



GOVERNEMENT

Liberté
Égalité
Fraternité

MISSION BOTHOREL

Pour une politique publique de la donnée

DECEMBRE 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

Ce document est publié sous une licence libre et ouverte (CC BY 3.0), qui rend obligatoire la mention de la paternité.

Sommaire

Synthèse	6
Recommandations	13
Introduction	17

Partie 1

Une politique au service de toutes les autres 18

1. Les données et les codes sources : de quoi parle-t-on ? 19
2. L'ouverture des données et des codes sources : pour quoi faire ? 22
3. L'ouverture des données et des codes sources : quels risques ? 34
 - CAS D'USAGE – Données et modèles épidémiologiques dans le cadre de la gestion de crise de la Covid19 43
 - CAS D'USAGE – Les statistiques de la délinquance 48

Partie 2

L'ouverture des données et des codes sources publics 53

1. Un cadre juridique à l'avant-garde européenne, qui demeure cependant complexe 54
2. Une dynamique d'ouverture des données et codes sources publics à relancer 66
3. Renforcer le portage politique et la gouvernance 81
 - CAS D'USAGE – Infogreffe et les données de la justice commerciale 98
 - CAS D'USAGE – La base SIRENE 100

Partie 3

Pour une donnée ouverte à tous les usages 102

1. Pour une donnée plus accessible et de meilleure qualité 103
2. Des « hubs » indispensables mais qui doivent être interopérables 113
3. Faciliter l'accès aux données pour les chercheurs 121
4. Adapter le cadre juridique national et européen en conciliant innovation et protection des droits fondamentaux 125
 - CAS D'USAGE – Namr 136
 - CAS D'USAGE – Le projet VidéoProtection ouverte et intégrée (VOIE) 137

Partie 4

Se donner les moyens de nos ambitions 139

1. Renforcer les compétences 140
2. Développer l'utilisation des logiciels libres..... 146
3. Investir dans les infrastructures 150
 - CAS D'USAGE – Trois succès de logiciels libres
d'information géographique..... 159

Partie 5

Les données d'intérêt général 162

1. Une notion imprécise et dont la traduction juridique
manque de cohérence 163
2. Un passage à l'échelle nécessaire mais juridiquement complexe 170
3. Pour une approche par la confiance, incitative et européenne..... 180
4. Privilégier une extension méthodique, progressive et concertée
du partage de données..... 188
 - CAS D'USAGE – Les données du secteur privé utilisées
par la statistique publique 195
 - CAS D'USAGE – Le secteur des assurances et le fichier
des véhicules assurés (FVA)..... 198
 - CAS D'USAGE - Bilan de l'application de la législation en matière
de données de mobilité..... 201

La consultation publique205

Liste des sigles208

Liste des personnes rencontrées211

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'utilisateur. 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 cas¹, 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 imposé à une API par certains producteurs, 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 pluralisme².

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 exsistantes 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 tous³, 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 recueillis, 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