives concrètes conduites par les acteurs, de façon volontaire ou contrainte. Il existe en effet de nombreux secteurs où l'intervention du législateur est légitime et a pu s'appuyer sur des motifs d'intérêt général suffisamment robustes pour limiter les droits et libertés des personnes privées. Cette construction juridique a le mérite du pragmatisme car elle permet bien souvent de répondre à des enjeux très spécifiques des secteurs concernés et de se placer au plus près des usages, mais la mission regrette qu'elle ait été élaborée sans vision d'ensemble ni principes méthodologiques communs.

Dans le domaine du transport, l'article 4 de la loi du 6 août 2015 pour la croissance, l'activité et l'égalité des chances économiques a suivi les recommandations du rapport Jutand en prévoyant une obligation de diffusion libre et gratuite des données des services réguliers de transport public et des services de mobilité à des fins d'information du voyageur sur les horaires, prix, tarifs et itinéraires[238]. **En 2019, la loi d'orientation des mobilités[239] a renforcé cette dynamique d'ouverture en élargissant le champ des données concernées**, pour y inclure les données statiques et dynamiques sur les déplacements et la circulation, ainsi que les données des bornes de recharge électrique (localisation, puissance, tarification, etc) et des services de covoiturage, de free-floating et des gestionnaires de l'infrastructure ferroviaire (cf. cas d'usage sur le bilan de l'application de la législation en matière de données de mobilité).

Dans le domaine de la **santé**, la loi du 26 janvier 2016 de modernisation de notre système de santé permet l'intégration des données publiques de l'assurance maladie et des hôpitaux avec des données privées des complémentaires santé. La création du Health Data Hub en 2019 a constitué un saut qualitatif substantiel dans l'enrichissement du catalogue de données de santé exploitables pour la recherche, l'appui au personnel soignant, le pilotage du système de santé, l'information des patients et le développement des technologies d'intelligence artificielle en santé.

Dans le domaine de la **formation professionnelle**, la loi du 8 août 2016 relative au travail, à la modernisation du dialogue social et à la sécurisation des parcours professionnels prévoit la mise à disposition des données des organismes de formation sur les entrées et les sorties des formations et les taux de retour à l'emploi.

[238] https://www.legifrance.gouv.fr/jorf/id/JORFARTI000030978647

[239] Loi n° 2019-1428 du 24 décembre 2019 d'orientation des mobilités.

Dans le domaine du **logement**, la loi pour une République numérique et la loi ELAN[240] ont prévu des obligations de coopération renforcée entre les plateformes de location de meublés touristiques et les communes, afin de faire respecter la limite de location d'un meublé fixée à 120 jours par an. La commune peut ainsi exiger de la plateforme qu'elle lui fournisse une liste détaillée des biens loués par son intermédiaire, avec à chaque fois l'adresse du logement, son numéro de déclaration et le nombre de jours au cours desquels il a fait l'objet d'une location, selon des modalités et une fréquence fixées par décret[241].

Dans le domaine de l'**énergie**, certaines dispositions de la loi relative à la transition énergétique pour la croissance verte[242] permettent au fournisseur d'électricité de recevoir des gestionnaires de réseaux leurs données de comptage, des systèmes d'alerte liés au niveau de consommation, ainsi que des éléments de comparaisons pour qu'ils soient ensuite mis à la disposition gratuitement du consommateur. Cette loi a également créé un registre national des installations de production et de stockage d'électricité, alimenté et tenu à jour par les gestionnaires du réseau, qui est mis à disposition du ministre chargé de l'énergie.

Malgré leur caractère non exhaustif, force est de constater que ces exemples de dispositions sectorielles permettent déjà de mettre en évidence la diversité et la fragmentation de notre ordre juridique en matière de données d'intérêt général. Sur le fond, cette hétérogénéité n'est pas nécessairement préjudiciable en soi, car l'approche par cas d'usage permet de calibrer les dispositifs de façon adaptée aux différentes problématiques. S'agissant de la méthode en revanche, le processus de construction de ces dispositions pourrait être davantage concerté et discuté, afin de structurer et donner une meilleure cohérence aux initiatives.

La mission estime qu'un effort de généralisation à partir des cas d'usage permettrait d'accélérer et de rehausser les ambitions d'une politique publique des données d'intérêt général, qui ne serait plus mobilisée de façon accessoire ou incidente mais construite de manière réflexive et volontariste.

Pour ce faire, la mission considère qu'il est de la responsabilité de l'État d'entreprendre une démarche inductive afin d'identifier systématiquement ce qui, dans chacun de ces cadres sectoriels existants et futurs, pourrait relever de bonnes pratiques susceptibles de nourrir une doctrine d'aide à la décision. La puissance publique pourrait alors décliner ce cadre, sans avoir à se poser les mêmes questions à chaque nouvelle problématique, ce qui permettrait à l'action publique de mieux suivre le rythme des évolutions rapides des usages et technologies numériques.

Aujourd'hui, beaucoup trop de besoins demeurent insatisfaits dans la sphère publique, comme le révèle la pratique grandissante et non encadrée de *scraping* par certains acteurs (le *web scraping* consistant à extraire, à l'aide d'un programme, les données d'un site web pour les réutiliser). Ces derniers ont besoin d'un recours et d'une médiation pour exprimer ce besoin et pouvoir développer des partenariats avec les acteurs privés en vue de partager davantage d'informations.

La définition d'une doctrine d'action partagée entre tous les acteurs est d'autant plus impérative que les données d'intérêt général constituent un levier puissant de transformation et de résolution des problèmes complexes auxquels nos sociétés sont confrontées. Il est désormais temps de passer à l'échelle.

[240] Article 145 de la loi n° 2018-1021 du 28 novembre 2018 portant évolution du logement, de l'aménagement et du numérique

[241] Décret n° 2019-1104 du 30 octobre 2019 pris en application des articles L. 324-1-1 et L. 324-2-1 du code du tourisme et relatif aux demandes d'information pouvant être adressées par les communes aux intermédiaires de location de meublés de tourisme (https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000039296575)

[242] Article 28 de la loi n° 2015-992 du 17 août 2015 relative à la transition énergétique pour la croissance verte

2. Un passage à l'échelle nécessaire mais juridiquement complexe

2.1 L'accès aux données d'intérêt général par la puissance publique : une nécessité pour affronter les crises et inventer de nouvelles formes de régulation à l'ère numérique

Crise sanitaire, changement climatique, lutte contre la désinformation en ligne : les trois cas développés ici constituent autant d'exemples de la légitimité de la puissance publique à accéder et à utiliser les données du secteur privé. L'objectivation des phénomènes que les données privées permettent les rend indispensables à l'action publique pour traiter ces problèmes majeurs avec précision et efficacité. L'absence d'alternative (une source de données non privées par exemple) et l'importance de ces enjeux justifient le passage à l'échelle des politiques sur les données d'intérêt général.

Une meilleure utilisation des données d'intérêt général aurait sans doute permis de gérer la crise sanitaire encore plus efficacement

La crise de la COVID19 a conduit à donner une impulsion très positive au numérique en santé, au sens le plus large du terme, en catalysant les différentes énergies, aussi bien au niveau des acteurs du secteur sanitaire, que dans les autres secteurs.

Le numérique s'est en effet trouvé au cœur d'une réflexion sur l'organisation de la réponse à apporter à cette crise sanitaire au travers du système de soins : la fiabilisation de l'information sur les moyens disponibles s'est posée, en particulier pour disposer de données à un niveau de granularité plus fin. Mais il est apparu rapidement que les données utiles à la gestion de cette crise n'étaient pas toutes disponibles.

L'éclatement des acteurs du système de soins, et surtout de leurs systèmes d'information, a nui à la remontée efficace des données en appui à la gestion de la crise sanitaire. De nombreux moyens humains ont dû être mobilisés pour identifier les données utiles et les utilisateurs potentiels. Les contractualisations nécessaires à la remontée de certaines données ont été longues et ont plusieurs fois avorté. Certains acteurs ont dû s'organiser pour reconstruire une information qui existait par ailleurs mais qui ne leur était pas disponible.

Certaines données privées, telles que celles d'un service de prise de rendez-vous comme Doctolib, Qare, Livi ou H4D, peuvent avoir une utilité indéniable dans le cadre de la crise pour compléter les données publiques. Ainsi, selon les informations communiquées à la mission, certaines sociétés ont pu transmettre des données sur le niveau d'activité des cabinets de médecine de ville, sur les délais d'accès et le renoncement au soin ainsi que sur le nombre de tests PCR prescrits et le temps d'attente pour les réaliser. Mais leur partage peut aussi être freiné par la crainte de la part de l'entreprise que ses clients, aussi bien les professionnels de santé que les patients, ne soient pas en accord avec cette utilisation de leurs données. De plus, cette communication auprès des pouvoirs publics s'effectue aujourd'hui de manière informelle, sans s'appuyer sur une disposition légale spécifique à ce type de pratique. Il n'existe pas non plus d'inventaire précis des données potentiellement utiles à la gestion de crise (à l'image des « opérateurs d'importance vitale », dont font partie des entreprises privées, et qui font l'objet d'un inventaire et d'un suivi spécifique par l'État), inventaire dont la constitution serait complexe mais devrait rester suffisamment dynamique par rapport à l'évolution de la situation et des besoins.

Plus globalement, la crise sanitaire a révélé le potentiel et les tensions de l'utilisation des données du secteur privé par le secteur public. Les données anonymisées des opérateurs de téléphonie mobile ont permis de mieux suivre les déplacements de population pendant les périodes de confinement. Les données de transaction de cartes bancaires ont éclairé la décision publique, en permettant de mesurer, en quasi temps-réel, l'impact de la crise sanitaire pour différents secteurs d'activité. Mais les conditions de cet accès ont été improvisées, sur des bases non pérennes, contestées notamment du fait des conditions économiques de leur mise à disposition, jugées inacceptables sur une longue période (cf. cas d'usage sur les données du secteur privé utilisées par la statistique publique).

L'exploitation des données environnementales : un impératif majeur pour la bonne compréhension et résolution des défis climatiques

Les données environnementales sont toujours plus massives, hétérogènes et multi-sectorielles. La croissance continue et la diversité des modes de production des données environnementales s'explique aussi bien par le développement des technologies numériques que par l'hétérogénéité des facteurs de désordre environnemental[243].

D'une part, la modernisation d'un certain nombre d'activités économiques (industrie 4.0, agriculture connectée) et de service public (gestion de l'eau, des déchets, des réseaux d'énergie) grâce aux outils numériques s'accompagne nécessairement d'une génération accrue de données de pilotage et de suivi, dont certaines concernent directement l'environnement. D'autre part, de plus en plus d'activités humaines et de politiques sectorielles (santé, mobilité, aménagement du territoire, alimentation) entretiennent des boucles de rétroaction avec l'environnement, qui doit être appréhendé de façon globale et multidimensionnelle.

En conséquence, **les données environnementales sont actuellement fragmentées entre de nombreux acteurs** (administrations, entreprises, associations, citoyens) et secteurs. Elles sont également « disponibles à différentes échelles régionales et temporelles, peuvent être gratuites ou payantes, comporter des restrictions sur l'usage ou la diffusion »[244], ce qui complexifie leur identification et utilisation. Ce constat plaide pour une définition large des données environnementales, nécessaire pour en saisir toute la richesse et la diversité, et justifie le besoin d'une typologie.

Dans son avis dédié de juillet 2020, **le Conseil national du numérique propose de distinguer les données environnementales par nature**, qui se rattachent directement au domaine de l'environnement (par exemple les données géographiques), **des données environnementales par destination** qui, sans relever spécifiquement du domaine de l'environnement (par exemple la mobilité, la consommation d'eau ou d'énergie), peuvent servir la transition écologique lorsqu'elles sont croisées avec d'autres données. Il en retient finalement une définition extensive, « comme toute donnée, par nature ou par destination, relative à l'environnement, à son état et/ou à ses flux d'interaction », qui présente l'avantage de pouvoir multiplier les jeux de données dont l'utilisation peut être mise au service de l'intérêt général.

Des finalités d'intérêt général incontestables : éclairer l'action publique (B2G), informer le citoyen et mettre l'efficacité économique au service de la transition écologique (B2B)

Une plus grande mise à disposition et traitement des données environnementales permet d'améliorer la connaissance sur l'environnement afin de mieux le protéger, de garantir le droit à l'information du public et de maîtriser l'empreinte environnementale des agents économiques.

[243] Les développements de cette partie sont inspirés de l'avis du Conseil national du numérique (*Faire des données environnementales des données d'intérêt général*, Juillet 2020, 60 p.)

[244] Gemma Cirac Claveras et Cédric Gossart, « Enjeux et perspectives des données environnementales massives », Terminal, 117, décembre 2015.

Pour la puissance publique, l'utilisation de données privées permet aussi bien d'enrichir l'ouverture des données publiques, d'aider à la prise de décision que de mesurer en temps réel certains phénomènes afin de mieux calibrer les dispositifs de soutien à la préservation de l'environnement. Les exemples sont nombreux. Ainsi, la loi n° 2006-1772 du 30 décembre 2006 sur l'eau et les milieux aquatiques a instauré l'obligation pour les distributeurs de produits phytosanitaires de déclarer leurs ventes annuelles auprès des agences et offices de l'eau, afin de mieux évaluer et gérer le risque « pesticides ». En matière de protection de la biodiversité, l'entreprise britannique The World Bee Project CIC a permis le déploiement d'un réseau de ruches connectées, en partenariat avec plusieurs universités et l'entreprise Oracle Cloud, afin de suivre en temps réel les colonies d'abeilles et mieux comprendre les causes de leur disparition. Les données sont partagées avec des chercheurs, apiculteurs, agriculteurs et certains gouvernements pour l'élaboration de mesures correctrices.

Dans le domaine du transport, la création d'un système de registre de preuves de covoiturage hébergé sur une plateforme numérique, permettant à l'ensemble des opérateurs de covoiturage, volontaires (par exemple BlablaLines, Karos ou Klaxit), de faire converger et d'attester les trajets effectués par les utilisateurs (cf. cas d'usage sur le bilan de l'application de la législation en matière de données de mobilité). Ce registre donne ainsi la possibilité aux collectivités et aux entreprises de distribuer des incitations (monétaires ou non monétaires) pour favoriser la pratique du covoiturage au quotidien. Il trouve toute son utilité pour l'application de plusieurs mesures de la loi d'orientation des mobilités, comme l'article 35 (versement d'allocations par les autorités organisatrices pour des trajets effectués en covoiturage) ou l'article 82 (forfait mobilités durables).

Pour le citoyen, un meilleur accès et partage de données détenues par des acteurs privés permet d'éclairer les choix de consommation et de contribuer à la co-production des politiques environnementales. Il en va notamment ainsi des projets Open Food Facts et Numalim qui fournissent une meilleure transparence sur la composition des produits alimentaires et donnent aux citoyens la possibilité de soutenir les filières et entreprises soucieuses de leur performance sociétale et environnementale.

Pour les entreprises, une circulation accrue des données au sein de leur écosystème ou à destination de start-up soutient l'innovation au service de l'optimisation des processus de production et la croissance verte. Nombre d'entreprises de la GreenTech[245] proposent ainsi des solutions fondées sur les données destinées à permettre le pilotage de la transition écologique de leurs clients. Par exemple, la start-up ViJi propose un système de traçabilité et de valorisation des actions écoresponsables à destination des marques, fournisseurs et clients du secteur de l'habillement, à partir de la collecte des données RSE. Quant à la société Agdatahub, elle met à la disposition des acteurs de la filière agricole une plateforme SaaS permettant aux acteurs d'échanger et valoriser leurs données agricoles en toute sécurité, pour améliorer les pratiques actuelles ou créer de nouveaux services innovants en matière d'agroécologie, d'agriculture de précision ou de performance économique.

La diversité des exemples précédents illustre bien à quel point les données environnementales sont présentes partout, et donc nulle part. Tant la nature des données que les finalités d'intérêt général, les modes de partage, les acteurs concernés et les équilibres économiques des filières sont multiples, si bien que la puissance publique peut être fondée à intervenir pour accompagner, structurer et mettre en cohérence les initiatives existantes.

Le cas emblématique pour un modèle de gestion par projets et à travers le soutien à la constitution de communs numériques

Au regard de la complexité des matières et de leurs interdépendances fortes, la mission estime qu'une généralisation excessive du concept de données d'intérêt général appliquée à l'environnement ne paraît pas pertinente, au risque de compromettre son caractère opérant. Dans le domaine de l'environnement, plus qu'ailleurs, c'est bien à travers le soutien à des projets concrets et en voie de structuration que le rôle de la puissance publique sera le plus efficace.

[245] Voir : https://greentechverte.fr

À cet égard, la mission déconseille l'approche retenue par le Conseil national du numérique lorsqu'il recommande de transcender la distinction entre données publiques et données privées pour « *englober les données privées et publiques dont l'ouverture se justifie pour répondre à un motif d'intérêt général relatif à la transition écologique et solidaire*. » Elle maintient à l'inverse la nécessité de conserver la distinction actuelle pour réserver la notion de données d'intérêt général aux seules données détenues par les acteurs privés, qu'ils participent ou non à une mission de service public, afin de préserver la clarté des débats et des sécurités juridiques spécifiques à prévoir pour les données privées.

Pour autant, elle estime que l'approche du Conseil via le soutien aux communs numériques est tout à fait pertinente et considère également que « les pouvoirs publics pourraient jouer un rôle essentiel de "catalyseur" en fournissant une aide déterminante à la constitution des communs sur les données environnementales d'intérêt général, en trouvant pour chaque situation et pour chaque territoire, des formes adaptées ». Il est en effet ressorti des auditions conduites par la mission que les acteurs de ces communs sont très demandeurs d'une plus grande implication de l'État et des collectivités dans l'accompagnement des projets, que ce soit par du financement initial ou une prise de participation au capital, un rôle d'animation, de conseil et de fourniture d'expertise, ou la clarification du régime réglementaire de certaines briques technologiques indispensables (par exemple l'identité numérique).

Open Food Facts : un exemple de communs à soutenir

Open Food Facts est une base de données collaborative d'informations alimentaires, indépendante, transparente et librement réutilisable (open data sous licence ODBL). Sorte de Wikipédia de l'alimentation, ce commun numérique répertorie plus de 1 500 000 produits collectés par plus de 20 000 contributeurs. Open Food Facts informe le grand public, aide la recherche scientifique et accompagne les politiques publiques visant à une alimentation plus saine et plus durable. Il a accompagné le déploiement du Nutri-Score en France, puis en Espagne, Suisse, Belgique et maintenant en Allemagne. Sur ce modèle, Open Food Facts s'apprête à lancer l'Eco-Score, qui vise à renseigner sur l'impact environnemental des produits alimentaires.

Né et piloté en France, le projet motorise (via son API) ou a donné naissance à plus de 100 applications mobiles généralistes ou spécialisées (compteur de calories, aide aux diabétiques, allergies, gestion de stock...) — dont la création d'une pépite française Yuka — ainsi que des dizaines d'usages professionnels ou amateurs (articles, cartographies, applications métiers, etc.). Si l'industrie et la recherche française sont les premières bénéficiaires du projet, Open Food Facts est avant tout un projet international, traduit dans plusieurs dizaines de langues et fédérant des dizaines de nationalités et d'usages.

Plusieurs dizaines d'articles scientifiques mentionnent Open Food Facts comme source de leurs données, dont une douzaine comme leur source principale voire exclusive.

Depuis 2019, Santé publique France soutient Open Food Facts, à hauteur de 260 000€ sur 4 ans, pour dé-velopper sa plateforme « producteurs » permettant aux fabricants d'automatiser la publication de leurs produits dans la base citoyenne. Outre le fait d'engager les entreprises dans une œuvre de transparence bénéfique au débat public, Open Food Facts a développé des algorithmes offrant aux industriels des con-seils d'amélioration de leurs produits et de leur Nutri-Score.

Dans le même temps, l'État a financé en 2019, via Bpifrance et le Programme Investissement d'Avenir, une base exclusivement nationale portée par l'Association nationale des Industries alimentaires (ANIA) pour un montant de 3 M€, dans sa première tranche. Depuis la signature du contrat d'aide avec Bpifrance (octobre 2019), Numalim a collecté environ 35 000 produits — le nombre mensuel de nou-veaux produits collectés par Open Food Facts. À ce jour (décembre 2020), la base Numalim ne peut être téléchargée entièrement et n'est accessible que *via* une API payante.

Source : Mission ; Banque mondiale

Enfin, la mission considère l'approche par projet particulièrement adaptée au domaine environnemental, compte tenu de la multiplicité des enjeux et parties prenantes. Dans cette optique, il ne s'agirait pas d'identifier et de qualifier des données d'intérêt général a priori mais plutôt de mettre en lumière les configurations associant certains acteurs, les données qu'ils partagent et les finalités poursuivies dont le résultat contribue à l'intérêt général. Une fois ce travail de veille et d'identification effectué par la puissance publique, elle pourrait accompagner le projet par la mise à disposition d'outils juridiques (contrats types), des infrastructures d'échanges (à l'image du *Green Data Hub* – cf. partie 3), et des incitations au partage volontaire de données.

Une asymétrie d'information de la puissance publique avec les plateformes numériques qui invite à davantage de régulation par la donnée

L'expérience récente montre que les plateformes numériques deviennent structurantes dans le débat public, l'équilibre économique de certains marchés et dans les relations entre les acteurs publics, la société civile et les citoyens. Leurs modèles d'affaires fondés sur la collecte et le traitement de données leur permet de procéder à des analyses très fines des phénomènes et comportements au sein de leurs écosystèmes, qui leur offrent un degré de connaissance supérieur à celui des utilisateurs ou des pouvoirs publics.

Faute d'élaborer de nouveaux modes de régulation qui permettent de rompre avec cette asymétrie d'information, le risque est celui d'une incapacité de l'État à exercer ses pouvoirs de contrôle et de supervision en matière régalienne et économique. Cette problématique est particulièrement prégnante s'agissant de la lutte contre les fausses informations, cas d'usage que la mission a souhaité approfondir au cours de ses travaux.

Au cœur de l'agenda européen et français, la lutte contre la désinformation en ligne a fait l'objet d'une première réponse qui montre ses limites.

Au niveau européen, la Commission a présenté les difficultés que posent les fausses informations pour nos démocraties dans une communication dédiée[246] et a d'abord souhaité y répondre via l'adoption d'un code de bonnes pratiques visant à responsabiliser les plateformes numériques. Comme en témoigne l'évaluation[247] produite un an après, cette incitation à l'auto-régulation a révélé certaines carences liées notamment à l'absence de transparence des acteurs sur les moyens mis en oeuvre et les résultats obtenus. La Commission regrette en particulier l'absence d'indicateurs clés de performance pertinents pour évaluer l'efficacité des politiques adoptées, un manque d'accès aux données permettant une évaluation indépendante des tendances émergentes et des menaces que représente la désinformation en ligne, ainsi qu'une absence de coopération entre les plateformes et la communauté des chercheurs. Afin de faire progresser la réponse, la Commission a souhaité agir à travers une nouvelle législation, le *Digital Services Act*, et un plan d'action pour la démocratie européenne.

[246] Réf. Communication du 26 avril 2018 : « Lutter contre la désinformation en ligne : une approche européenne »

[247] https://ec.europa.eu/digital-single-market/en/news/assessment-code-practice-disinformation-achievements-and-areas-further-improvement

Au niveau français, la loi contre la manipulation de l'information du 20 novembre 2018 prévoit un certain nombre d'obligations de moyens destinées à freiner la viralité des fausses nouvelles et à jeter les bases d'une meilleure coopération entre les plateformes et le régulateur. Dans son premier bilan[248] de l'application de cette loi, réalisé à partir des déclarations obligatoires des plateformes sur les dispositifs de modération des contenus mis en œuvre, le CSA relève également l'insuffisance de données fournis sur des points clés d'appréciation de la bonne application de la réglementation.

Plus précisément, **le régulateur pointe plusieurs séries de données auxquelles il souhaiterait avoir accès** pour apprécier le niveau de conformité des plateformes à leurs obligations :

- les données relatives à la politique suivie à l'égard des publications propageant des fausses informations ;
- les données relatives aux moyens humains et financiers déployés en matière de modération des contenus ;
- les données relatives à l'intelligibilité et la redevabilité des algorithmes de présentation et de hiérarchisation des informations ;
- les données relatives aux procédures de traitement des signalements ;
- les données relatives à l'exercice des droits de recours interne contre les décisions prises par les plateformes (conditions, délai, modalités de traitement du recours).

Faut-il aller vers des pouvoirs d'accès aux données renforcés pour le régulateur ? À l'évidence, l'autorégulation comporte ses limites et les pouvoirs publics en sont bien conscients. Le CSA a fait part à la mission d'un projet d'article législatif qui permettrait de moderniser les pouvoirs d'enquête et d'audit que lui confie la loi de 1986, mais qui ne trouve pas aujourd'hui de vecteur législatif pour être adopté. Cela étant, cette nouvelle compétence qui pourrait être confiée au CSA par le législateur suppose de pouvoir bénéficier d'un retour d'expérience suffisant pour évaluer précisément les besoins d'intervention et l'opportunité de retenir une approche plus coercitive. La régulation se construisant également par la pratique, le CSA préfère pour l'heure exploiter pleinement les moyens dont il dispose, préciser son questionnaire et mieux évaluer le niveau de coopération des plateformes, avant de proposer d'en tirer d'éventuelles conséquences sur le plan réglementaire dans le cadre du *Digital Services Act*.

Cela étant, le niveau d'asymétrie d'information dont pâtit actuellement le régulateur n'a pas manqué de retenir toute l'attention de la mission, qui s'est longuement interrogée sur la nécessité de proposer un renforcement des mesures d'accès aux données des plateformes. En tout état de cause, si une telle piste était choisie à terme, il conviendrait de privilégier une action au niveau européen et de bien distinguer les différents niveaux de transparence et d'ouverture, selon les bénéficiaire d'un tel accès : régulateur, chercheurs, public. Ces accès ne pourront pas être de même ampleur et présenter les mêmes garanties.

[248] https://www.csa.fr/Informer/Espace-presse/Communiques-de-presse/Lutte-contre-les-infox-le-CSA-publie-son-premier-bilan-sur-l-application-et-l-effectivite-des-mesures-mises-en-oeuvre-par-les-plateformes-en-2019

2.2 Un partage des données d'intérêt général qui doit préserver les droits économiques des personnes privées et sécuriser leurs initiatives de partage volontaire

Dans la mesure où les données d'intérêt général permettent de résoudre plus efficacement certaines problématiques complexes auxquelles nos sociétés sont confrontées, la légitimité de la puissance publique à intervenir pour organiser le passage à l'échelle en la matière est fondée. Pour autant, cette intervention ne peut avoir lieu dans n'importe quelles conditions. Diverses modalités d'action sont envisageables, de la plus coercitive à la plus incitative, et un équilibre subtil doit constamment être recherché entre l'intérêt général et les intérêts légitimes des acteurs économiques. En ce sens, la mission estime opportun de renforcer les démarches volontaires de partage de données d'intérêt général que les personnes privées pourraient entreprendre, ce qui nécessite de leur offrir une sécurité juridique suffisante, notamment en matière de protection des données à caractère personnel et de droit de réutilisation par l'administration. De même, toute décision d'obliger les personnes privées à partager leurs données doit nécessairement tenir compte de leurs droits économiques, au premier rang desquels le droit de propriété.

Les personnes privées sont titulaires de droits sur leurs bases de données qui contraignent les démarches de réquisition publique ou de partage forcé de données d'intérêt général[249]

En premier lieu, les personnes privées sont protégées au niveau européen par un droit *sui generis* sur leurs bases de données[250], qui n'est cependant pas de portée générale. Ce droit ne couvre en effet que les bases de données ayant fait l'objet d'un investissement « substantiel » et les seules bases de données structurées[251]. Il permet notamment à son titulaire d'interdire l'extraction ou la réutilisation de tout ou partie du contenu de la base, dans une logique de protection des investissements engagés pour la produire.

D'après le rapport relatif aux données d'intérêt général de 2015, il est raisonnable de penser que ce droit sui generis serait analysé comme un droit de propriété par le Conseil constitutionnel. Une lecture a contrario de la jurisprudence[252] conforte d'ailleurs cette analyse.

En deuxième lieu, les restrictions du droit sui generis ne signifient pas que les bases non structurées, exploitables notamment via les techniques de *Big Data*, ou qui n'ont pas fait l'objet d'un investissement substantiel échapperaient à toute protection. Il est en effet fort probable que **ces ensembles de données soient couverts par le droit de propriété en tant qu'actifs incorporels**, à condition que la personne privée démontre qu'ils ont une valeur économique, ce qui ne semble pas hors de portée à l'heure où les stratégies de valorisation sont multiples au sein de l'économie de la donnée.

En outre, la décision d'une entreprise de constituer et d'exploiter une base de données relève très certainement de sa **liberté d'entreprendre**, principe de valeur constitutionnelle[253] que le juge apprécie de façon large.

---

[249] Voir notamment l'analyse juridique figurant dans le rapport de 2015 sur les données d'intérêt général (p. 40-50) ou Yann PADOVA, « Entre patrimonialité et injonction au partage la donnée écartelée ? », Droit de l'immatériel, Revue Lamy, janvier 2019

[250] Directive 96/9/CE du Parlement européen et du Conseil du 11 mars 1996, dont les dispositions sont transposées aux articles L. 341-1 à L. 343-7 du code de la propriété intellectuelle.

[251] CJCE, 9 novembre 2004, Fixtures Marketing Ltd, C-444/02

[252] Dans sa décision n° 2015-715 DC (§ 107), le Conseil constitutionnel considère en effet que l'obligation de transmission des données relatives aux registres des sociétés par les greffiers des tribunaux de commerce à l'INPI, et non la base constituée dans le cadre de leur activité privée, ne constitue pas une atteinte au droit de propriété.

[253] Décision n°31-132 DC du 16 janvier 1982

Compte tenu de ce qui précède, toute disposition législative qui viendrait imposer une communication B2G ou B2B des données d'intérêt général d'une personne privée sera de nature à méconnaître sa liberté d'entreprendre et son droit de propriété. Saisi d'une telle disposition, **le Conseil constitutionnel ferait varier la nature de son contrôle selon qu'il identifierait une privation du droit de propriété ou une simple atteinte**. Dans le premier cas, il vérifierait si la nécessité publique impose une telle privation et si une juste et préalable indemnité a été prévue par la puissance publique. Dans le second cas, il examinerait si l'atteinte au droit de propriété est justifiée par des motifs d'intérêt général, proportionnée et suffisamment encadrée par la loi, un contrôle similaire étant prévu en matière d'atteinte à la liberté d'entreprendre.

En raison du caractère « non-rival » des données, il est probable que la majorité des dispositifs législatifs contraignants en matière de communication de données d'intérêt général soient qualifiés de simple atteinte au droit de propriété des personnes privées, à condition d'obéir à certains critères. En effet, une obligation de communication de données privées aux seules fins d'exercice de ses missions par la puissance publique, lorsqu'elle ne se comporte pas en tant qu'acteur économique, n'est pas susceptible de priver le titulaire de la base de données de la possibilité de la valoriser auprès de tiers.

Il en va en revanche différemment s'agissant d'une obligation d'ouverture gratuite et générale ou d'un partage contraint en B2B, qui aura pour conséquence de faire disparaître un marché pour le propriétaire des données et justifiera donc le versement d'une indemnisation préalable. De plus, même si elle ne devait pas être accompagnée d'un principe de gratuité, **une obligation de partage en B2B risquerait de perturber le fonctionnement des négociations commerciales entre acteurs et d'altérer les mécanismes de formation des prix sur le marché de la donnée**, si bien qu'une telle option paraît peu opportune au regard des impératifs d'attractivité de l'économie française. Si l'État devait intervenir sur le champ du partage B2B, il conviendrait ainsi de privilégier une approche incitative, d'accompagnement et d'animation d'écosystèmes visant à faire converger les intérêts économiques des acteurs privés.

Si le législateur devait malgré tout retenir des modalités d'intervention plus coercitives, il peut jouer sur plusieurs leviers afin de minimiser le risque constitutionnel et l'incidence sur les finances publiques de ses décisions, tout en préservant la confiance des acteurs privés. Une stratégie consistant à multiplier les dispositifs qualifiables de privations du droit de propriété semble, à cet égard, génératrice d'un coût potentiellement excessif et déséquilibré par rapport au motif d'intérêt général poursuivi, fût-il impérieux, comme la sécurité sanitaire en période d'épidémie.

Pour ces raisons, et dans le prolongement du rapport relatif aux données d'intérêt général de 2015, la mission recommande « **d'envisager en priorité les dispositifs qui ne seraient qualifiés que d'atteintes au droit de propriété**, c'est-à-dire qui imposeraient la communication de données mais sans l'accompagner d'une obligation de gratuité, ou qui réserveraient cette gratuité aux communications à la puissance publique ou aux réutilisations à des fins non commerciales »[254].

Dans cette logique, la mission estime qu'il pourrait être envisagé, **après étude d'opportunité, d'attribuer à certains acteurs publics le pouvoir dont disposent aujourd'hui plusieurs autorités de régulation** (ARCEP, CRE, AMF) pour accéder à des données d'entreprises privées sans indemnisation, **à condition que cet usage n'affecte pas les échanges économiques et obéisse à de solides garanties en matière de protection des données, notamment s'agissant des secrets légaux** (secret fiscal, secret industriel et commercial, en particulier). Ce pouvoir devrait, le cas échéant, être accompagné d'une clarification de la déontologie à laquelle sont soumis les agents publics, mais aussi de mesures d'organisation permettant de préserver les secrets commerciaux dont la puissance publique aurait connaissance.

Une mise à disposition de données d'intérêt général qui doit garantir la transparence et le contrôle de la protection des données personnelles

De nombreux jeux de données comportent en effet des données personnelles et le législateur devra donc être particulièrement attentif au respect des exigences du RGPD. Il en va de même pour tout acteur privé qui souhaiterait volontairement partager ses données à des fins d'intérêt général.

[254] Rapport 2015 page 45

En principe, le respect d'une obligation légale ou l'exécution d'une mission d'intérêt public suffiront à fonder la licéité d'un traitement de données personnelles par la puissance publique. En pratique, il reste néanmoins fondamental de préserver l'exercice du consentement libre et éclairé des individus, qui est facteur d'acceptabilité sociale de tout dispositif de communication obligatoire de données à des fins d'intérêt général, comme l'a révélé la consultation publique conduite par la mission.

Il est en outre nécessaire d'encourager le développement d'un modèle durable de partage des données, intégrant une forte composante éthique et soucieuse des droits fondamentaux, dont fait naturellement partie la protection de la vie privée.

En tout état de cause, chaque dispositif de communication obligatoire devrait comporter des garanties de nature à limiter les risques du point de vue de la protection des données personnelles, qui peuvent être de plusieurs ordres : pseudonymisation des données diffusées, interdiction des tentatives de réidentification, définition des réutilisations autorisées en fonction des données en cause, durée de conservation limitée, etc. Ces dispositions ne pourront rester toutefois que très générales, tant les modalités de partage des données et de respect des droits ne peuvent s'apprécier qu'à l'aune de projets concrets, pour lesquels un accompagnement du ou des régulateur(s) compétent(s) est sans doute nécessaire.

Pour des acteurs privés qui envisagent de partager des données personnelles à des fins d'intérêt général, la difficulté d'identification des conditions à remplir pour respecter le RGPD peut constituer un frein important, tant juridique que psychologique. C'est pourquoi, dans une optique de sécurisation de ces démarches volontaires, la mission recommande l'élaboration par la CNIL d'un guide juridique et pratique donnant aux acteurs une grille d'analyse et différentes options de mise en conformité. Ce guide pourrait notamment s'attacher à clarifier les questions d'identification des responsables de traitement, de compatibilité du traitement avec les finalités initiales de la collecte, de fondements légaux de la collecte, ou encore de conditions d'information et d'exercice des droits des personnes tout au long du processus.

> **Recommandation** : Sécuriser le cadre juridique du partage volontaire de données d'intérêt général concernant l'utilisation des données à caractère personnel

Le droit d'accès aux documents administratifs peut constituer une source d'insécurité juridique pour les acteurs privés souhaitant s'engager dans des démarches de partage volontaire de leurs données

Les entreprises qui envisageraient de partager volontairement leurs données à une personne publique pour faciliter l'exercice de missions d'intérêt général peuvent légitimement souhaiter être sécurisées sur le sort des données transmises, notamment s'agissant des repartages entre différentes administrations, des réutilisations ultérieures et du respect d'éventuels secrets protégés par la loi.

Or, l'article L. 300-2 du code des relations entre le public et l'administration (CRPA) définit la notion de document administratif de façon très large, comme tous les documents « produits ou reçus » par les personnes publiques dans le cadre de leur mission de service public. Cette définition englobante est nécessaire pour donner une large assise à la transparence de l'action publique, mais il en résulte une forme « d'effet contaminant » : **toute donnée reçue par la personne publique devient un document administratif, régie par une obligation de diffusion** sur demande ou spontanée.

Certes, **seules les informations publiques non grevées de secrets protégés ou de droit de propriété intellectuelle de tiers sont librement réutilisables**, y compris à des fins commerciales. Mais cette distinction est fine et échappe le plus souvent aux entreprises, qui peuvent alors être découragées de partager volontairement des données avec les personnes publiques. Les entreprises peuvent également estimer que la possibilité de réutilisation par d'autres administrations des données transmises à une première administration est déloyale, au sens où dans cette situation, elles n'auraient pas fourni spontanément les données.

Par ailleurs, il est ressorti des échanges entre la CADA et la mission que certaines autorités publiques peuvent également rencontrer **des difficultés pour apprécier avec certitude dans quelle mesure elles seraient tenues de communiquer des documents qu'une entreprise leur transmettrait à des seules fins de conseil**, qui devrait a priori échapper au cadre réglementaire d'une communication obligatoire. De plus, les textes et la jurisprudence en matière d'appréciation des secrets protégés, de l'aveu même de la CADA, vont plutôt dans le sens de la communication. Dans certaines configurations bien précises, cet état du droit et de son application peut exposer l'acteur privé à une insécurité juridique préjudiciable à son activité économique.

Il en va par exemple ainsi d'une start-up qui développerait un nouveau produit basé sur un traitement de données personnelles, qui n'existe pas sur le marché, et pour lequel elle souhaiterait recueillir l'avis informel de la CNIL en amont de la finalisation de son analyse d'impact sur la protection des données (AIPD). Bien que l'entreprise puisse considérer que la simple existence du document doit en soi être un élément protégé, dont la confidentialité lui permet de garder un temps d'avance sur ses concurrents, il sera qualifié de document administratif et donc communicable. L'occultation des éléments internes à l'AIPD protégés du secret industriel et commercial ne suffira donc pas à sécuriser un tel projet.

Cet exemple n'est pas purement théorique. Bien que l'administration ne communique pas sur une saisine à titre de conseil, en pratique, l'information relative au fait que des discussions sont en cours sur les conditions juridiques d'un projet peut circuler au sein des services de l'administration ou de l'entreprise et certains acteurs peuvent avoir un intérêt à en demander la communication.

Dès lors, la mission recommande de clarifier le cadre applicable pour tous les acteurs et de sécuriser davantage les acteurs privés dont les droits de propriété intellectuelle et le secret industriel et commercial sont protégés. Si la majorité des cas semble pouvoir être traitée par l'élaboration d'un guide fournissant une analyse juridique officielle, dont la rédaction pourrait être confiée à la CADA, l'autorité a souligné en audition que les cas les plus difficiles à apprécier semblent, en première analyse, ne pouvoir être traités que par une modification du CRPA.

La production de ce guide juridique devra donc s'accompagner, le cas échéant, d'une mise en évidence des problématiques pour lesquelles seule la loi est susceptible d'apporter une réponse. Si tel était le cas, l'intervention du législateur pourrait se traduire par la création dans la loi de la notion de « partenariat de données », qui viendrait encadrer les conditions de mise à disposition volontaire de données à une personne publique par une entreprise, à des fins d'intérêt général. Ce nouveau régime juridique pourrait notamment traiter la question du degré de contrôle à laisser aux entreprises sur les conditions de réutilisation et repartage des données, préciser le sort réservé aux données communiquées dans le cadre de rapports normaux de conseil avec l'administration, et clarifier les conditions de confidentialité associées au partenariat.

> **Recommandation** : Sécuriser le cadre juridique du partage volontaire de données d'intérêt général concernant le droit d'accès et de réutilisation applicable aux données du secteur privé reçues par les administrations

3. Pour une approche par la confiance, incitative et européenne

3.1 Inscrire la circulation des données d'intérêt général dans une logique d'autodétermination informationnelle, source de confiance des citoyens dans la puissance publique

Une consultation en ligne conduite par la mission qui révèle une certaine défiance des citoyens dans le partage des données d'intérêt général

Dans le cadre de la consultation publique en ligne, les participants ont été invités à réagir sur dix cas d'usage, par les administrations, de données du secteur privé, notamment dans le domaine de l'énergie, des télécommunications, de la mobilité, du logement ou encore de la santé. Près d'une centaine de réactions à ces cas ont été enregistrées sur le site de la consultation. Les biais de participation et de représentativité sont bien sûr à prendre en compte.

Cependant, il en ressort plusieurs éléments utiles au débat sur les données d'intérêt général et un grand nombre de ces éléments sont aussi confortés par d'autres enquêtes et dispositifs[255]. **Le point saillant est le manque de confiance – voire parfois la défiance – à l'égard de la puissance publique**. « *Inutile de renforcer la surveillance* » écrit l'un des contributeurs, résumant un sentiment partagé par de nombreux contributeurs. **Le cadre présenté par la mission pour ces cas d'usage, qui souligne qu'ils doivent être envisagés avec les garanties de l'anonymisation des données et de la limitation des analyses de données à des niveaux agrégés, à l'exclusion de toute donnée individuelle, ne semblent pas suffisant pour garantir la légitimité de l'usage envisagé** et pour répondre aux inquiétudes quant au respect des droits et libertés individuelles. La question du respect de la vie privée est prédominante, même quand les cas d'usage présentés excluent toute réidentification des individus.

Parmi les dix cas d'usage présentés, seuls deux font l'objet d'un accueil a priori plus favorable, celui concernant l'usage des données des assureurs pour l'amélioration de la sécurité des infrastructures routières, et l'accès aux données des sites de référencement immobilier dans une optique d'encadrement des loyers. Cela ouvre la porte à un autre élément du débat, la question de l'évaluation de l'intérêt général par les individus. En reformulant, la question n'est ainsi plus seulement **à qui fait-on confiance** (et pas à la puissance publique, si l'on en croit les réactions exprimées) mais aussi et surtout **: pour quel usage**, pour quelle finalité accorde-t-on sa confiance ? Les individus doivent être étroitement associés aux échanges entre le secteur privé et la puissance publique, *a fortiori* lorsque des données personnelles sont en jeu, et être considérés comme des acteurs à part entière de cet échange. Cette notion d'autodétermination informationnelle est au cœur du droit à la portabilité des données, comme souligné dans le rapport du Conseil d'État sur le numérique et les droits fondamentaux[256].

[255] Voir notamment l'enquête menée au niveau européen par l'Open Data Institute : « Attitudes towards data sharing », 2018

[256] Conseil d'État, Le numérique et les droits fondamentaux, étude annuelle de 2014.

De la portabilité des données à l'altruisme des données : faire des individus des acteurs du partage des données

Le droit à la portabilité des données a été introduit par le règlement européen sur la protection des données à caractère personnel (RGPD). Il offre aux utilisateurs la possibilité de récupérer une partie de leurs données afin de les stocker ou les transmettre à des tiers en vue d'en faciliter la réutilisation à des fins personnelles. Ce droit est souvent présenté comme un moyen, pour les utilisateurs, de contrôler et maîtriser l'usage des données les concernant. C'est aussi un outil de régulation économique dans la mesure où il vise à limiter les effets de verrouillage des clients dans un écosystème fermé. La portabilité est censée permettre une meilleure concurrence entre les fournisseurs de services numériques. La même disposition a donc une **double finalité** (concurrentielle et contrôle par les individus eux-mêmes des données les concernant). « Dans la perspective concurrentielle, le pouvoir conféré aux individus n'est qu'un moyen au service d'une fin plus vaste, qui est la régulation du marché » écrit Emmanuel Netter, qui propose une autre vision de la portabilité tout entière au service de l'individu, dans une optique de « souveraineté personnelle[257] ».

Deux ans après l'entrée en vigueur du RGPD, la Commission Européenne a publié en juin 2020 une évaluation des effets du nouveau règlement[258]. **La portabilité des données y est qualifiée de « potentiel non entièrement réalisé à ce jour »**. En cause notamment le manque de standards (tant au niveau des données que des interfaces) qui limite dans les faits la portabilité des données. Dans le domaine financier, la directive européenne DSP2 impose aux acteurs du secteur bancaire et financier des standards de partage et de portabilité des données en vue de stimuler la concurrence. Cette démarche d'open banking n'a pas encore produit les résultats escomptés en France, alors qu'une démarche similaire au Royaume-Uni permet à près de 2 millions de clients (individus et entreprises) d'utiliser leurs données bancaires personnelles sur des sites autres que celui de leur établissement bancaire.

En Europe, et particulièrement en France, plusieurs acteurs n'ont pas attendu la mise en œuvre du RGPD pour tester de nouvelles modalités de partage des données personnelles, sous le contrôle des utilisateurs et dans une logique d'**autodétermination informationnelle**. C'est notamment le cas de l'initiative Self Data, porté dans notre pays par la Fondation Internet nouvelle génération. Depuis 2017, la FING et ses partenaires ont expérimenté cette notion dans plusieurs domaines, dont l'énergie et la santé. Depuis 2019, trois territoires (Nantes, La Rochelle et le Grand Lyon) expérimentent le Self Data territorial, déclinaison locale et opérationnelle de la portabilité et du partage des données.

Il convient de souligner que les acteurs locaux, et en premier lieu les villes et métropoles européennes, sont bien souvent à l'avant-garde de ces enjeux, souvent en avance par rapport aux gouvernements nationaux comme l'illustre la coalition Cities for Digital Rights lancée par Amsterdam, New-York et Barcelone et maintenant rejointe par plusieurs villes françaises (Bordeaux, Grenoble, Lyon et Rennes).

L'altruisme – ou donation – de données consiste, pour les utilisateurs, à partager volontairement avec des tiers pour des projets d'intérêt général. Ce concept a notamment mis en œuvre dans le domaine de la recherche en santé, où les patients sont invités à partager les données les concernant. La Commission européenne a fait de l'altruisme de données l'un des enjeux de la stratégie européenne en matière de données.

La mission est pleinement convaincue du potentiel de la portabilité des données, qui peut permettre aux citoyens engagés de mettre leurs données personnes au service de l'intérêt général. Mais ce droit récent demeure insuffisamment connu et peu mis en œuvre. Pour insuffler une dynamique réelle aux pratiques citoyennes d'altruisme de la donnée, l'État dispose de plusieurs modalités d'intervention :

[257] « La portabilité, un droit à inventer », Emmanuel Netter, Dalloz IP/IT, Dalloz, 2020, pp.352-357

[258] « *Data protection as a pillar of citizens' empowerment and the EU's approach to the digital transition - two years of application of the General Data Protection Regulation* », Commission européenne, juin 2020

- recenser, accompagner et accélérer les expérimentations existantes au niveau local qui visent à redonner un contrôle aux individus sur leurs données et les mobiliser dans le pilotage de certaines politiques territoriales (énergie, mobilité, etc) ou pour des causes d'intérêt général (suivi des déplacements en période de crise sanitaire). Le bilan de ces expérimentations fournirait un ensemble de bonnes pratiques à généraliser pour multiplier ces initiatives sur l'ensemble du territoire national ;
- au niveau national, organiser et promouvoir des campagnes de mobilisation citoyenne sur un nombre limité de grande causes chaque année afin de montrer qu'elles peuvent être efficaces et susciter un effet d'entraînement ;
- soutenir l'émergence et l'adoption de solutions technologiques clés en main (exemple des « *MyData Operators* ») permettant d'actionner concrètement le droit à la portabilité (par exemple, l'accompagnement du projet, un financement PIA ou France Relance, mise en visibilité, etc.).

> **Recommandation** : Encourager les initiatives de portabilité citoyenne des données au service de l'intérêt général, notamment par l'organisation de campagnes de mobilisation citoyenne

3.2 Des spécificités de l'économie de la donnée qui s'accommodent mal d'une approche excessivement coercitive des pouvoirs publics

Avant tout, il convient de rappeler que l'utilité d'un jeu de données privées est difficilement identifiable a priori par la puissance publique et s'acquiert bien souvent par l'échange et l'exploitation au sein des éco-systèmes privés. Il n'existe d'abord aucun répertoire des bases de données privées, ni d'ailleurs de cartographie complète des bases de données détenues par l'ensemble des acteurs publics. Pour améliorer une politique publique ou en lancer une nouvelle, l'État doit donc expertiser les enjeux et identifier les données privées qui seront nécessaires à l'action publique. La découverte des données pertinentes est un exercice en soi, qui suppose des échanges constructifs et progressifs dans un cadre de confiance. De la même manière, les échanges de données entre acteurs privés ne s'improvisent pas, mais sont le fruit d'un historique de la relation commerciale, mais aussi d'échanges, de collaborations, de partenariats, qui permettent la co-construction d'un projet au sein d'un écosystème.

Des difficultés réelles pour la puissance publique à identifier les données dont elle a besoin (B2G)

Le premier élément qui fait obstacle à la pleine efficacité d'une approche coercitive tient à la difficulté pour la puissance publique de connaître l'existence et la nature des données privées qui pourraient servir un motif d'intérêt général.

Le volume de données produites dans le monde croît de manière soutenue et serait passé de 2 à 33 zettazoctets entre 2010 et 2018[259]. Avec la multiplication des terminaux, des modalités de stockage (*cloud*) et des systèmes d'information, l'entreprise elle-même rencontre des difficultés à faire l'inventaire de ses propres données. *A fortiori*, la puissance publique est donc bien souvent placée en situation d'asymétrie d'information.

[259] Statista Digital Economy Compass

La puissance publique a une vision partielle du fonctionnement de l'entreprise. Elle dispose de données sur les entreprises (données sociales, fiscales et douanières, en particulier), mais ces données sont déterminées par les processus internes qui les génèrent et intégrées dans des processus métier des administrations concernées, outre le fait qu'elles sont protégées par des secrets légaux. En dehors de dispositions spécifiques, notamment en matière d'environnement, la puissance publique n'a pas de capacité de connaître les détails de l'activité et l'ensemble du fonctionnement de l'entreprise. Par exemple, la nomenclature d'activités par secteur ne reflète qu'imparfaitement les situations des entreprises, et deux entreprises enregistrées dans le même secteur peuvent avoir des modèles commerciaux très différents.

Même dans les cas où l'entreprise est tenue de communiquer des données, cette communication peut être limitée, parfois même nulle. Par exemple, le concessionnaire d'un service doit fournir à l'autorité concédante les données et les bases de données collectées ou produites à l'occasion de l'exploitation de la concession, conformément à l'article 17 de la loi pour une République numérique. Or, la majorité des collectivités territoriales auditionnées par la mission ont rapporté qu'elles n'accèdent qu'à des données parcellaires et selon des modalités décevantes. Cette fourniture pour le moins limitée de données serait en effet la pratique la plus courante des concessionnaires. Un renforcement de l'effectivité de la loi est donc indispensable pour garantir un meilleur partage des données d'intérêt général au niveau territorial. Plusieurs voies d'action sont recommandées : mettre à disposition des acteurs publics un clausier type à décliner systématiquement dans les contrats ; faire de l'application de l'article 17 de la loi pour une République numérique un point d'attention systématique du contrôle de légalité et des missions des chambres régionales des comptes.

> **Recommandation** : Garantir l'effectivité de la loi pour une République numérique en matière de données détenues par les concessionnaires et délégataires du service public

En dehors de la fourniture de données à différentes administrations prévues par la loi, **les entreprises sont généralement peu enclines à révéler leurs données**, qu'elles soient sensibles ou non. Comme l'ont montré les auditions conduites par la mission, certaines collectivités territoriales ont négocié avec des entreprises du numérique comme Airbnb, Waze ou Bookings, mais sans grand succès et avec des doutes sur la qualité des données diffusées.

La crise sanitaire a accéléré les conventions de fourniture de données (cf. cas d'usage sur les données du secteur privé utilisées par la statistique publique). Mais ces communications ponctuelles ne sont pas nécessairement amenées à se poursuivre dans le temps et de façon gratuite. Car la question de fond est bien la nature des données espérées par la puissance publique : celles qui intéressent sont souvent celles que l'entreprise commercialise ou tient à garder secrètes pour préserver son modèle d'affaires.

Si une catégorie d'acteurs devait contribuer à fournir des données, leur standardisation et leur uniformisation serait un obstacle et représenterait un coût, qu'il soit assumé par les entreprises ou la puissance publique. La charge financière est donc potentiellement importante pour la fourniture de données, dont il est parfois difficile d'évaluer *a priori* l'apport réel pour l'intérêt général. Si l'on comprend bien l'intérêt pour une collectivité locale d'analyser l'activité touristique sur son territoire, est-ce que la connaissance des déplacements de population sur le territoire national concomitamment au confinement a une valeur importante ? Cette connaissance n'est-elle pas accessible par d'autres voies ?

Toute demande ou exigence de fourniture de données d'intérêt général devrait donc être subordonnée à un travail d'analyse approfondi qui tienne compte des spécificités de chaque écosystème, permette d'identifier les cas d'usage pertinent pour la puissance publique et de dresser un bilan coût-avantages de l'accès aux données pour résoudre la problématique d'intérêt général.

Une souplesse d'intervention publique requise pour garantir une circulation vertueuse des données entre acteurs privés (B2B) et respectueuse des modèles d'affaires privés

L'établissement d'un cadre général contraignant de partage de données entre acteurs privés soulève un certain nombre de problèmes qui rendent l'exercice risqué et difficilement souhaitable pour diverses raisons, par exemple :

- toutes les données détenues par une entreprise ne relèvent pas nécessairement de sa propriété et peuvent appartenir à ses fournisseurs ou ses clients, ce qui complexifie le partage des responsabilités si l'obligation de mise à disposition est insuffisamment précise ;
- les données peuvent être considérées, parfois à tort, comme un actif stratégique par les entreprises sur lesquelles elles souhaitent conserver une maîtrise et prévenir la réutilisation par des concurrents, en particulier les grandes entreprises étrangères du numérique, qui pourraient les dépositionner de leur marché ;
- la constitution de jeux de données représente un investissement et un coût pour l'entreprise, parfois même le fondement de son activité, qui lui permettent de générer une valeur économique que le partage obligatoire peut amoindrir ;
- le partage de données B2B peut être détourné de ses finalités initiales (ex : dynamiser l'innovation ou la concurrence) et renforcer les positions dominantes sur un marché ;
- le partage de données B2B peut conduire à des équilibres sous-optimaux, faute de culture de la donnée suffisante chez les entreprises concernées ou d'expérience des pratiques collaboratives en la matière, ou de standards communs de qualité des données ;
- le partage de données B2B peut se heurter à des obstacles techniques, tenant notamment à l'absence d'exigences en matière d'interopérabilité et de sécurité.

Inversement, **l'expérience montre que les écosystèmes dans lesquels les entreprises du numérique se sont développées et ont rapidement atteint une taille critique**, en Chine et aux Etats-Unis notamment, **bénéficient de contraintes moins lourdes sur le partage et la circulation des données** qu'en Europe.

La création d'un **terrain de jeu plus équitable pour les entreprises européennes semble donc devoir passer davantage par la souplesse** que par l'obligation. Une entreprise leader sur son marché peut en effet souhaiter s'engager dans une démarche de partage de données pour dynamiser son écosystème ou sa filière et en tirer une valeur économique, mais elle préférera toujours choisir quelles données partager, avec qui, comment et dans quels termes (prix/intéressement, réciprocité, etc.). **Nombre d'acteurs économiques plébiscitent le cadre contractuel et le levier de l'auto-régulation**, qui leur offrent la flexibilité nécessaire pour déterminer, dans le respect du secret des affaires, les données susceptibles d'avoir un intérêt pour les tiers et les solutions les mieux adaptées à leurs objectifs commerciaux, tout en tenant compte des contextes concurrentiels et d'innovation mouvants[260].

[260] Conseil national du numérique, « Données d'intérêt général », *États généraux des nouvelles régulations numériques - synthèse de la consultation*, (Mai 2020. 49 p.)

Par ailleurs, selon la plupart des observateurs[261], **c'est en fonction des usages que la circulation des données doit être pensée**, l'approche *bottom-up* étant recommandée. Une organisation satisfaisante pour les parties prenantes, notamment les entreprises, procèdera en effet d'un accord structurant et d'une lecture commune pour les divers maillons de la chaîne de valeur considérée. Le volet « Traçabilité » du projet ATLAS de l'AFNET[262], qui vise le développement et la diffusion de standards numériques au sein de 15 filières[263], mais selon une méthodologie commune, s'inscrit dans cette logique.

Airbus Skywise[264]

La plateforme Airbus Skywise lancée en 2017 par l'avionneur européen se concentre sur le partage de données sur le fonctionnement de l'avion avec les compagnies aériennes. Les données concernées sont très nombreuses (un appareil de type A 350 embarque plus de 250 000 capteurs) et proviennent de l'ensemble des composants de l'avion. La plateforme ne propose pas seulement des données, mais aussi des capacités de visualisation, de gestion d'alerte, de prédiction et d'apprentissage automatique, des analyses prédictives, notamment en matière de maintenance et d'entretien des appareils. Aujourd'hui, plus de 80 compagnies aériennes dans le monde sont déjà connectées à Skywise. Ce modèle, centré sur une filière autour d'un acteur dominant, rappelle celui de l'échange de données informatisées (EDI) très présent dans le domaine industriel (automobile ou aéronautique) pour l'approvisionnement de la chaîne de production. La plateforme Skywise met en relation la filière amont (les fabricants des composants) et l'aval (les compagnies aériennes et les sociétés de maintenance) autour des données d'usage et non plus de production. Skywise permet de gérer une flotte d'avions sur toute sa durée de vie, en intégrant l'ensemble de ses opérations d'exploitation et de maintenance. Il s'agit notamment pour une compagnie aérienne de maximiser la disponibilité d'une flotte d'avions pour accroître la performance opérationnelle et économique.

Cette illustration montre bien à quel point la circulation des données apparaît de manière croissante au cœur de l'agenda stratégique des acteurs privés. Pour autant, les **niveaux de maturité et d'acculturation aux enjeux de la donnée varient fortement d'une filière à l'autre et en fonction des tailles d'entreprises**, les PME pouvant accuser un retard si elles ne bénéficient pas d'un effet d'entraînement via leur écosystème ou l'appui d'un grand groupe. Cette hétérogénéité se vérifie à travers certaines comparaisons internationales, aussi bien sur des cas d'usage spécifiques des données, comme l'intelligence artificielle[265], que de façon plus générale[266]. Les études démontrent notamment que les entreprises les plus matures en matière de partage des données B2B gagnent en compétitivité via l'optimisation de leurs processus d'organisation interne et un accroissement de leurs opportunités commerciales.

Plusieurs modèles d'affaires sont identifiables au sein de l'économie de la donnée :

[261] Cf. par exemple le rapport de l'ANRT « Prix et valeur des données dans la plateformisation numérique - Repères pour les relations interentreprises », Les Cahiers FutuRIS, Octobre 2019

[262] Source : audition de l'AFNET par la mission.

[263] Ce projet concerne 15 des 16 filières du Conseil national de l'industrie : Aéronautique, Agroalimentaire, Automobile, Bois, Chimie, Construction, Déchets, Eau, Electronique, Ferroviaire, Mines & Métallurgie, Mode & Luxe, Naval, Nucléaire, Santé.

[264] Id. & Audition Cap Gemini

[265] Capgemini Research Institute, *The AI-powered enterprise. Unlocking the potential of AI at scale*, 2020 (https://www.capgemini.com/wp-content/uploads/2020/07/State-of-AI_Report_Web.pdf)

[266] European Commission, Directorate-General of Communications Networks, Content & Technology, *Study on data sharing between companies in Europe*, 2018 (https://op.europa.eu/en/publication-detail/-/publication/8b8776ff-4834-11e8-be1d-01aa75ed71a1/language-en)

- la monétisation des données : une approche unilatérale par laquelle les entreprises engrangent des revenus additionnels grâce aux données qu'elles partages avec d'autres entreprises, ou via la fourniture de services ;
- les places de marché de données : des tiers de confiance jouent le rôle d'intermédiaire entre les producteurs de données et les utilisateurs via la mise à disposition d'une plateforme sécurisée, et génèrent des revenus à partir des transactions réalisées sur la plateforme ;
- les plateformes industrielles d'échange de données : à l'image de la plateforme Skywise, il s'agit d'une approche collaborative et stratégique visant à échanger de la donnée au sein d'un groupe restreint d'entreprises, souvent appartenant à la même filière, qui participent à ces initiatives pour améliorer leur efficacité interne et développer des nouveaux produits et services ;
- les facilitateurs techniques : des entreprises spécialisées et dédiées spécifiquement à la facilitation du partage de données B2B via la mise à disposition, la gestion et l'entretien de solutions techniques ;
- l'*open data* privé : certaines entreprises peuvent décider d'ouvrir certains jeux de données pour permettre de le développer de nouveaux produits et services.

Sur l'ensemble de ces segments, **la puissance publique a un rôle à jouer pour compenser les carences de l'initiative privée et résorber les coûts d'opportunité des entreprises les moins matures**. Mais son intervention, nécessairement délicate compte tenu des blocages identifiés précédemment, doit rester précautionneuse, souple et mesurée pour tenir compte des besoins des entreprises.

En ce sens, et dans le prolongement de l'approche engagée au niveau européen, la mission invite la puissance publique à se comporter essentiellement comme un **facilitateur et un tiers de confiance dans l'orchestration du partage de données B2B**, afin de dynamiser les initiatives existantes et en faire émerger de nouvelles lorsque cela est nécessaire. Les leviers du droit souple, la clarification des contraintes réglementaires, la mise en valeur des bonne pratiques et l'accompagnement financier des projets les plus structurants sont, à cet égard, particulièrement pertinent (cf. partie 4.1).

3.3. Une démarche incitative qui doit permettre à la France d'assurer un rôle de leadership en Europe

Une action européenne permet de garantir une politique cohérente sans nuire à la compétitivité et à l'attractivité de l'économie française

Si la France souhaite s'engager dans une démarche plus volontariste, elle a tout intérêt à porter cette action au niveau européen et à rechercher une coopération de ses partenaires internationaux dans ce domaine. En particulier, le droit européen ne permet vraisemblablement pas en l'état d'imposer la communication d'informations depuis des établissements situés en dehors de France.

La directive commerce électronique interdit d'imposer des règles nationales à des services de la société de l'information fournis depuis un établissement situé dans un autre État membre, sauf pour les finalités qu'elle énumère (ordre public, santé publique, protection des consommateurs). Suivant cette directive, il est probable qu'une loi relative au partage obligatoire de données d'intérêt général ne pourrait s'appliquer qu'aux entreprises fournissant leurs services depuis un établissement situé en France.

La France a un rôle de leadership à prendre en Europe

Le partage et la maitrise des données d'une filière ou d'un écosystème deviennent un facteur clé de succès du développement économique, de l'efficience des entreprises et de l'innovation. L'Union européenne s'est fixé comme objectif de faire de l'économie de la donnée un levier de croissance économique. « La valeur de l'économie des données […] a dépassé le seuil des 400 milliards d'euros en 2019 pour l'UE27 plus le Royaume-Uni, avec une croissance de 7,6% par rapport à l'année précédente. ». Le scénario optimiste envisage « un taux de croissance annuel composé de 11,5% entre 2025 et 2020. L'économie des données croîtra plus rapidement que le marché des données, atteignant une valeur de 827 milliards d'euros dans l'UE27, avec une incidence sur le PIB de l'UE de 5,9 %, contre 4,0 % dans le scénario de référence ». La Commission européenne s'efforce de construire un cadre adéquat (*Digital Single Market*) et de donner des garanties de protection des données personnelles (RGPD).

La Commission européenne a publié en février 2020 sa feuille de route pour une stratégie européenne pour les données avec une ambition élevée : l'Union européenne peut « devenir un modèle de premier plan pour une société à laquelle les données confèrent les moyens de prendre de meilleures décisions, tant dans les entreprises que dans le secteur public ». Cette approche globale vise à traiter l'ensemble des freins à la pleine exploitation du potentiel des données au niveau européen, parmi lesquels le manque d'un cadre harmonisé pour le partage et l'exploitation des données tant du secteur public que privé.

La stratégie européenne sur les données se traduit par un ensemble d'initiatives, dont :

- le *Data Governance Act* (4ème trimestre 2020), cadre législatif générique pour la gouvernance des espaces européens communs de données ;
- l'acte d'exécution sur les séries de données de forte valeur (1er trimestre 2021), dans l'application de la directive de 2019 (anciennement « directive PSI ») ;
- le *Data Act* (fin 2021), loi sur les données qui viserait à adresser plusieurs sujets dont celui de l'accès aux données des entreprises.

Enfin, il faut souligner que la régulation des données des plateformes numériques figurent parmi les problématiques abordées par le *Digital Services Act* (DSA). Il convient de noter que le niveau de maturité est plus ou moins avancé selon les questions : celles qui relèvent de l'*open data* et des évolutions de la directive PSI est bien plus mature que les sujets du partage de données entre entreprises (B2B).

Concernant le partage de données B2G, les questions à traiter pour la préparation du *Data Act* ont été clairement identifiées par le groupe d'experts de haut-niveau, et notamment le caractère volontaire ou obligatoire du partage de données, les modèles de compensation, le traitement préférentiel pour les autorités publiques, la définition des critères permettant de caractériser l'intérêt général de l'accès à certaines données. Sur ce dernier point, il faut souligner que, bien que la notion d'intérêt général ne soit pas définie dans le droit européen, elle sert toutefois de justification à nombre de politiques publiques menées au niveau européen (dont la directive sur les services).

Au niveau européen comme au niveau national, la crise sanitaire a joué le rôle de révélateur et d'accélérateur. Elle a révélé l'absence de cadre harmonisé au niveau européen, plusieurs États membres prenant des initiatives propres. En réponse, le Joint Research Center (JRC), centre de recherche de la Commission a ainsi lancé un programme de partage de données des opérateurs mobiles à des fins de lutte contre la crise sanitaire au niveau européen. De même, la crise sanitaire est un accélérateur de la démarche de données d'intérêt général, la lutte contre la pandémie figurant parmi les motifs d'intérêt général communément admis.

On le voit, les prochains mois vont être décisifs pour l'émergence d'un ensemble de règles communes pour les données d'intérêt général au niveau européen. La France doit jouer pleinement son rôle dans cette période cruciale, à l'instar de celui qu'elle joue dans l'évolution de la directive données publiques ouvertes (anciennement directive PSI). L'existence d'un certain nombre de règles juridiques concernant l'accès des données du secteur privé par les autorités pu-bliques (dont l'article 19 de la loi pour une République numérique) donne à notre pays la possibilité de faire entendre sa voix sur ce sujet et de contribuer ainsi à définir un cadre commun ambitieux.

4. Privilégier une extension méthodique, progressive et concertée du partage de données

4.1. Inciter au partage des données, et contraindre les acteurs en ultime recours

Depuis 2015, le sujet des données d'intérêt général a fait l'objet de nombreuses analyses qui ont mis en évidence **les freins, les opportunités et les leviers d'action** envisageables pour lui insuffler une dynamique durable et lui apporter une réponse définitive. À plusieurs reprises, la littérature administrative **a écarté la proposition de créer un cadre réglementaire général**, transversal et obligatoire pour lui préférer des approches sectorielles, progressives et différenciées, qui ont conduit à la **multiplication de cas d'usage et de dispositions réglementaires, aujourd'hui dispersées et sans cohérence**.

Les données d'intérêt général connaissent un regain d'actualité sur le plan politique, en particulier au niveau européen, et la récurrence des débats autour des nouveaux usages de la donnée, de l'autonomie stratégique de la France et de l'Europe en matière numérique ou de notre capacité à gérer des crises en s'appuyant sur la donnée.

En particulier, **certains travaux récents[267] présentent différentes options de politiques publiques envisageables selon plusieurs scénarios** qui permettent de trancher entre un certain nombre d'alternatives : niveau de coercition pour les acteurs, caractère transversal ou sectoriel de l'approche, nature des finalités et des modalités de partage (B2G ou B2B), typologie des cas de réutilisation des données par les acteurs publics.

La mission propose de structurer la démarche de manière cohérente et générale pour répondre à la variété des enjeux et des secteurs. Pour le partage de données d'intérêt général, la puissance publique doit **d'abord privilégier une approche incitative avant d'envisager une voie coercitive**. Dans ce dernier cas, le législateur doit respecter une doctrine claire pour garantir la cohérence entre les obligations sectorielles et se poser les questions-clés pour traiter au mieux le sujet particulier.

Le principe : privilégier une approche incitative

La première étape consiste à encourager le partage de données sur une base volontaire. La puissance publique doit montrer le caractère souhaitable du partage des données à des fins d'intérêt général. Elle se positionne d'abord dans un rôle de sécurisation juridique, de facilitation, de détection des bonnes pratiques et d'essaimage.

C'est ce type de démarche qui a été initiée jusqu'à présent en France et qui s'inscrit également dans les **orientations proposées par la Commission européenne**. C'est aussi la première des recommandations du rapport du député Cédric Villani (« Inciter les acteurs économiques à la mutualisation de données »).

Cette démarche, qui repose avant tout sur des **instruments incitatifs (appels à projet, recommandations, incitations financières, guides de bonnes pratiques, accompagnement, etc.),** permet de se placer au plus près des usages et d'agir avec pragmatisme pour tenir compte de la spécificité des modèles d'affaires des entreprises, selon les secteurs et les finalités poursuivies.

[267] Conseil national du numérique, « Données d'intérêt général », Synthèse de la consultation sur les données d'intérêt général des États généraux des nouvelles régulations numériques - synthèse de la consultation, (Mai 2020. 49 p.)

Cette démarche comporte des avantages substantiels sur les plans de la **faisabilité** et de **l'efficacité**. Le partage des données s'effectue sur la base du volontariat et présente donc de **moindres difficultés juridiques** (hors respect du RGPD), en particulier s'agissant du droit de propriété, de la liberté d'entreprendre et du secret des affaires. Il s'inscrit dans un relatif consensus politique à l'échelle nationale et européenne et permet, en outre, de minimiser le risque de désincitation à investir pour les acteurs économiques.

En revanche, il nécessite une **médiation active de la puissance publique** pour la promotion du partage de données, la diffusion des bonnes pratiques et l'engagement des acteurs, qui doivent être garantis dans la durée et au-delà des changements de circonstances. C'est pourquoi la mission recommande de confier cette responsabilité au réseau de l'AGDAC et des AMDAC.

> **Recommandation** : Confier au réseau de l'AGDAC et des AMDAC une mission de facilitation et de médiation de l'accès et de l'utilisation des données du secteur privé par le secteur public (B2G), en lien avec la direction générale des entreprises (DGE)

Pour le partage de données entre acteurs privés (B2B), la mission a mis en évidence qu'une **intervention excessivement coercitive de la puissance publique pouvait se heurter à des risques juridiques importants et nuire au modèle économique des entreprises** investies dans l'économie de la donnée. Parallèlement, elle a souligné les vertus du « mode projet » pour sa souplesse et sa capacité à fédérer des acteurs variés autour d'une cause commune. C'est pourquoi elle considère que si l'État devait intervenir sur le champ du partage B2B, il conviendrait de **privilégier une approche incitative, d'accompagnement et d'animation d'écosystèmes pour faire converger les intérêts économiques des acteurs privés vers l'intérêt général**. Diverses modalités d'action concrète sont envisageables à cet égard, et elles mériteraient d'être employées simultanément.

La puissance publique pourrait **participer, animer et subventionner des dispositifs de partage de données, en lançant des projets fédérateurs pour des filières ou des écosystèmes, avec un amorçage humain et financier**. Les comités stratégiques de filière mais aussi certaines associations privées pourraient être le lieu de telles dynamiques, compte tenu de leur expérience pour fédérer des acteurs variés autour d'un projet sectoriel. Des fédérations professionnelles proposent par exemple déjà à des acteurs privés de collaborer et de partager des données au niveau européen pour lutter contre la contrefaçon.

Par ailleurs, il serait opportun de généraliser l'introduction d'un volet sur le partage des données dans les appels à projet publics (PIA) et de renouveler régulièrement les appels à manifestation d'intérêt sur la donnée dont la DGE a la responsabilité.

> **Recommandation** : Développer le partage de données privées au service d'intérêts partagés (B2B) au sein des comités stratégiques de filières, dans les appels à projets publics (PIA), et en soutenant les initiatives associatives et privées

Si l'incitation ne suffit pas pour atteindre certains objectifs, le recours à des instruments plus contraignants pourrait alors être envisagé, **une fois l'appel au volontariat épuisé** et à condition que l'accès aux données privées soit un besoin impossible à satisfaire autrement.

La coercition doit rester l'exception

Cette démarche de partage B2G doit reposer sur finalité d'intérêt général expertisée pour l'amélioration de la conduite des politiques publiques. En la matière, il est possible de capitaliser sur un certain nombre d'expériences :

- celle des autorités administratives indépendantes et des autorités de régulation qui ont été dotées par le législateur de prérogatives spécifiques de collecte de données (AMF, ACPR, CRE, ARJEL, ARCEP) ;
- celle de la statistique publique sur le fondement de l'article 19 de la loi pour une République numérique

- celle des collectivités territoriales ayant acquis des données auprès d'entreprises privées sur une base contractuelle, contre rémunération ou en échange de données publiques ou d'accès à des services

Après étude d'opportunité, si la puissance publique est contrainte de prévoir une législation particulière pour obliger la communication de données à destination d'une autorité administrative déterminée, ses **modalités d'application devront répondre à certaines questions** auxquelles la mission n'est pas en mesure d'apporter une réponse univoque tant il est nécessaire de les traiter *in concreto* (cf. partie 4.2).

Sur le plan de l'efficacité de l'action publique, l'approche incitative serait donc complétée au cas par cas par un instrument contraignant si elle échoue à atteindre certains objectifs d'intérêt général. Cela étant, les limites sont plus nombreuses et importantes que pour l'approche purement incitative. Un devoir de communication à la puissance publique **risque de désinciter les acteurs économiques à l'innovation et à l'investissement** dans des services basés sur l'utilisation de données. De plus, les contraintes techniques liées à la mise à disposition de données seront vraisemblablement plus **difficiles à supporter pour les petites entreprises**. En outre, une mise en œuvre sectorielle risque de se concentrer sur des acteurs nationaux qui sont déjà largement régulés (télécoms, transport, énergie, etc.).

Pour ces raisons, la mission estime nécessaire que la création de dispositifs coercitifs ne soit jamais envisagée comme un réflexe premier mais bien après motivation du besoin et analyse coût-avantages de l'obligation par rapport à l'incitation au regard de l'objectif d'intérêt général poursuivi.

> **Recommandation** : Privilégier une approche incitative et concertée, le recours à d'éventuels dispositifs coercitifs devant être dûment justifié et faire l'objet d'une évaluation préalable

Malgré les avantages de la progressivité de cette approche, qui fait de l'incitation la règle et de la coercition l'exception, permettant ainsi de s'adapter finement à la multiplicité des cas d'usage, la puissance publique peut également être confrontée à des circonstances exceptionnelles qui la rendrait moins efficace. Dans ces cas de figure très limités, où la situation est d'une gravité majeure et l'urgence absolue, il peut devenir impérieux que la puissance publique prenne des mesures d'exception.

Évaluer les pouvoirs de réquisition existants et l'opportunité de leur extension

La crise sanitaire constitue un exemple caractérisé de circonstances exceptionnelles pour lesquelles une approche progressive rencontrerait des limites en matière d'efficacité. La rapidité de progression d'une pandémie est en effet peu compatible avec une démarche progressive, volontaire et concertée, dont les résultats interviendraient trop tard.

Pour ne pas laisser la puissance publique démunie, la mission s'interroge sur la possibilité de mobiliser des pouvoirs de réquisition similaires à ceux de certaines autorités de régulation sectorielle, de manière ponctuelle et encadrée. Le sujet est particulièrement sensible, complexe et, s'il devait être tranché positivement, cet élargissement des pouvoirs de réquisition devrait s'accompagner de solides garanties. La mission est en effet particulièrement sensible aux craintes exprimées par les citoyens lors de la consultation en ligne qu'elle a conduite, ainsi qu'à la nécessité de développer un modèle éthique de l'utilisation des données, basé sur le respect des libertés et droits fondamentaux. Elle souligne à cet égard qu'un régime de réquisition élargi ne devrait pas avoir vocation à être pérennisé dans le droit commun mais devrait demeurer concomitant des circonstances exceptionnelles qu'il vise à affronter.

À droit constant, l'article L. 2215-1 du Code général des collectivités territoriales semble déjà ouvrir la voie à la réquisition de données, en convoquant la notion de « tout bien ou service », si un certain nombre de circonstances sont réunies et que des modalités strictes sont respectées.

Il dispose en effet qu' « en cas d'urgence, lorsque l'atteinte constatée ou prévisible au bon ordre, à la salubrité, à la tranquillité et à la sécurité publiques l'exige et que les moyens dont dispose le préfet ne permettent plus de poursuivre les objectifs pour lesquels il détient des pouvoirs de police, celui-ci peut, par arrêté motivé, pour toutes les communes du département ou plusieurs ou une seule d'entre elles, réquisitionner tout bien ou service, requérir toute personne nécessaire au fonctionnement de ce service ou à l'usage de ce bien et prescrire toute mesure utile jusqu'à ce que l'atteinte à l'ordre public ait pris fin ou que les conditions de son maintien soient assurées. »

Cette réquisition doit être précise, dûment motivée, limitée dans le temps et doit compenser les frais qui en découlent. En effet, « lorsque la prestation requise est de même nature que celles habituellement fournies à la clientèle, le montant de la rétribution est calculé d'après le prix commercial normal et licite de la prestation ».

Cet article est conforté par la loi n° 2020-290 du 23 mars 2020 d'urgence pour faire face à l'épidémie de covid-19 qui prévoit à son article 2 :

« Dans les circonscriptions territoriales où l'état d'urgence sanitaire est déclaré, le Premier ministre peut, par décret réglementaire pris sur le rapport du ministre chargé de la santé, aux seules fins de garantir la santé publique : […]

« 7° Ordonner la réquisition de tous biens et services nécessaires à la lutte contre la catastrophe sanitaire ».

En revanche, la généralité des termes retenus par la loi ne permet pas de conclure avec fermeté que les données des acteurs privés peuvent aujourd'hui entrer dans le champ de la réquisition. A cet égard, il convient de relever que ces dispositions n'ont, à la connaissance de la mission, jamais été actionnées pour permettre à la puissance publique d'accéder à des données privées pendant la crise sanitaire. **Compte tenu de ces difficultés d'appréciation, la mission considère que cet alinéa pourrait être complété par la réquisition de données** afin de sécuriser les demandes que la puissance publique pourrait formuler en cas d'urgence. Les dirigeants d'une entreprise doivent en effet respecter les règles de gouvernance de leur société sur la cession des stocks et des actifs, et le mécénat (don à l'État et ses établissements publics) est encadré par la loi.

Ces réquisitions devraient être motivées par la nécessité et respecter le principe de proportionnalité.

En première analyse, l'indemnisation pourrait être moins favorable que la jurisprudence existante sur la réquisition de biens et de services pour tenir compte de l'atteinte plus limitée au droit de propriété par le simple accès aux données (absence de perte du bien), à la manière d'une servitude de passage. Cet aspect mériterait une analyse juridique approfondie et devrait se confronter aux cas d'espèces pour être confirmée.

> **Recommandation** : Clarifier le régime juridique de la réquisition pour permettre à la puissance publique d'accéder à des données du secteur privé en cas de motif impérieux d'intérêt général et d'urgence

4.2. Une doctrine claire et définie en cas de partage de données obligatoire

Des principes à respecter[268]

L'obligation du partage de données ne doit pas conduire à la mise en place brutale de processus dis-proportionnés ou inappropriés. Au contraire, la puissance publique doit se fixer des grands principes, dont le respect donnera aux acteurs privés une prévisibilité et stabilité suffisantes pour intégrer les nou-velles obligations dans leurs modèles d'affaire en minimisant leur coût réglementaire. Ces principes visent à clarifier les sujets à traiter lorsqu'il s'agit de définir des règles d'accessibilité obligatoires et prendre les options utiles pour leur conception. L'analyse de la littérature en la matière et les précisions recueillies en audition ont conduit la mission à sélectionner les principes qui lui paraissent s'inscrire le plus harmonieu-sement dans son approche des données d'intérêt général.

D'abord, la mission souligne la pertinence des propositions formulées par la Commission européenne en matière de partage des données B2G, qui recommande de respecter les principes de **proportionnalité**, de **détermination des finalités**, de **non-préjudice** pour les personnes privées concernées, de **détermination de la compensation financière** en fonction de l'intérêt public poursuivi, de **transparence** sur ces partenariats et de **participation de la société civile**.

Plusieurs de ces principes ont déjà été abordés en filigrane de ce rapport, à l'instar des principes de **proportionnalité** et de **détermination des finalités** qui visent, pour le premier, à sécuriser l'intervention du législateur sur le plan constitutionnel et les personnes privées dans l'exercice de leurs droits économiques et, pour le second, à garantir que la décision est élaborée méthodiquement et dans le respect du RGPD. De même, l'attention au principe de **non-préjudice** conduit la mission à privilégier une approche progressive, qui fait de l'incitation la règle et de l'obligation l'exception.

Ensuite, la mission souhaite mettre l'accent sur la nécessité pour la puissance publique d'observer les trois principaux fondamentaux suivants de façon rigoureuse, pour susciter la confiance des acteurs privés dans les dispositifs de partage obligatoire de données :

- principe de **transparence** et de **redevabilité** : il s'agit pour la puissance publique, d'une part, de donner des gages et un accès facile aux informations concernant, par exemple, l'usage des données transmises, les services habilités à les traiter ainsi que les traitements effectués (dans le respect des secrets légaux et de la vie privée) ; d'autre part, il s'agit pour la puissance publique de rendre compte de son accès et utilisation des données auprès d'instances de contrôle tierces (régulateur, Parlement, etc.) ;
- principe de **justification de l'intervention** : au niveau économique et politique, il s'agit pour la puis-sance publique de démontrer la nécessité de recourir à l'obligation plutôt qu'à l'incitation, via l'objectivation des limites de l'approche incitative et l'évaluation a priori des avancées permises par la contrainte pour satisfaire le motif d'intérêt général poursuivi ;
- principe de **proximité** : les obligations visées doivent idéalement s'appuyer sur la préexistence d'une communauté de partage entre acteurs privés qui inclue l'État en tant qu'animateur ou acteur écono-mique, afin de minimiser les perturbations économiques potentielles.

[268] Ces principes sont notamment inspirés du rapport de Heiko Richter « The Law and Policy of Government Access to Private Sector Data ('B2G Data Sharing') », Max Planck Institute for Innovation and Competition Research Paper No. 20-06, Mai 2020

Questions clés pour fonder et organiser les règles d'accès

Il ne suffit pas à la puissance publique de structurer ses modes de décision selon des principes clairs pour garantir durablement la confiance dans son action et en maximiser l'efficacité. Le contenu même de ces décisions doit permettre de traiter un certain nombre d'interrogations qui se poseront dans toutes les situa-tions et pour lesquelles les réponses sont variables. Selon Heiko Richter[269], les cinq questions clés à se poser pour élaborer des règles d'accès obligatoire aux données privées sont les suivantes : Pour quoi ? Pour qui ? Qui est impacté ? Accès à quoi ? Comment ?

Les auditions conduites par la mission lui permettent de mettre en avant les orientations qu'il est judicieux de conserver à l'esprit au moment d'aborder certaines de ces questions :

- **pour quelle finalité d'intérêt général ?** Cette question constitue la clé de voûte du raisonnement et doit permettre, selon la mission, de décider si le recours à l'obligation est légitime. Ainsi qu'il a été évoqué précédemment, des droits économiques sont en jeu la finalité d'intérêt général doit être caractérisée de façon suffisamment précise pour fonder la constitutionnalité du dispositif. À cet égard, si des motifs tirés, par exemple, de l'efficacité de l'action publique en matière de régulation sectorielle, de préservation de la sécurité ou de la santé publiques semblent suffisamment précis, il en va différemment de catégories aussi larges que le soutien au développement économique et à l'innovation. Pour ces dernières catégories, la mission considère particulièrement opportun d'éviter autant que possible le recours à un dispositif coercitif ;
- **pour quel bénéficiaire ?** Pour des données sensibles et/ou protégées par des secrets légaux, l'ouverture et le repartage large des données est inopportun et il peut être pertinent, au contraire, de restreindre volontairement le nombre de personnes publiques et d'agents publics habilités à manipuler les données. En fonction du degré de sensibilité du projet et des données, il sera donc nécessaire de s'interroger sur l'opportunité de se contenter des exceptions existantes dans le CRPA ou bien prévoir des exceptions spécifiques pour sécuriser davantage les acteurs privés (capacité à garder un con-trôle sur les conditions de réutilisation), via l'éventuelle création d'un régime juridique spécifique (cf. partie 2.2).

Quels types de données ? Il s'agit ici d'identifier le niveau pertinent de traitement dans la chaîne de valeur (données brutes, données retraitées ; granularité, temporalité, fréquence). Les règles d'accès peuvent également porter sur les métadonnées et organiser la façon d'identifier la manière dont les données sont ré-parties au sein d'un écosystème ou d'une entreprise. Elles peuvent mandater l'acteur privé pour évaluer la qualité des données qu'il transmettrait. Enfin, elles peuvent être conçues selon une logique progressive : accès à l'information sur les jeux de données ; accès à des échantillons pour s'assurer de leur utilité au regard de l'objectif public ; accès au jeu de données pertinent pour l'utiliser à cette fin.

Enfin, il est nécessaire de s'interroger sur les modalités de mise à disposition des données, qui peuvent être d'ordre technique et économique.

Sur le plan **technique**, le législateur et/ou le régulateur peuvent définir des règles d'accès basées sur l'utilisation d'infrastructures ou protocoles existants, ou bien imposer la création de nouvelles infrastructures. L'accès signifie ici le transfert ou l'échange de données. En général, l'accès direct par l'État à des bases de données privées est proscrit pour des raisons de sécurité. La loi oblige plutôt l'entreprise à transmettre ses données, par exemple via une API, un tiers de confiance, la mise à disposition dans un entrepôt dédié. Il peut également exister des solutions techniques qui ne nécessitent pas d'extraction à proprement parler (CASD). Enfin, une réflexion doit être engagée sur les formats d'échange et les durées d'accès, au cas par cas, qui peuvent varier en fonction de considérations liées au respect du principe de proportionnalité et des règles de protection des données personnelles.

[269] Op. cit.

Enfin, la question **économique** de l'indemnisation de l'acteur privé est extrêmement sensible, comme cela est clairement apparu lors des auditions, et peut être résolue de différentes manières. Généralement, le niveau d'indemnisation peut être organisé selon la typologie suivante : gratuité ou absence d'indemnisation ; indemnisation au coût marginal ; indemnisation au coût moyen ; indemnisation au prix de marché. La détermination du bon niveau d'indemnisation pour l'acteur privé dépend d'un certain nombre de paramètres qui sont parfois très difficiles à articuler en pratique. L'indemnisation a notamment vocation à varier selon que l'accès demandé concerne des données brutes, des données retraitées, une offre de service basée sur l'exploitation de données, etc. Elle doit également tenir compte des coûts de mise en œuvre de l'accès que les acteurs public et privé supporteront respectivement. Enfin, elle peut présenter un lien avec la nature du motif d'intérêt général poursuivi et l'importance pour la puissance publique d'accéder aux données, comme le propose notamment le groupe d'experts mandaté par la Commission européenne sous forme d'arbre décisionnel (cf. encadré).

Analyse du groupe d'experts mandaté par la Commission européenne

Le groupe d'experts mandaté par la Commission européenne propose un arbre décisionnel pour déterminer les étapes à suivre dans le cadre d'un partage de données en B2G[270].

La première étape consiste à déterminer le problème et à s'interroger s'il y a un objectif d'intérêt général. Si ce n'est pas le cas, il n'y a pas lieu d'envisager un transfert. Si c'est le cas, plusieurs autres étapes doivent être envisagées dans le raisonnement, dans l'ordre suivant, pour déterminer l'opportunité du partage :

- le secteur public a-t-il besoin de données du secteur privé pour atteindre l'objectif d'intérêt général visé, de manière efficace et efficiente ?
- quels sont les critères qui doivent déterminer les conditions de réutilisation ? il peut s'agir de la probabilité des bénéfices, de l'intensité des bénéfices probables, la probabilité des dommages, l'intensité des dommages probables, l'immédiateté et l'urgence de la situation, le potentiel dommage de ne pas avoir recours aux données ;
- quatre situations peuvent être déterminées à partir de là : i) des conditions coercitives et gratuites (par exemple après un désastre naturel) ; ii) des conditions coercitives avec une compensation au coût marginal (changement climatique) ; iii) des conditions coercitives mais avec une compensation négociable (par exemple en matière d'urbanisme) ; iv) des conditions de volontariat (par exemple dans le tourisme).

À partir de là, le partenariat de partage en B2G peut être établi et les modalités d'accès définies (modèle opérationnel, moyens techniques). Ce partenariat permet de conduire une politique publique fondée sur des preuves (*evidence-based policymaking*) et d'augmenter la qualité du service public rendu. Il implique un principe de transparence et de responsabilité sur les usages qui sont fait des données, auprès des parties prenantes, y compris des citoyens.

Dans l'ensemble de cette chaîne, les principes suivants doivent s'appliquer : proportionnalité, limitation de l'usage des données, limitation des risques et garde-fous, compensation, absence de discrimination, atténuation des limitations des données du secteur privé, transparence et participation de la société civile, responsabilité, usage éthique de la donnée.

[270] « Towards a European strategy on business-to-government data sharing for the public interest / Final report prepared by the High-Level Expert Group on Business-to-Government Data Sharing », Publications Office of the European Union, 2020, p. 48.

CAS D'USAGE – Les données du secteur privé utilisées par la statistique publique

Les données massives (Big Data) issues du secteur privé constituent une source nouvelle et complémentaire pour la statistique publique. De longue date, les services statistiques ont utilisé des données issues des déclarations des entreprises pour mener des enquêtes et produire des indicateurs, par exemple sur les niveaux d'emploi dans chaque secteur d'activité. Les données dont il est ici question sont différentes, dans la mesure où elles sont le plus souvent produites par des entreprises dans le cadre de leur activité et ne font pas l'objet d'obligation de déclaration. On pense notamment aux données de transactions des cartes bancaires ou encore aux données issues des réseaux de téléphonie mobile.

À l'occasion de la pandémie, de nombreuses collaborations ont pu voir le jour entre des administrations publiques et des acteurs privés, à l'image de l'INSEE qui a pu accéder à des données des opérateurs de téléphonie mobile afin d'évaluer les déplacements de population lors de la première période de confinement. Cela a permis de confirmer que des données du secteur privé pourraient utilement informer les décideurs publics. Ces collaborations, qui ont vu le jour dans plusieurs pays européens, posent de nombreuses questions sur le cadre encadrant l'accès et l'utilisation, par la puissance publique, de données produites par le secteur public.

La loi pour une République numérique a posé des bases juridiques à l'utilisation de données du secteur privé à des fins de statistique publique

L'article 19 de la loi pour une République numérique (2016) donne un cadre juridique à l'accès et à l'utilisation à des fins de statistique publique des données produites par le secteur privé. Cet article prévoit notamment que le Ministre chargé de l'économie peut réclamer l'accès à certaines données nécessaires à la statistique publique.

La procédure prévue par le législateur comporte de nombreux garde-fous: les données ainsi collectées ne peuvent faire l'objet d'aucune autre finalité que celle définie initialement, elles ne peuvent pas être partagées avec des tiers même au sein de l'administration (pas d'open data ni de circulation des données), elles ne doivent pas permettre l'identification des individus. De même, une concertation avec les acteurs concernés doit être organisée et une enquête d'opportunité et de faisabilité doit être produite et rendue publique. L'ensemble du dispositif (type de données collectées, modalités de collecte, de traitement et le cas échéant de destruction des données collectées) fait l'objet d'un texte réglementaire (arrêté).

Il convient de noter que cet article 19 est envié par de nombreux instituts nationaux de statistiques en Europe qui ne disposent pas d'une telle faculté.

Une première mise en oeuvre pour la mesure de l'inflation

La première mise en oeuvre de cette disposition introduite par la loi de 2016 concerne la construction de l'Indice des prix à la consommation (IPC) qui constitue la mesure de l'inflation par l'INSEE. La statistique publique, tant au niveau national qu'européen, a depuis longtemps identifié les données de transactions de la grande distribution comme une source potentielle et complémentaire pour le calcul de l'IPC. Depuis janvier 2020, l'Indice des prix à la consommation intègre ainsi les données de caisse des enseignes de grande distribution. L'accès à ces données a fait l'objet d'un arrêté paru en avril 2017, accompagné d'une étude d'opportunité et de faisabilité. Il convient de s'arrêter un moment sur ce premier exemple pour bien comprendre l'intérêt et les limites des dispositifs juridiques existants. Le cas d'usage des données de caisse est déjà relativement bien connu de l'INSEE lors de l'adoption de la loi pour une République numérique en 2016. Plusieurs années auparavant, quelques enseignes ont contractualisé avec l'institut national

statistique pour permettre de valider l'intérêt de cette source d'information pour l'IPC. C'est un point-clé en matière de fiabilité des résultats: la statistique publique doit s'assurer de bien comprendre non seulement les données elles-mêmes, mais aussi et surtout de maîtriser le contexte et les modalités de leur production. Il en va de la qualité de l'information statistique qu'elle produit in fine et, en dernier ressort, de la souveraineté de la décision publique. On verra plus tard avec l'exemple des données des opérateurs de télécommunications, que le degré de transparence ainsi requise sur la production des données peut constituer une barrière à la généralisation de l'usage des données du secteur privé à des fins de statistique publique.

L'autre enseignement de l'utilisation des données de caisse est que la démarche prend du temps, à chacune des étapes. Il y a d'abord le temps de l'expérimentation (en l'occurrence avec des enseignes volontaires et sur une base contractuelle), celui de la loi (le suivi de la procédure telle que définie par l'article 19) et enfin celui de la mise en oeuvre opérationnelle.

Pour un sujet relativement mature comme celui des données de caisse, et malgré toute l'expertise préalable du sujet par l'institut national statistique, il faut ainsi compter plus de 3 ans entre l'adoption de la loi pour une République et la première mise en œuvre concrète sur un périmètre bien défini.

La crise sanitaire a révélé le potentiel des données du secteur privé pour la conduite des politiques publiques … mais aussi les tensions existantes concernant les modalités d'accès à ces données

La capacité à évaluer et mesurer la population présente dans chaque territoire avant, pendant et après les périodes de confinement constitue un élément essentiel en matière de santé publique, notamment pour dimensionner l'offre de soins dans chaque région. Dès le 8 avril 2020, soit moins de 3 semaines après la mise en place du premier confinement, l'INSEE a été en mesure de fournir de premiers résultats à partir de l'analyse des données de téléphonie mobile fournies par l'opérateur Orange. "Ces données sont des statistiques de comptage, agrégées territorialement, collectées au niveau des antennes relais. Ce ne sont ni des données GPS de localisation des téléphones portables, ni des données issues d'application téléchargées. Elles ne permettent pas de suivre le déplacement des personnes, mais d'effectuer des comptages par zones à différentes dates, ce qui est l'unique objectif de ces travaux" explique l'INSEE. Ces travaux ont ensuite été poursuivi par l'institut national statistique, en utilisant notamment des données fournies par deux autres opérateurs opérant sur le territoire national (Bouygues et SFR).

Les données des cartes bancaires sont aussi utilisées pour évaluer notamment les niveaux de transaction dans certains secteurs d'activité durement impactés par la crise (notamment la restauration). La Direction générale du Trésor a ainsi accédé à des données du Groupement Cartes bancaires. Il est important de souligner que l'accès à ces données a le plus souvent été réalisé sur une base volontaire et bénévole (pro-bono) par les entreprises concernées.

Enfin, on peut noter que la situation n'est pas propre à la France, car de nombreuses initiatives comparables ont vu le jour, en Europe et dans le Monde.

La crise sanitaire agit aussi comme un révélateur des tensions existantes. En effet, l'accès à ces données s'est fait bien souvent avec une grande réactivité, sur une base volontaire et dans tous les cas hors du cadre juridique posé par l'article 19 de la Loi pour une République numérique. L'urgence de la crise, sa brutalité ont entraîné une mobilisation de l'ensemble du pays. Cette mobilisation des acteurs publics et privés a permis de donner vie à ces initiatives d'accès à des données du secteur privé. Il en relève aussi les tensions.

Protection des investissements réalisés par les entreprises pour valoriser les données vs. accès aux données brutes par la statistique publique

Le cas de l'opérateur de téléphonie mobile Orange illustre la tension entre la volonté des entreprises de valoriser les données de leur activité et celle de la puissance publique de pouvoir accéder à des données utiles pour la décision publique. Orange propose, via son service Flux Vision l'accès à des données agrégées, anonymisées, permettant d'évaluer notamment la population présente à un

instant donnée dans un lieu. Il s'agit par exemple de mieux connaître les flux de visiteurs sur un site touristique, afin d'en comprendre le profil et le comportement. Flux Vision est aujourd'hui une offre commercialisée dans plusieurs pays européens. Orange a consenti depuis plusieurs années des investissements notamment pour anonymiser, traiter et transformer les données issues du réseau mobile en informations utiles pour la prise de décision des clients de Flux Vision. La question posée aujourd'hui est du cadre applicable au-delà de la période actuelle. L'opérateur fait valoir que d'autres instituts statistiques nationaux sont déjà des clients de l'offre Flux Vision. L'INE (institut statistique national) a ainsi déboursé près d'un demi-million d'euros pour accéder des données des trois opérateurs présents sur le territoire espagnol (Telefonica, Vodafone et Orange). A ce titre, Orange estime que l'institut national statistique français doit à l'avenir payer pour accéder à l'offre Flux Vision, comme tous les autres clients publics de la plateforme (et en particulier les collectivités territoriales). Selon l'opérateur, il est impossible de donner accès aux données brutes (et non au service d'analyse de ces données retraitées pour Flux Vision) pour des raisons de complexité des données (adhérence des données au métier) et de protection de la vie privée. Or, comme le montre l'exemple de l'Indice des prix à la consommation, un certain niveau de transparence est requis pour permettre un usage critique et raisonné de ces données à des fins de statistique publique. De même, la question de l'effectivité, dans le cas d'espèce, de l'article 19 de la Loi pour une République numérique est posée.

CAS D'USAGE – Le secteur des assurances et le fichier des véhicules assurés (FVA)

Une voiture sur la route qui n'est pas assurée représente un danger pour la société, car « les conducteurs non assurés se révèlent plus dangereux que les autres, en particulier parce qu'ils cumulent les infractions » . Ce danger est croissant (8711 victimes corporelles blessées ou décédées, en hausse de 6,6% depuis 2014, alors que nombre des accidents corporels recensés par l'Observatoire National Interministériel de la Sécurité Routière a baissé de 2,8% sur la même période) et coûteux (116 millions d'euros d'indemnités en 2019) . C'est le Fonds de Garantie des Assurances Obligatoires de dommages (FGAO) qui indemnise les victimes d'accidents de la circulation provoqués par des personnes non assurées ou non identifiées, il est financé par l'ensemble des assurances et des assurés.

Le FVA est un outil d'identification de l'assureur d'un véhicule impliqué dans un accident, mais aussi de lutte contre la non assurance, c'est un archétype du progrès par la donnée d'intérêt général. Les assureurs mettent en commun leurs données dans une base, grâce à un tiers de confiance, respectueux des règles de la concurrence, pour le bien commun. Les forces de l'ordre qui procèdent aux contrôles routiers pourront accéder à ces données et lutter plus efficacement contre la non assurance.

Le cadre européen pour améliorer l'indemnisation des accidents de la route

La Directive 2000/26/CE du Parlement européen et du Conseil du 16 mai 2000 concernant le rapprochement des législations des États membres relatives à l'assurance de la responsabilité civile résultant de la circulation des véhicules automoteurs et modifiant les directives 73/239/CEE et 88/357/CEE du Conseil, plus communément désignée comme la quatrième directive automobile, vise notamment à améliorer l'indemnisation des victimes d'accidents de la circulation survenus à l'étranger avec un Organisme d'information chargé d'indiquer l'assureur du véhicule responsable d'un accident. La pratique française est alors l'échange de données plutôt que la constitution d'un fichier. L'Association pour la Gestion des Informations sur le Risque en Assurance (AGIRA)[271] a été désignée comme l'Organisme d'information par un arrêté du 13 janvier 2004.

La route jusqu'au déploiement

En octobre 2010, François WERNER avance dans son rapport à la Ministre de l'Économie que 1 à 2 % des véhicules roulants ne sont pas assurés et estime alors le coût pour Fonds de garantie à 80 millions d'euros. Il propose « un répertoire des assurances automobiles, croisé avec le répertoire des immatriculations et qui constituerait un outil efficace en matière de sécurité routière afin d'empêcher les plus dangereux de prendre le volant, à l'image de ce qui se fait déjà chez nos voisins européens tels que l'Angleterre, l'Allemagne ou l'Italie »[272].

En mars 2016, l'intérêt du partage des données des assureurs est bien identifié dans les travaux de la mission sur les données d'intérêt général. Cette mission souligne à l'époque que l'indemnisation par le FGAO a connu une hausse de 28,4% entre 2008 et 2013. La création du FVA avait d'ailleurs

[271] http://www.agira.asso.fr/content/qui-sommes-nous

[272] https://www.argusdelassurance.com/a-la-une/automobile-les-propositions-du-rapport-werner-fgao-pour-lutter-contre-le-defaut-d-assurance.51819

été annoncée le 26 janvier 2015 à l'occasion du Plan national de sécurité routière. Effectivement, les discussions entre les parties prenantes (assurances et DCSR) sont constructives.

La loi n° 2016-1547 du 18 novembre 2016 de modernisation de la justice du XXIème siècle[273] « a pour ambition d'améliorer pour tous la justice du quotidien en la rendant plus proche, plus simple et plus efficace »[274]. Elle vise notamment à accélérer et renforcer la répression des délits routiers avec une sanction plus rapide et plus sévère, tout en désengorgeant les tribunaux (sanction à partir d'un contrôle automatisé ou vidéo, sanction du délit de conduite sans assurance par une peine forfaitisée). Son article 35 modifie l'article L451-1 du code des assurances et prévoit la mise en place d'un fichier des véhicules terrestres à moteur assurés et d'un fichier des véhicules susceptibles de ne pas satisfaire à l'obligation d'assurance. Ce dernier est consultable par les forces de l'ordre pour contrôler l'obligation d'assurance et par le fonds de garantie des assurances obligatoires. Cette loi modifie aussi l'article L. 233-2 du code de la sécurité intérieure qui autorise ainsi le traitement automatisé de données relatives à l'assurance des véhicules.

Le 1er janvier 2019, le FVA est mis en place. Une expérimentation des forces de l'ordre est en cours sur le « nouvel équipement opérationnel » (NEO), qui permet aussi par exemple d'interroger le fichier des objets et des véhicules volés (FOVeS) lors des contrôles dits « bord de route ». L'AGIRA a mis en place une infrastructure et des serveurs dédiés et cryptés (duplication de la base FVA) pour une mise à disposition gracieuse et sécurisée aux forces de l'ordre. L'intégration et le déploiement devraient être effectifs vers la fin de l'année 2020. Les conducteurs de véhicules non assurés pourront alors être verbalisés et, le cas échéant, la voiture immobilisée, après une vérification informatique ; au contraire, la bonne foi des assurés n'étant pas en mesure de présenter leur certificat d'assurance sera étayée.

L'AGIRA est une association loi 1901 créée par la Fédération Française de l'Assurance (FFA). L'AGIRA regroupe les sociétés d'assurance exerçant sur le marché français et les organisations professionnelles intervenant dans le secteur. L'AGIRA exerce d'autres fonctions (médiation de l'assurance par exemple) et gère d'autres fichiers (Fichier des Victimes Indemnisées par exemple), répondant ainsi à diverses dispositions du code des assurances et aux besoins de la profession.

L'exigence soutenue et partagée sur la qualité et la gestion des données

Le Ministère de l'Intérieur attend une qualité des données à la hauteur de l'enjeu. Le délai de 72 h prévu par le décret d'application pour transmettre les modifications au FVA a connu des retards se comptant en semaines, compromettant ainsi la valeur du contrôle (si le véhicule n'est en fait plus assuré, mais que la vignette est toujours apposée sur le pare-brise) et la bonne foi des assurés (si la souscription de l'assurance du véhicule n'est pas encore enregistrée dans le FVA).

Les flottes professionnelles des loueurs et des entreprises sont assurés au titre de contrats de groupes commercialisés par les courtiers et les assureurs, qui pouvaient ne pas avoir une connaissance parfaite en temps réel de l'évolution de ces flottes. L'immatriculation des véhicules qui entrent et sortent des flottes n'étant pas enregistrée au moment exact de ce mouvement (ni même la marque et le modèle), la couverture de l'assurance reposant sur la seule composition du parc par catégories. Or ces véhicules représentent une part significative du parc en France (environ 2,9 millions de véhicules sont détenus par des personnes morales[275] sur 38,2 millions de voitures en circulation en France[276]).

[273] NOR JUSX1515639L

[274] https://www.gouvernement.fr/action/la-justice-du-21e-siecle

[275] https://www.ecologie.gouv.fr/sites/default/files/Th%C3%A9ma%20-%20Les%20flottes%20de%20v%C3%A9hicules%20des%20personnes%20morales.pdf rapport d'août 2019 du CGDD

[276] https://www.statistiques.developpement-durable.gouv.fr/382-millions-de-voitures-en-circulation-en-france?list-chiffres=true

Tous les acteurs de la filière (constructeurs, concessionnaires, garages, courtiers d'assurances, assureurs, etc.) doivent donc continuer de renforcer leur organisation et accélérer leurs processus pour respecter ce délai de 72h. Cette exigence de qualité est dans l'intérêt des assureurs qui s'efforcent de ne pas décevoir leurs clients (sur qui pèse un risque d'amende et d'immobilisation du véhicule) et d'éviter tout risque pour leur réputation.

À long terme, des sanctions pour non assurance pourraient intervenir dans le cadre des contrôles automatisés pour excès de vitesse (ANTAI), sous réserve de conforter la procédure (vérification de l'historique et la cohérence) et de résoudre les difficultés techniques. La volumétrie des requêtes du FVA pourrait connaitre une croissance forte avec Lecture Automatique de plaques d'immatriculation (LAPI).

CAS D'USAGE - Bilan de l'application de la législation en matière de données de mobilité

Le modèle historique de la mobilité a changé avec l'émergence de solutions concurrentes pour l'accès à la mobilité, les applications de covoiturage, de VTC ou de billettique. La Direction Générale des Infrastructures, des Transports et de la Mer (DGITM) accompagne la transformation numérique des transports de voyageurs et de marchandises pour répondre aux enjeux de mobilités durables et innovantes. La concertation avec les parties prenantes est cruciale et les discussions sont essentielles pour établir un modèle équilibré, permettant de concilier l'intérêt général et l'émergence de nouveaux services.

L'ouverture des données, engagée avec la loi République numérique, porte sur les données qui ont vocation à être diffusées largement, comme les données du réseau routier national (data.gouv.fr et le site internet de Bison Futé, qui est le point d'accès national aux données routières). La Loi Macron avait posé des obligations de publication (données numériques « relatives aux arrêts, aux tarifs publics, aux horaires planifiés et en temps réel, à l'accessibilité aux personnes handicapées, à la disponibilité des services, ainsi qu'aux incidents constatés sur le réseau et à la fourniture des services de mobilité et de transport »), mais tous les décrets en Conseil d'État n'avaient pas été pris.

L'ouverture des données s'est accélérée avec la Loi d'Orientation des Mobilités (LOM), qui transpose le règlement délégué européen 2017/1926 du 31 mai 2017, avec la décision pionnière d'intégrer les données dynamiques. Les données « statiques » concernent la description de l'infrastructure ou des lignes de transport et leurs caractéristiques, les données « dynamiques » concernent leur fréquentation ou leur régularité. La RATP a investi environ 1 million d'euros pour la mise à disposition de ces données et déclare ne pas profiter de retour sur cet investissement alors que d'autres acteurs en font une utilisation commerciale.

En outre, la LOM prévoit le partage de données privées, comme celle des constructeurs et équipementiers des véhicules terrestres connectés, des gestionnaires de parking, des services de covoiturage ou des compagnies de transports ferroviaires, à des fins d'amélioration de l'offre et des services de transport (y compris l'intermodalité), de statistique publique (y compris la mesure des effets sur l'environnement), de régulation (Autorité de Régulation des Transports, ART), de dématérialisation et de simplification des procédures administratives pour les professionnels. L'autorité organisatrice de la mobilité anime le partage de données sous le contrôle de l'ART, qui dispose de moyen d'action et de sanction.

La géolocalisation et la disponibilité des taxis et des VTC

Pour les transports publics particuliers de personnes, que sont les taxis et les VTC, l'article 46 de la LOM prévoit la dématérialisation des procédures administratives au 1er janvier 2022, leur ouverture après anonymisation et structuration sera étudiée à partir de 2022.

Le projet Le.Taxi vise à doter la France d'un registre national de géolocalisation et de disponibilité des taxis. Le but est de fédérer l'offre de taxis disponibles et de développer la pratique de la maraude électronique. À partir du 26 décembre 2020, tous les taxis en activité devront se connecter à une plateforme de collecte des données de géolocalisation en temps réel, plateforme accessibles aux seuls services de mise en relation agréés. La réflexion sur l'ouverture de certaines données sera engagé en 2021, en concertation avec les acteurs, pour une ouverture des données en 2022 (ou avant si le nombre de taxis connectés est suffisant).

Le projet d'API « stations de taxi » est conduit en lien avec le point d'accès national pour la diffusion d'au moins deux jeux de données : un jeu statique avec la position des stations au niveau national, un jeu temps réel avec les taxis qui sont disponibles.

Le partage volontaire des données du covoiturage

Les Assises nationales de la mobilité et les travaux menés par la Fabrique des Mobilités soutenue par l'Ademe ont préconisé la création d'un « registre de preuve de covoiturage », plateforme numérique permettant à l'ensemble des opérateurs de covoiturage, volontaires, de faire converger et d'attester les trajets effectués par les utilisateurs. Ce registre donne la possibilité aux collectivités et aux entreprises de distribuer des incitations (monétaires ou non monétaires, comme des avantages en matière de stationnement), pour favoriser le covoiturage.

Ce « registre de preuve de covoiturage » entre en application de plusieurs mesures de la LOM, dont l'article 35 (versement d'allocation par les AOM pour des trajets effectués en covoiturage) ou l'article 82 (justificatifs pour le forfait mobilités durables). Depuis le 1er septembre 2019, les collectivités peuvent ainsi avoir accès à l'ensemble des trajets réalisés en covoiturage sur leur territoire, effectués via les opérateurs liés au registre, dont BlaBlaLines, Karos, Klaxit. Le modèle d'affaires fondé sur la mise en relation n'est donc pas mis en cause par un partage trop large qui aurait pu favoriser la domination progressive du marché par un seul acteur. Un groupe de travail (DGITM, ADEME, GART, CEREMA, DINUM) sera chargé de définir les indicateurs pertinents pour le suivi l'évolution des pratiques et l'évaluation des mesures prises. Un observatoire pourra s'appuyer sur les données du registre de preuve de covoiturage pour évaluer le développement du covoiturage par territoire ou par type d'incitation.

L'exploitation délicate du gisement des données des voitures connectées

L'article 32 de la LOM prévoit la fourniture des données produites par les véhicules connectés (par les systèmes intégrés aux véhicules terrestres à moteur équipés de moyens de communication permettant d'échanger ces données avec l'extérieur du véhicule) à des acteurs ciblés (gestionnaires d'infrastructures routières, autorités organisatrices de la mobilité, forces de l'ordre, services d'incendie et de secours, assureurs, constructeurs…). Les modalités seront précisées par voie d'ordonnance d'ici fin 2020, donnant un cadre règlementaire aux collectes de données. Il s'agit d'organiser le délicat partage des données de ce gisement de données considérable, qui attise la convoitise de nombreux acteurs privés et publics.

La discussion sur le partage des données du transport ferroviaire

L'article L. 2131-1 du code des transports prévoit que l'ART « concourt au suivi et au bon fonctionnement, dans ses dimensions techniques, économiques et financières, du système de transport ferroviaire national, notamment du service public et des activités concurrentielles, au bénéfice des usagers et clients des services de transport ferroviaire. Elle exerce ses missions en veillant au respect de la loi n°2010-788 du 12 juillet 2010 portant engagement national pour l'environnement, notamment des objectifs et dispositions visant à favoriser le développement des modes alternatifs à la route pour le transport de marchandises »[277].

Pour être en mesure d'assurer les missions qui lui sont attribuées, l'ART doit disposer d'informations fiables, précises et détaillées (par zone géographique, par type d'activité et de trafic, par entreprise). Le recueil des informations fait l'objet d'une décision. En application des décisions de l'ART du 10 mai 2017, du 5 juillet 2017 et du 11 décembre 2017 et du 11 avril 2019, l'Observatoire des transports et de la mobilité est chargé de collecter des informations à un rythme trimestriel, semestriel ou annuel, de les conserver, les traiter et les utiliser à des fins d'analyses économiques et statistiques. Ces analyses peuvent donner lieu à des publications, dans le respect du secret des affaires.

[277] https://www.legifrance.gouv.fr/codes/id/LEGIARTI000038884827/2019-10-01/

Les nouveaux services attendent des données dans l'ensemble des pays où ils souhaitent opérer, à une échelle européenne voire mondiale, en particulier pour la construction d'itinéraires internationaux ou l'historique de circulation des trains. Les données déjà ouvertes ne le sont pas nécessairement dans des formats identiques d'un pays à l'autre, ce qui montre que l'ouverture des données peut se heurter au manque de standardisation. En outre, le partage des données en temps réel est crucial pour les nouveaux acteurs, par exemple sur les perturbations du trafic dont la complétude ne serait pas assurée en France selon Trainline. Les attentes sur les modalités de gestion des données sur les retards et les compensations sont aussi discutées, les interprétations de la LOM divergent sur certains points opérationnels comme la programmation des travaux.

Mobility as a Service (Maas)

Les articles 25, 27 et 28 de la LOM visent à permettre la mise en place de « services numériques multimodaux » qui assurent la planification des trajets, la réservation, le paiement et la délivrance du titre de transport. Ce type de service est souvent désigné sous le terme de Mobility as a Service (Maas). Les articles L 1115-8 à 12 du Code des transports prévoient que les autorités organisatrices des mobilités veillent à l'existence d'un service d'information multimodale. Les services de transport doivent ouvrir l'accès de leur service numérique de vente aux opérateurs de billettique multimodale. Ces dispositions sont destinées à faciliter la mise en place de services dits Mobility as a Service (MaaS), qui combinent l'information et la billettique multimodale. La Commission européenne a prévu une initiative sur la billettique dans le cadre de la révision de la directive « Intelligent Transport Services », elle examine en particulier l'accès non discriminatoire des voyageurs ferroviaires aux données nécessaires pour planifier des voyages et acheter des billets. Elle présentera un rapport au Parlement européen et au Conseil sur la disponibilité de tels systèmes communs d'information et de billetterie directe, assorti, au besoin, de propositions législatives.

Un des enjeux est le droit des autorités publiques et des opérateurs délégataires d'opérer des services dits MaaS et vendre les billets du transport public. Le rôle des autorités organisatrices de mobilité est discuté, un équilibre doit être trouvé entre les politiques publiques de mobilité et de décarbonation des transports et l'ouverture à la concurrence des acteurs MaaS. Le point d'équilibre tient aussi aux licences de réutilisation des données de mobilités et aux contraintes qu'elles emportent (identification du réutilisateur et absence de limite inutile imposée aux possibilités de réutilisation notamment). Certains acteurs comme la Métropole de Lyon s'inquiètent de la capture des données par des acteurs monopolistique et s'interrogent sur l'absence de dialogue avec les réutilisateurs, y compris sur la réutilisation des données dans un sens divergent de l'intérêt général, par exemple avec le déport de trafic routier sur des voies protégées à proximité des écoles ou des hôpitaux. Des producteurs et réutilisateurs de données travaillent actuellement à la préparation qu'ils qualifient de « licence d'intérêt général », pour que la réutilisation des données de mobilité soit conforme avec les politiques publiques (comme la sécurité ou l'environnement) et qu'un certain niveau de partage des enrichissements par les réutilisateurs, nécessairement identifiés, soit garanti.

Comment ça se passe à l'étranger ?

À Taïwan[278], en 2015, une start-up appelée « TMS Technologies Corporation» a développé la plateforme DATABAR en intégrant des données de transport, comme la géolocalisation des flottes publiques ou l'emplacement des parkings privés et des stations-service.

À Tokyo[279], la multiplicité des opérateurs de transport a conduit au partage de données entre concurrents, notamment pour répondre à des objectifs de politique publique, sans rediffusion

[278] Source : DGT, Service économique de Taipei

[279] Source : audition Benjamin Jean

simple au-delà du cercle de ces opérateurs. Mais la publication de données émanant des entreprises commence avec des informations sur les horaires de passage et la position des bus en temps réel.[280]

Dans l'Euro-région Meuse-Rhin[281] (Aix-La-Chapelle – Maastricht – Hasselt – Liège), une plateforme MaaS partage des données provenant d'opérateurs privés, d'autorités publiques et de citoyens (partage de leur historique de déplacements). Le but est de faciliter la circulation entre l'Allemagne, les Pays-Bas, le Luxembourg et la Belgique avec une application multimodale et des services et des notifications.

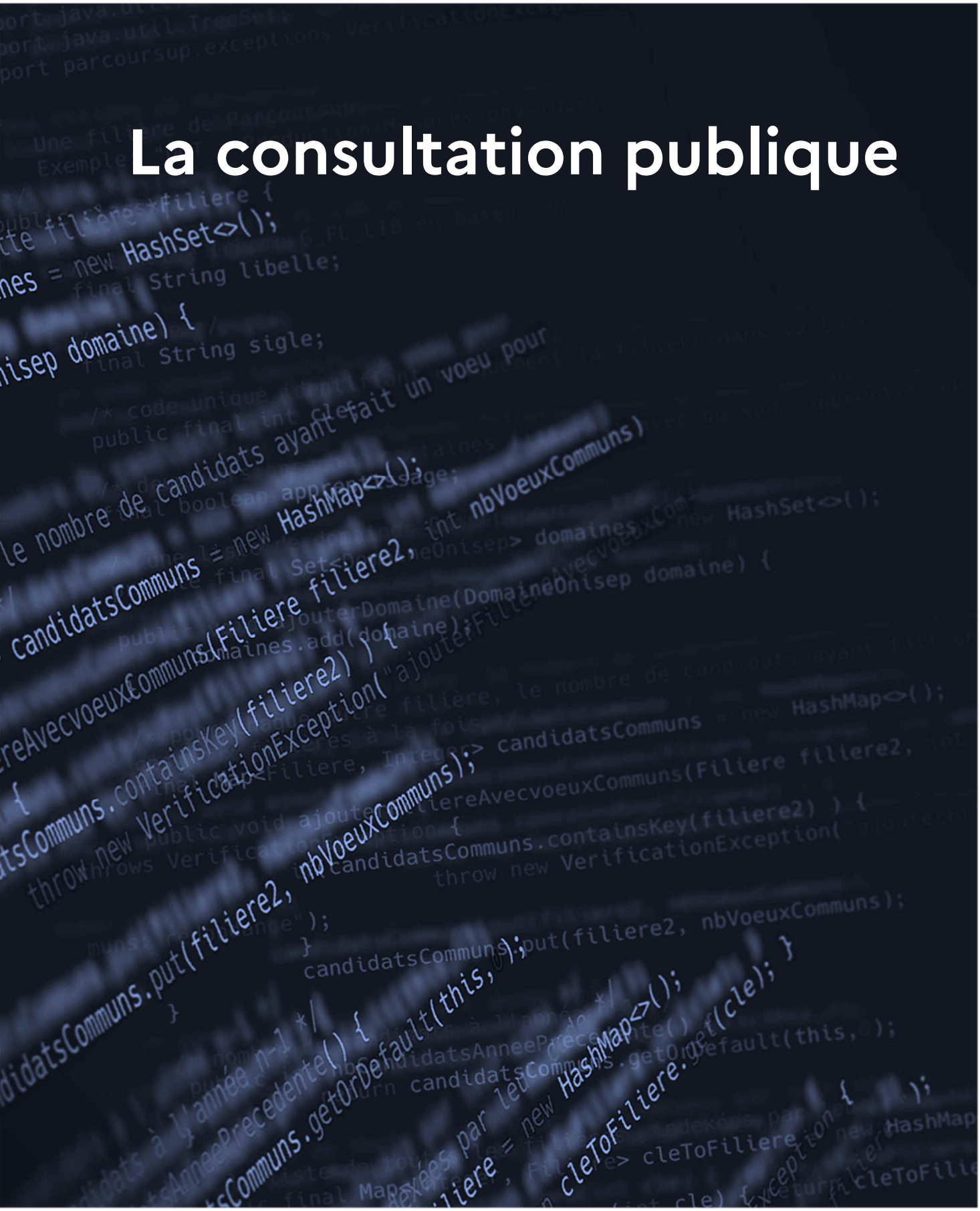
Aux États-Unis[282], le Bureau of Transportation Statistics (BTS) a lancé un programme sur la mise en commun de certains jeux de données des compagnies aériennes sur la fréquentation de lignes de vol domestiques. Après traitement statistiques, les compagnies aériennes peuvent utiliser la base constituée pour construire leurs stratégies.

[280] Source : DGT, Service économique de Tokyo

[281] Source : « Partage des données personnelles changer la donne par la gouvernance », Matthias de Bièvre et Olivier Dion, www.thedigitalnewdeal.org , juillet 2020

[282] Source : « États généraux du numérique, synthèse de la consultation, Données d'intérêt général », mai 2020

La consultation publique



Contexte de la consultation

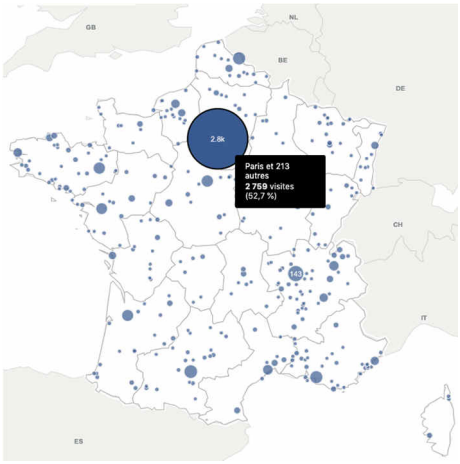
La mission a conduit une consultation entre le 8 octobre et le 9 novembre 2020, sur un site Internet accessible à tous (*www.mission-open-data.fr*), pour recueillir de la part de citoyens et d'organisations i) des réactions à son rapport d'étape, mis en ligne sur cette plateforme et présentant les premiers constats de la mission, ii) des propositions d'action pour répondre à ces problèmes identifiés, iii) des réactions à dix situations (« cas d'usage ») où la puissance publique pourrait être amenée à utiliser des données d'acteurs privés.

La consultation s'est déroulée en deux étapes, d'abord par le recueil des réactions aux dix principaux constats du rapport d'étape et des contributions libres, ensuite par le recueil des contributions aux cas d'usage.

Les chiffres de la participation

En l'espace de 33 jours de consultation :

- **545 comptes utilisateurs** validés ont été créés sur le site ;
- **108 contributions libres** ont été formulées dans la partie dédiée aux recommandations ;
- **418 commentaires** sur les contributions ont été publiés ;
- **1 753 soutiens** aux contributions ont été recueillis ;
- **5 954 visites** ont été enregistrées au total sur le site, soit **180 visites par jour** en moyenne.



Les consultations du site se répartissent dans l'ensemble du territoire français, même si la région Île-de-France concentre 52,7 % des visites (chiffres estimés d'après les informations de localisation des utilisateurs lorsqu'ils se connectent, qui peuvent avoir des biais liés aux infrastructures de connexion). À l'étranger, le site a été consulté aux États-Unis (346 vues), en Allemagne (36 vues) et en Belgique (31 vues) notamment.

Parmi les constats du rapport d'étape de la mission, les plus soutenus ont été :

- le manque de culture et de compétences limite la libération du potentiel de la donnée et des codes sources (50 soutiens) ;
- l'ouverture est perçue comme un danger pour les acteurs publics, à tort ou à raison (46 soutiens) ;
- le manque de portage managérial et le besoin d'une formation des cadres dirigeants aux enjeux de la donnée (40 soutiens).

Les contributions

La consultation publique a fait ressortir cinq préoccupations principales.

Premièrement, la consultation a fait ressortir une forte mobilisation autour du logiciel libre : les promoteurs d'une action plus ferme de l'État en faveur du logiciel libre se sont très largement mobilisés, à l'invitation notamment de l'April et du CNLL. Leurs propositions concernent notamment la priorité à accorder au logiciel libre, le manque d'effectivité de l'article 16 de la loi pour une République numérique ou encore l'application du référentiel général d'interopérabilité.

La thématique a ainsi recueilli les soutiens les plus nombreux aux contributions libres (recommandations suggérées par les contributeurs) :

- Priorité au logiciel libre et aux formats ouverts dans le secteur public, modifier l'article 16 de la loi république numérique (181 soutiens) ;
- L'État doit créer et maintenir une forge publique des logiciels libres (174 soutiens) ;
- Création d'une agence ou mission interministérielle pour accompagner à l'usage du logiciel libre (156 soutiens).

Deuxièmement, en matière d'*open data*, la plupart des contributions concernent l'animation de la démarche et l'effectivité de la loi. Les contributions relèvent davantage de la mise en action (comment rendre l'open data effectif ?) plutôt que de prises de position « doctrinales ».

Troisièmement, en matière de données d'intérêt général, le plus marquant est la méfiance affichée par les contributeurs vis-à-vis de l'utilisation par la puissance publique des données du secteur privé. Le risque de surveillance, l'attachement à la *privacy* reviennent souvent dans les commentaires sur les cas d'usage listés par la mission.

Les 3 cas d'usage proposés par la mission les plus commentés ont été :

- les données des opérateurs télécoms (21 commentaires) ;
- les données de prise de rendez-vous et de gestion des consultations de santé (13 commentaires) ;
- les données de consommation énergétique (compteus communicants Linkyu et Gazpar) (10 commentaires).

Quatrièmement, la santé est le domaine de l'action publique le plus souvent cité, loin devant les autres missions de l'État. Le thème santé est évoqué sous plusieurs dimensions (Covid19 bien sûr, mais aussi souveraineté, rôle des hubs, etc.). Ce n'est pas une surprise au regard du contexte sanitaire, mais c'est un fait marquant.

Enfin, la question du rôle des individus en tant que citoyens revient à plusieurs reprises dans les contributions, les commentaires et les réactions aux constats. Il s'agit notamment de mieux les intégrer dans la démarche d'ouverture des données, d'en faire des acteurs à part entière du partage des données (DIG), de les intégrer dans les démarches d'IA.

Liste des sigles

ADEME	Agence de l'environnement et de la maîtrise de l'énergie
AFNET	Association française des utilisateurs du net
AGDAC	Administrateur général de la donnée, des algorithmes et des codes sources
AGD	Administrateur général des données
AMDAC	Administrateur ministériel des données, des algorithmes et des codes source
AMD	Administrateur ministériel de la donnée
AMF	Autorité des marchés financiers
ANR	Agence nationale de la recherche
ANSM	Agence nationale de la sécurité du médicament
ANSSI	Agence nationale de la sécurité des systèmes d'information
AOM	Autorité organisatrice de la mobilité
API	*Application Programming Interface* (interface de programmation d'application)
ARCEP	Autorité de régulation des communications électroniques et des Postes
ART	Autorité de régulation des transports
BRGM	Bureau de recherches géologiques et ministères
CADA	Commission d'accès aux documents administratifs
CASD	Centre d'accès sécurisé aux données
CEA	Commissariat à l'énergie atomique et aux énergies alternatives
CEPD	Comité européen de la protection des données
CEREMA	Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement
CESREES	Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé
CNAF	Caisse nationale des allocations familiales
CNIL	Commission nationale de l'informatique et des libertés
CNRS	Centre national de la recherche scientifique
CRE	Commission de régulation de l'énergie
CRPA	Code des relations entre le public et l'administration
CSA	Conseil supérieur de l'audiovisuel
CSNS	Code statistique non signifiant
CVS	Cadre de vie sécurité
DARES	Direction de l'animation de la recherche, des études et des statistiques
DDT-M	Direction départementale des territoires (et de la mer)
DGCCRF	Direction générale de la concurrence, de la consommation et de la répression des fraudes
DGCL	Direction générale des collectivités locales
DGE	Direction générale des entreprises
DGFiP	Direction générale des finances publiques
DGITM	Direction générale des infrastructures, des transports et de la mer
DGT	Direction générale du travail
DINSIC	Direction interministérielle du numérique et des systèmes d'information et de communication
DINUM	Direction interministérielle du numérique
DITP	Direction interministérielle de la transformation publique
DREES	Direction de la recherche, des études, de l'évaluation et des statistiques
DVF	Base des Demandes de valeurs foncières
EHPAD	Établissement d'hébergement pour personnes âgés dépendantes
EPCI	Établissement public de coopération intercommunale
ETP	Équivalent temps plein
FDA	*Food and drug administration*
FEDER	Fond européen de développement régional
FVA	Fichier des véhicules assurés
HDH	*Health Data Hub*
IA	Intelligence artificielle
IGN	Institut national de l'information géographique et forestière
INPI	Institut national de la propriété industrielle
INRIA	Institut national de recherche en sciences et technologies du numérique
INSEE	Institut national de la statistique et des études économiques
Inserm	Institut national de la santé et de la recherche médicale
MDPH	Maisons Départementales des Personnes Handicapées
NIR	Numéro d'identification au répertoire
OCDE	Organisation de coopération et de développement économiques
OFB	Office Français de la Biodiversité

ONDRP	Observatoire national de la délinquance et des réponses pénales
PIA	Programme d'investissement avenir
PLF	Projet de loi de finances
PRADA	Personne responsable de l'accès aux documents administratifs
RATP	Régie autonome des transports parisiens
RGPD	Règlement général sur la protection des données
RNA	Répertoire national des associations
RSA	Revenu de solidarité active
RSE	Responsabilité sociétale des entreprises
SHOM	Service hydrographique et océanographique de la marine
SIV	Système d'immatriculation des véhicules
SNDS	Système national des données de santé
SpF	Santé publique France
SSM	Service statistique ministériel
SSP	Système statistique public
SSMSI	Service statistique ministériel de la sécurité intérieure

Liste des personnes rencontrées

Shéhérazade Abboub, Parme Avocats · **Henri d'Agrain**, CIGREF · **Guillaume Allenet**, MAIF · **Bernard Allouche**, Cerema · **Benjamin André**, Cozy Cloud · **Théophile Anglès d'Auriac**, PMP Conseil · **Daniel Ansellem**, Ministère de l'intérieur · **Fabien Antoine**, Ministère de la Justice · **Isabelle Antoine**, AXA · **Thierry Aouizerate**, INSEE · **Clémence Arto,** Doctrine · **Charles-Pierre Astolfi**, CNNum · **Stéphane Athanase**, AMUE · **Anne-Marie Atlan**, Nice Métropole · **Valérie Attia**, Ellisphere · **Jean-Gabriel Audebert-Lasrochas**, Trainline · **Servane Augier**, PlayFrance.digital · **Warren Azoulay**, Juris'prédis · **Ludivine Azria**, FFA · **Emmanuel Bacry**, Namr · **Marie-Caroline Baerd**, Cap Gemini Invent · **Emanuele Baldacci**, Commission européenne, DIGIT · **Jérôme Balmes**, FFA · **Murielle Barlet**, DREES · **Julien Barreteau**, Direction générale de l'énergie et du climat · **Anne-Sophie Bartez**, Direction générale de l'enseignement supérieur et de l'insertion professionnelle · **Vincent Bataille**, L'Ouvre Boîte · **Valérie Beaudouin**, UTP · **Sophie Beaupère**, Unicancer · **Laure Bédier**, Direction des affaires juridiques des ministères économiques et financiers · **Michael Benesty**, Open Law* Le droit ouvert · **Gérard Berry**, Académie des sciences · **Pascal Berteaud**, Cerema · **Denis Berthault**, GFII · **Céline Berthon**, Direction générale de la police nationale · **Pierre-Henri Bertoye**, Unicancer · **Pierre-Louis Biaggi**, Orange · **Claire Bieth**, Ministère de l'écologie · **Pierre Bitard**, ANRT · **François-Emmanuel Blanc**, Caisse centrale de la mutualité sociale agricole · **Michel Blancard**, L'Ouvre Boîte · **Jean-Yves Blay**, Unicancer · **Gaël Blondelle**, Eclipse Foundation · **Alizée Bombardier**, Cabinet de la ministre de la Mer · **Thomas Borel**, CNRS · **Zoheir Bouaouiche**, Délégation à la sécurité routière · **Luc Bougé**, Société informatique de France · **Nadi Bou Hanna**, DINUM · **Jacques Bouldoires**, Caisse centrale de la mutualité sociale agricole · **Gaël Bouquet**, MEDEF · **Jean-Marie Bourgogne**, OpenDataFrance · **Cédric Bourillet**, Direction générale de la prévention des risques · **Thomas Bouix**, Office national de la biodiversité · **Jean Brangé**, Boost Conseil · **Luc Brière**, DGCL · **Bastien Brillet**, CADA · **Hélène Brisset**, Directrice du numérique des ministères sociaux · **Jérôme Buzin**, Métropole européenne de Lille · **Christine Cabuzel**, SGAE · Frédéric Cantat, GFII · **Jean-Yves Capul**, Ministère de l'enseignement supérieur · **Bruno Caron**, MFP Services · **Patrick Carrié**, Boost Conseil · **Simon Cauchemez**, Institut Pasteur · **Didier Célisse**, Banque des Territoires · **Clément Chaix**, Agence nationale de la cohésion des territoires · **Claire Chalvidant**, Orange · **Pascal Chambreuil**, Orange · Guillaume Chaslot, Center for Humane Technology · **Pierre Chastanet**, Commission européenne, DG Connect · **Alain Chateau**, Nice Métropole · **Lionel Chaudanson**, Nice Métropole · **Natacha Chicot**, Ministère de l'éducation nationale, de la jeunesse et des sports et ministère de l'éducation supérieure, de la recherche et de l'innovation · **Rémy Choquet**, Laboratoires Roche · **Jennyfer Chrétien**, Renaissance numérique · **Pierre Chrzanowski**, Banque Mondiale · **Olivier Clatz**, Secrétariat général pour l'investissement · **Benoît Claveranne**, AXA · **Sigrid Clavieras**, UTP · **Nathalie Colin**, Direction générale de l'administration et de la fonction publique · **Thomas Cottinet**, Commissariat général au développement durable · **Laurence Comparat**, OpenDataFrance · **Aurelien Conraux**, Service interministériel des archives de France · **Laurent Coudercy**, Office national de la biodiversité · **Thomas Courbe**, Direction générale des entreprises · **Thierry Courtine**, Commissariat général au développement durable · **Bénédicte de Boisgelin**, DGMIC · **Henry de Courtois**, AXA · **Laurent Cytermann**, Conseil d'État · **Marin Dacos**, Direction générale de la recherche et de l'innovation · **Emmanuel de Lanversin**, Direction générale de l'aménagement, du logement et de la nature · **Romain Delassus**, Ministère de la culture · **Nicolas Deloge**, Commission de régulation de l'énergie · **Pierre Demeulemeester**, Cap Gemini Invent · **Thomas Denfer**, Conseil national des greffiers des tribunaux de commerce · **Marie-Laure Denis**, CNIL · **Alexandre De Pablo**, API Entreprise · **Laurent Depommier-Cotton**, Banque des territoires · **Thierry Desponts**, Météo-France · **Roberto Di Cosmo**, INRIA · **Olivier Dion**, Onecub & aNewGovernance · **David Djaïz**, Agence nationale de la cohésion des territoires · **Philippe Dobrinski**, Centre Energy for climate (Institut Polytechnique de Paris) · **Romain Drosne**, Justice.cool · **Emmanuelle Dubée**, Cabinet du ministre de l'intérieur · **Clarisse Dubert**, Secrétariat général pour les affaires européennes · **Claudine Duchesne**, Conseil général de l'économie · **Anne Dufour**, INPI · **Arnaud Dumourier**, Le Monde du Droit · **Tam Kien Duong**, Etalab · **Laure Durand-Viel**, DGMIC · **Eric Durieux**, RENATER · **Luc d'Urso**, PlayFrance.digital · **Laurent Eckert**, Direction générale de police nationale · **Alexis Eidelman**, DARES · **François Elie**, ADULLACT · **Coline Emmel**, Gotham City · **Olivier Esper**, Google France · **Céline Faivre**, Région Bretagne · **Pierre Faure**, AFNeT · **Floriane Fay**, Google France · **Karine Feige**, Apidae tourisme · **Julia Fenart**, France Digitale · **Stéphane Fermigier**, Conseil national du logiciel libre · **Serge Ficheux**, UTAC · **Fabien Fieschi**, Ministère de l'Europe et des affaires étrangères · **Marie-Gabrielle Fournet**, Cabinet du ministre des comptes publics · **Églantine Fraisse**, Direction générale des collectivités locales · **Adrien Friez**, Direction générale de l'administration et de la fonction publique · **Jérôme François**, Numalim · **Claire-Elisabeth Fritz**, Ellisphere · **Alain Ferraton**, Ministère de la transition écologique · **Clémentine Furigo**, MEDEF · **Emile Gabrié**, CNIL · **Kamel Gadouche**, Centre d'accès sécurisé aux données · **Thomas Gageik**, Commission européenne, DIGIT · **Yoann Gantch**, BNF · **Alain Garnier**,

PlayFrance.digital · **Marie Gautier-Melleray**, Délégation à la sécurité routière · **Carole Gay,** Orange · **Pascal Gayat**, PlayFrance.digital · **Edouard Geffray**, Direction générale de l'enseignement scolaire · **Jean-Frédéric Gerbeau**, INRIA · **Mehdi Gharsallah**, Direction générale de l'enseignement supérieur et de l'insertion professionnelle · **Claire Giry**, Inserm · **Claude Gissot**, Caisse nationale de l'assurance maladie · **Marion Glatron**, Rennes Métropole · **François Godineau**, DSS · **Fabrice Gombert**, CNAM · **Etienne Gonnu**, April · **Christine Gonzalez-Demichel**, Ministère de l'Intérieur · **Marine Gossa**, CSA · **Etienne Grass**, Cap Gemini Invent · **Mélisande Grieu**, Wellcom, Conseil de Braincube · **Françoise Guegot**, Région Normandie · **Christelle Guichard**, CADA · **Régis Guyonnet**, IHEMI · **Tiphaine Havel**, CNIL · **Loïc Haÿ**, FNCCR · **Christophe Hazard**, Conseil national des greffiers des tribunaux de commerce · **Alice Herreye Baxter**, Laboratoires Roche · **Matthieu Heurtel**, Cabinet du secrétaire d'État au numérique · **François Hissel**, Office national de la biodiversité · **Brince Huet**, Direction générale de l'aménagement, du logement et de la nature · **Henri Isaac**, Renaissance numérique · **Pierre-Alain Jachiet**, Haute autorité de santé · **Clément Jaquemet**, Commissariat général au développement durable · **Pierre Januel**, Journaliste indépendant · **Benjamin Jean**, Inno3 · **Estelle Jong-Necand**, Cour de cassation · **David Jonglez**, ESRI · **Nadia Joubert**, Caisse centrale de la mutualité sociale agricole · **Elodie Jousset**, Ifremer · **David Julliard**, Délégation à la sécurité routière · **Frédéric Julien**, Infolegale · **Francis Jutand**, Institut Mines Télécom · **Isabelle Kabla-Langlois**, Ministère de l'enseignement supérieur, de la recherche et de l'innovation · **Michelle Kelly-Irving**, Inserm · **Emilie Kerdelhué**, Ministère de l'enseignement supérieur, de la recherche et de l'innovation · **Bernadette Kessler**, Rennes Métropole · **Marine Kettani**, Ministère de la justice · **Pascal Kuczynski**, ADULLACT · **Grégory Labrousse**, Namr · **Laurent Lafaye**, Dawex · **Xavier Lafon**, Cabinet de la ministre de la mer · **Sylvie Lagarde**, INSEE · **Eric Lajarge**, Cerema · **Thierry Lambert**, Direction interministérielle de la transformation publique · **Laurent Laporte**, Braincube · **Tanguy Larher**, SGAE · **Bernard Larroutouru**, Direction générale de la recherche et de l'innovation · **Sylvain Latarget**, IGN · **Laurentino Lavezzi**, Orange · **Macaire Lawin**, Caisse nationale de la solidarité et de l'autonomie · **Noam Léandri**, ADEME · **Arnaud Le Bas**, Délégation à la sécurité routière · **Alexandre Léchenet**, La Gazette des communes · **Alexis Leclerc**, API Entreprise · **Thierry Ledroit**, Cabinet du ministre de l'éducation nationale, de la jeunesse et des sports · **Sylvestre Ledru**, Mozilla · **Gwendal Le Grand**, CNIL · **Michèle Léridon**, CSA · **Cécile Le Guen**, Direction générale des douanes et droits indirects · **Fabrice Lenglart**, DREES · **Xavier Leroy,** Collège de France · **Eloise Lehujeur**, Syntec Numérique · **Ludovic Le Moan**, Sigfox · **Sophie Le Pallec**, Syntec Numérique · **Yann Le Strat**, Santé publique France · **Thomas Lesueur**, Commissariat général au développement durable · **Franck Lethimonnier**, Inserm · **Laura Létourneau**, Délégation ministérielle du numérique en santé · **Guillaume Levieux**, Direction générale de l'aménagement, du logement et de la nature · **Romain Liberge**, MAIF · **Alexandre Liccardi**, Office national de la biodiversité · **Antoine Magnan**, Hôpital Foch · **Selma Mahfouz**, DARES · **Patrick Maison**, ANSM · **Virginie Magnant**, Caisse nationale de la solidarité et de l'autonomie · **Nicolas Marchand**, Direction générale de l'aménagement, du logement et de la nature · **Benoit Marichal**, RATP · **Anne-Claire Marquet**, GFII · **Guillaume Martin**, Métropole européenne de Lille · **Hélène Martin**, DGCL · **Louis Marty**, Doctolib · **Luc Mathis**, Cerema · **Sandrine Mathon**, OpenDataFrance, Mairie de Toulouse & Toulouse Métropole · **Fabrice Mattatia**, Ministère de l'intérieur · **Boris Melmoux-Eude**, Cabinet de la ministre de la transformation et de la fonction publiques · **Boris Mericskay**, Université Rennes-2 · **Denis Merigoux**, Chercheur · **Christophe Merlin**, SNCF · **Jean-Marc Merriaux**, Direction du numérique pour l'éducation · **Antoine Michon**, Cabinet de la ministre de la transformation et de la fonction publiques · **David Miodownick**, Cabinet de la ministre du travail, de l'emploi et l'insertion · **Marina Molin**, MFP Services · **Catherine Mongenet**, FUN MOOC · **Emmanuel Monnet**, Cabinet du ministre de l'économie, des finances et de la relance · **François Moreau**, Ministère de l'agriculture · **Christophe Morel**, Météo-France · **Julien Morel**, Lysios Public Affairs, Trainline · **Laurent Morice**, ADEME · **Gabriel Morin**, Cabinet de la Ministre de la Cohésion des Territoires · **Laura Motet**, Le Monde · **Mathilde Muñoz**, chercheuse · **John-David Nahon**, RATP · **Bruno Nédélec**, Certificare · **Jean-Luc Nevache**, CADA · **Javier Nicolau**, DREES · **Bertrand Nicolle**, Cabinet de la ministre de la Cohésion des Territoires · **Rémy Nollet**, DGGN · **Sébastien Oliveau**, PROGEDO · **Nicolas Orsini**, Ministère de la culture · **Nicolas Osouf**, Ministère de l'écologie · **Akim Oural**, Métropole européenne de Lille · **Bernard Ourghanlian**, Microsoft · **Bertrand Pailhès**, CNIL · **Pauline Pannier**, Cabinet de la ministre de la transformation et de la fonction publiques · **Pierre Paradinas**, Société Informatique de France · **Nikos Paragios**, TheraPanacea · **Nicolas Paris**, InterHop · **Adrien Parrot**, InterHop · **Emmanuel Passilly**, Banque des territoires · **Michel Paulin**, OVH · **Claude Pénicand**, IGN, Institut national de l'information géographique et forestière · **Soizic Penicaud**, Etalab · **Valerie Perhirin**, Cap Gemini Invent · **Patrick Perrot**, DGGN · **Cécile Pertruisot**, Ifremer · **Lucile Petit,** CSA · **Louis Petros**, Namr · **Sébastien Picardat**, Agdatahub · **Olivier Platz**, Secrétariat Général à l'Investissement · **Lionel Ploquin**, Direction générale des finances publiques · **Éric Pol**, aNewGovernance · **Guillaume Poupard**, ANSSI

· **Valérie Porcherot**, Ministère des armées · **Frédéric Pradeilles**, CNES · **Guillaume Pressiat**, InterHop · **Jacques Priol**, CIVITEO · **Renaud Prouveur**, SPALLIAN · **Fadoua Qachri**, MEDEF · **Sara Rami**, Commission de régulation de l'énergie · **Guillaume Rince**, MAIF · **Loïc Rivière**, TECH IN France · **Mathieu Robain**, Unicancer · **Baptiste Robert**, Hacker · **Laurent Romary**, INRIA · **Fabienne Rosenwald**, Ministère de l'enseignement supérieur, de la recherche et de l'innovation · **Gérard Rouicarol**, ANRT · **Juliette Rouilloux-Sicre**, MEDEF · **Khalil Rouhana**, Commission européenne, DG Connect · **Sylvain Rouri**, PlayFrance.digital · **Sylvie Rousset**, CNRS · **Philippe Roux**, FGAO · **Guillaume Rozier**, CovidTracker · **Jean-Renaud Roy**, Microsoft · **Marie Ruault**, DARES · **Christian Quest**, OpenStreetMap · **Sammy Sahnoune**, Inserm · **Laurence Samelson**, Laboratoires Roche · **Ali Saïb**, Cabinet de la ministre de l'enseignement supérieur, de la recherche et de l'innovation · **Sumi Saint-Auguste**, Open Law* Le droit ouvert · **Christophe de Saint-Viance**, Direction générale des douanes et droits indirects · **Jean-Luc Sallaberry**, FNCCR · **Olivier Sanz**, SNCF · **Alice Schoenauer-Sebag**, Inspection générale des finances · **Alain Schuhl**, CNRS · **Virginie Schwarz**, Météo-France · **André Schwob**, DGCCRF · **Béatrice Sédillot**, Commissariat général au développement durable · **Christine Sisowath-Bleiberg**, DARES · **Nicolas Siegler**, MAIF · **Jean-Michel Sommer**, Cour de cassation · **Luc Sonké**, Idemia · **Nathalie Sonnac**, Comité d'éthique des données d'éducation · **Pascal Staccini**, CHU Nice · **Jean-François Sulzer**, consultant · **Camille Sztejnhorn**, Lefebvre Sarrut · **Benoît Tabaka**, Google France · **Amal Taleb**, SAP, membre de Renaissance numérique · **Christine Tartanson**, Onecub & aNewGovernance · **Jean-Luc Tavernier**, INSEE · **Yann Padova**, Backer McKenzie · **Fabrizio Papa Techera**, Lexbase · **Jérôme Teillard**, Parcoursup · **Cécile Teissedre**, Centre Energy for climate (Institut Polytechnique de Paris) · **Olivier Thereaux**, Open Data Institute · **Cédric Thomas**, OW2 · Stéphane Tisserand, MAIF · **Nathalie Thouny**, BNF · **Emmanuelle Tixier**, Région Normandie · **Fabrice Tocco**, Dawex · **Julien Tognola**, cabinet de la ministre de la Transition Ecologique · **Marianne Tordeux**, France Digitale · **David Tortel**, Cabinet du ministre de l'Intérieur · **Marianne Trillat**, PMP Conseil · **Carole Vachet**, Cabinet du secrétaire d'État au Numérique · **Laurent Vachey**, Inspection générale des finances · **Jacques Lévy Véhel**, Case Law Analytics · **Pierre Vercauteren**, Direction du numérique des ministères sociaux · **Henri Verdier**, Ambassadeur pour les affaires numériques · **Denis Veynante**, CNRS · **Florian Veyssilier**, Direction générale de la prévention des risques · **Alice Vieillefosse**, Direction générale de l'énergie et du climat · **Pierre Vigné**, CEREMA · **Cédric Villani**, Assemblée nationale · **Philippe Vimard**, Doctolib · **Patrick Vincent**, Ifremer · **Roberto Viola**, Commission européenne, DG Connect · **Marc Viot**, ADEME · **Yvo Volman**, Commission européenne, DG Connect · **Thomas Wanecq**, Haute autorité de santé · **François Weil**, président du comité du secret statistique · **Renaud Wetzel**, Cabinet du ministre de la Santé et des Solidarités · **Misoo Yoon**, Pôle Emploi · **Stefano Zacchiroli**, Software Heritage · **Philippe Zamora**, Cabinet de la ministre de l'emploi, du travail et de l'insertion · **Ismaël Ziani**, Luxia

Contacts presse
Eric Bothorel
eric.bothorel@assemblee-nationale.fr